The Dissertation Committee for Emily Jane Bell McTavish certifies that this is the approved version of the following dissertation:

Estimating population histories using single-nucleotide polymorphisms sampled throughout genomes

Committee:

David M. Hillis, Supervisor

Martha K. Smith

C. Randal Linder

David C. Cannatella

Thomas Juenger

Estimating population histories using single-nucleotide polymorphisms sampled throughout genomes

by

Emily Jane Bell McTavish B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2013

Acknowledgements

This work would not have been possible without the support and contributions of my advisor David M. Hillis. My committee, Martha K. Smith, Randy Linder, Tom Juenger and David Cannatella provided invaluable discussion and advice. Discussion and collaboration with the Hillis-Bull lab group including Shannon Hedtke, Tracy Heath, Jeremy Brown, Will Harcombe, Rick Heineman, Thomas Keller, Jeanine Abrams, April Wright and Ben Liebeskind has greatly improved my work. My unofficial committee member Jim Bull provided essential advice. Jerry Taylor, Jared Decker, and Bob Schnabel were a great collaborators who made these projects achievable. I thank Scott Edwards, Robert Wayne, Mike Heaton, Geneviève K. Smith and Roz Eggo for helpful manuscript suggestions. I thank Debbie Davis and the Texas Longhorn Cattleman's Association for research support and genetic samples. The Ecology, Evolution, and Behavior department at the University of Texas was an ideal environment for my research. The encouragement of my friends and family was vital as well.

I was supported by research fellowships awarded by the Graduate Program in Ecology, Evolution, and Behavior at the University of Texas at Austin, The Houston Rodeo, Texas EcoLabs, and NSF BEACON. This material is based in part upon work supported by the National Science Foundation under Cooperative Agreement No. DBI-0939454. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Estimating population histories using single-nucleotide polymorphisms sampled throughout genomes

Emily Jane Bell McTavish, PhD

The University of Texas at Austin, 2013

Supervisor: David M. Hillis

Genomic data facilitate opportunities to track complex population histories of divergence and gene flow. We used 47,506 single-nucleotide polymorphisms (SNPs) to investigate cattle population history. Cattle are descendants of two independently domesticated lineages, taurine and indicine, that diverged 200,000 or more years ago. We found that New World cattle breeds, as well as many related breeds of cattle in southern Europe, exhibit ancestry from both the taurine and indicine lineages. Although European cattle are largely descended from the taurine lineage, gene flow from African cattle (partially of indicine origin) contributed substantial genomic components to both southern European cattle breeds and their New World descendants. We extended these analyses to compare timing of admixture in several breeds of taurine–indicine hybrid origin. We developed a metric, scaled block size (*SBS)*, that uses the unrecombined block size of introgressed regions of chromosomes to differentiate between recent and ancient admixture. By comparing test individuals to standards with known recent hybrid ancestry, we were able to differentiate individuals of recent hybrid origin from other

admixed individuals using the *SBS* metric. We genotyped SNP loci using the bovine 50K SNP panel. The selection of sites to include in SNP analyses can influence inferences from the data, especially when particular populations are used to select the array of polymorphic sites. To test the impact of this bias on the inference of population genetic parameters, we used empirical and simulated data representing the three major continental groups of cattle: European, African, and Indian. We compared the inference of population histories for simulated data sets across different ascertainment conditions using F_{ST} and principal components analysis (PCA). Ascertainment bias that results in an over-representation of within-group polymorphism decreases estimates of F_{ST} between groups. Geographically biased selection of polymorphic SNPs changes the weighting of principal component axes and can bias inferences about proportions of admixture and population histories using PCA. By combining empirical and simulated data, we were able to both test methods for inferring population histories from genomic SNP data and apply these methods to practical problems.

Table of Contents

List of Tables

List of Figures

Chapter 1: New World Cattle Show Ancestry from Multiple Independent Domestication Events

The development of genomic tools has given biologists the ability to analyze variation among DNA sequences to reconstruct population history on a fine scale. Given the close interaction of humans with domesticated species, and the economic importance of domesticated organisms, it is not surprising that humans have developed many of these species as model organisms. Over the past few years, genomic data have been used to reconstruct the domestication history of many of these species, including dogs (Larson *et al.* 2012, Vonholdt *et al.* 2010), horses (Achilli *et al.* 2012), sheep (Kijas *et al.* 2012), and cattle (The Bovine HapMap consortium 2009, Decker *et al.* 2009). The global economic importance of cattle, in combination with the anthropological interest in the shared history of cattle and humans over the past 10,000 years, make cattle an ideal target for spatial genetic research. The first assembly of the cattle genome sequence was published in 2009 (The Bovine Genome Sequencing and Analysis Consortium 2009, Zimin *et al.* 2009). This achievement enables biologists to use genetic variation across breeds and the linkage relationships between those markers to trace the global history of cattle domestication and breed development.

Despite the history of artificial selection in cattle by humans, we here report that genomic data can be used to reconstruct broad aspects not only of breed structure, but also of the global spatial history of domesticated cattle. Similarly to the strong correlation of genetic variation and geography in European human populations (Novembre *et al.*

2008), we also find geographic patterning of genetic variation in cattle. Reconstructing the population history of domesticated species is particularly interesting because historical information can be used to realistically constrain parameter estimates in the modeling process. In addition, although the within- versus among-breed partitioning of genetic variation varies widely across different domesticated species (Freeman *et al.* 2004; The Bovine HapMap Consortium 2009), most established breeds of cattle can be distinguished using genetic markers (Kuehn *et al.* 2011). Thus, the population history — including movement, population subdivision, hybridization, and introgression — of breeds of domesticated species can be tracked using genetic tools.

Domesticated cattle were introduced to the Caribbean in 1493 by Christopher Columbus, and between 1493 and 1512 Spanish colonists brought additional cattle in subsequent expeditions (Rouse 1977). Spanish colonists rapidly transported these cattle throughout southern North America and northern South America. In the intervening 520 years, they have adapted to the novel conditions in the New World. The descendants of these cattle are known for high feed- and drought-stress tolerance in comparison to other European-derived cattle breeds (Clutton-Brock 1999; Barragy 2003). Genetic variation found within these breeds may be especially valuable in the future adaptation of cattle breeds to climate change. Using genomic tools we can reconstruct the global population structure of domesticated cattle, and determine how different lineages contributed to this group's evolution.

Domesticated cattle consist of two major lineages that are derived from

independent domestications of the same progenitor species, the aurochs (*Bos primigenius*). The aurochs was a large wild bovine species found throughout Europe and Asia, as well as in North Africa; it has been extinct since 1627 (Mona *et al.* 2010). These two primary groups of domesticated cattle are variously treated by different authors as subspecies (*Bos taurus taurus* and *Bos taurus indicus*) or as full species (*Bos taurus* and *Bos indicus*). For simplicity, we refer here to these two groups as taurine and indicine cattle, respectively. The most obvious phenotypic differences between these groups are the noticeable hump at the withers (i.e. the shoulders of a four-legged mammal) and the floppy rather than upright ears of indicine cattle (Grigson 1991).

The taurine lineage was probably first domesticated in the Middle East, with some later contributions from European aurochsen; the indicine lineage was domesticated on the Indian sub-continent (MacHugh *et al.* 1997). Although archaeological evidence suggests these domestication events likely occurred only 7,000-10,000 years ago (Perkins 1969; Grigson 1991; Loftus *et al.* 1994; Clutton-Brock 1999), there was already pre-existing spatial genetic structure in the aurochs population at that time. As a result, the taurine and indicine groups are thought to share a most-recent common ancestor 200,000 or more years ago (Hiendleder *et al.* 2008). However, aurochsen and domesticated cattle co-existed in Europe until 1627, and ancient DNA sequencing of aurochs fossils suggests that some large divergences within European domesticated cattle mtDNA may be driven by the repeated incorporation of wild aurochsen into domesticated herds (Bailey *et al.* 1996; Achilli *et al.* 2009). European cattle breeds are largely taurine in origin, whereas

cattle from the Indian subcontinent are indicine. Generally, indicine cattle are more feed-stress and water-stress tolerant, and are more tropically adapted, compared to taurine breeds (Frisch and Vercoe 1977). European taurine cattle have been subjected to more intensive selection for milk and meat production, as well as docility and ease of handling. Taurine and indicine cattle have both contributed genetically to cattle herds in much of Africa (MacHugh *et al.* 1997; Cymbron *et al.* 1999; Loftus *et al*. 1999; Hanotte *et al.* 2002; Freeman *et al.* 2004), and microsatellite analyses show a cline of decreasing indicine heritage from east to west and from north to south across the continent (MacHugh *et al.* 1997). Some researchers have suggested that African taurine cattle are derived from a third independent domestication, from North African aurochsen (Grigson 1991; Bradley *et al.* 1996; Hanotte *et al.* 2002), although there is also archeological and biological support for post-domestication population structuring within North African herds (Loftus *et al.* 1994). The major mitochondrial haplogroups within taurine cattle distinguish European from African cattle, but show patterns of gene flow north across the Mediterranean, particularly at the strait of Gibraltar and from Tunisia into Sicily (Cymbron *et al.* 1999; Beja-Pereira A *et al.* 2006). Wild aurochsen in southern Europe and northern Africa, which likely crossed with the domesticated cattle there, may have carried indicine-like haplotypes, but aurochsen mtDNA sampled from Europe to date groups with extant taurine lineages (Bailey *et al.* 1996).

The first cattle in the Americas were brought to the Caribbean island of Hispaniola, from the Canary Islands, by Christopher Columbus on his second voyage

across the Atlantic in 1493, and Spanish colonists continued to import cattle until approximately 1512 (Barragy 2003). The descendants of these cattle are the main focus of this paper. The cattle from the Canary Islands were descended from animals of Portuguese and Spanish origin, introduced 20 years earlier by early Spanish explorers (Barragy 2003). Therefore these cattle likely shared some ancestry with Northern African breeds of cattle, and thus may have included an indicine genetic component, via earlier gene flow from Africa to the Iberian Peninsula.

The imported cattle reproduced rapidly in the Caribbean, and by 1512 importation of cattle by ship was no longer necessary (Barragy 2003). Caribbean cattle were introduced into Mexico in 1521, and had been moved north into what is now Texas and south into Colombia and Venezuela within a few decades (Barragy 2003). The Spanish settlers relied on these cattle for meat, but largely allowed them free range in the unfenced wilderness. Artificial selection was occasionally imposed by the choice of which individuals to castrate for steers, and which to leave as bulls, except in completely feral herds. Although population sizes plummeted in the late 1800s and herds became more highly managed (Dobie 1941; Barragy 2003), natural selection had driven the evolution of this group for 400 years (Rouse 1977), or between 80 and 200 generations (Kantanen *et al.* 1999). Although precise generation time of feral populations of cattle is unknown, Texas Longhorns in captivity today reproduce by age two.

The mostly feral Spanish cattle were the ancestors of the present day New World breeds including Corriente cattle from Mexico, Texas Longhorns from northern Mexico

and the southwestern United States, and Romosinuano cattle from Colombia (Rouse 1977). This long period of natural selection left these groups better adapted to these landscapes than breeds of more recent European origin. Texas Longhorns are known to be immune to a tick-borne disease known as "Texas fever" or "Cattle tick fever," caused by the protozoan *Babesia bigemina* (Figueroa 1992). This pathogen's vector genus *Boophilus* is known to have been imported with cattle into the New World (George *et al.* 2002). Texas Longhorns have also been described to have far greater drought resistance in comparison to more recently imported European breeds (Dobie 1941).

Research on the genetic diversity that was captured by Spanish colonists in the cattle they chose to bring to the New World has been limited. Some African mtDNA haplotypes and microsatellite alleles are also found in Creole (Caribbean) and Brazilian cattle (Magee *et al.* 2002). Although some references suggest that cattle may have been brought directly from West Africa to the Caribbean and South America as part of the slave trade, there is no direct historical evidence for this hypothesis (Rouse 1977).

Genomic studies have been conducted on cattle breed population structure (The Bovine HapMap Consortium 2009; Decker *et al.* 2009), but the Iberian lineage of New World cattle has not been investigated in depth. In a phylogenetic analysis on a subset of the SNP data set used here, Decker et al. (2009) found New World cattle to be the sister-group of all other European taurine cattle when heterozygous genotypes were treated as ambiguous characters. However, when genotypes were coded as allele counts (0 for AA, 1 for AB, 2 for BB), the New World cattle were placed within the European clade.

For several hundred years the only cattle present in North America were those introduced by the Spanish, but indicine cattle were introduced to North America via Jamaica by the 1860s (Hoyt 1982). In the mid-1900s, indicine cattle were imported into Brazil, and now there are "naturalized" Brazilian indicine (Nelore) and indicine/taurine hybrid (Canchim) breeds. In some samples of Spanish-derived breeds from South America, mtDNA haplogroups and a Y chromosome microsatellite marker suggest indicine introgression in New World cattle (Mirol *et al.* 2003; Ginja *et al.* 2010). In particular, recent male-mediated introgression of indicine alleles into taurine breeds appears common in Brazil (Giovambattista *et al.* 2000).

In this study, we sampled individuals and markers both within New World cattle, and from across the globe, to study the hybrid history of New World cattle. By analyzing nuclear single nucleotide polymorphisms (SNPs) scored in cattle from distinct evolutionary lineages, we were able to estimate introgression on a genomic scale. Previous work on New World cattle relied on mtDNA and Y chromosome markers (Ginja *et al.* 2010). These sequences each reflect the history of a single locus and thus do not have the power to track complex histories of introgression and admixture of genomes. The 47,506 nuclear loci we examined can reflect independent coalescent histories due to recombination and assortment, so they are able to provide much finer resolution of population history than mitochondrial DNA or other single locus markers (Edwards and Bensch 2009).

Results

Our samples of New World cattle included Texas Longhorn cattle (n=114), Mexican Corriente cattle (n=5), and Colombian Romosinuano cattle (n=8). To place these individuals in a global phylogeographic context, we also included previously published data from individuals of 55 other breeds (Decker *et al.* 2009; n=1332; Table 1.1). These cattle were genotyped for nuclear SNP loci across all 29 autosomal chromosomes using the Illumina BovineSNP50 BeadChip, the Illumina 3K chip, or 6K chip. We analyzed two datasets: one (termed the 1.8k dataset) included 1,814 SNP loci present on all three chips, and the other (termed the 50k dataset) included 47,506 SNP loci from the Bovine SNP50 chip. The 1.8k dataset included more extensive sampling of Texas Longhorn cattle (n=114) compared to the 50k dataset (n=40), but a less thorough sampling of the genome.

Average heterozygosity within breeds ranged from 15% (standard deviation 1%) in the indicine breed Gir, to 30% (s.d. 1%) in the taurine Belgian Blue cattle (Table 1.2). The highest heterozygosity was 31% (s.d. 1%) in the recent hybrid Beefmaster. Generally, as expected from the ascertainment panel for the SNP chip (Matukumalli *et al.* 2009), taurine breeds had higher heterozygosity. Breeds of taurine origin averaged heterozygosity of 27%, whereas breeds of indicine origin averaged heterozygosity of 16%. Across New World cattle, average heterozygosity was 28% (Texas Longhorns, 29% s.d. 2%; Corriente, 27% s.d. 2%; Romosinuano, 27% s.d. 1%).

PCA analyses. For both the 50k and the 1.8k datasets, the first axis of our principal components analysis was associated with the indicine–taurine split (Fig. 1.1, Fig. 1.4). This axis accounted for 9% of the variance in genotypes in the 1.8k dataset and 13% in the 50k dataset. The second PC axis was associated with the divergence between European and African taurine cattle, and accounted for 2.6% (1.8k dataset) to 3.2% (50k dataset) of the variance in genotypes. The placement of African cattle reflected both the gradient of indicine introgression across the continent along PC1 and the divergence between European and African taurine cattle along PC2. N'Dama cattle exhibited the most distinct African taurine ancestry. The New World cattle exhibited intermediate ancestry along both of these axes, with both more indicine-like and African-like ancestry than most other European breeds.

The full 50k SNP data set overemphasized genetic diversity in British breeds of cattle (especially Herefords, Fig. 1.4A). Therefore, we re-analyzed the 50k PCA excluding those individuals (Fig. 1.4B), which resulted in the same patterns seen for the 1.8k data (Fig. 1.1).

The first 90 PC axes in the 1.8k dataset, and the first 154 axes in the 50k dataset, were statistically significant based on the Tracy-Widom test (Patterson *et al.* 2006 but see discussion in Methods).

Model-based clustering. In the STRUCTURE analyses of the 1.8k data set (Fig. 1.2), we found strong support for two population subdivisions (K), consistent with the deep

division of indicine and taurine lineages. The 'Hybrid' section shown in Figure 1.2 contains the two cattle breeds derived from recent taurine–indicine crosses: Santa Gertrudis (Brahman/Shorthorn) and Beefmaster (Brahman/Hereford/Shorthorn). The STRUCTURE ancestry estimates of these groups reflect their hybrid origins. At $K = 2$, all New World cattle were estimated to have some indicine ancestry (Fig. 1.2). Romosinuano cattle from Colombia (n = 8) averaged 14% (s.d. 3%) indicine introgression, Corriente cattle from Mexico (n = 5) exhibited 10% (s.d. 3%) indicine introgression, and Texas Longhorns (n = 114) averaged 11% (s.d. 6%) indicine introgression. An ANOVA showed no significant differences in the extent of indicine introgression among these three groups (P=0.16).

Increasing K beyond two subdivisions resulted in only marginal increases in likelihood scores, which suggested possible model over-parameterization. At $K = 3$, the population subdivisions were roughly consistent with groups of indicine cattle, European taurine cattle, and African cattle (the latter represented by N'Dama cattle; group 48). However, the African subdivision was also present in Mediterranean and New World cattle breeds. At higher values of K, among-breed genetic structure predominated. Levels of indicine introgression varied across individual Texas Longhorns. In agreement with Decker et al. (2009), some groups (e.g., Jersey: group 35) consistently showed complex ancestry that was consistent across a range of K values from 3 to 8 (Fig. 1.2).

Correlation between latitude and genotype. For breeds originally developed within

Europe, we found a significant negative correlation ($r = -0.502$; $P = 0.002$) between latitude of country of origin and estimated percent indicine introgression. Percent indicine introgression was estimated from the 1.8k STRUCTURE analyses with $K = 2$.

Discussion

Simulations have demonstrated that inference of complex historical migration models using PCA is difficult as multiple processes can result in the same patterns (Novembre *et al*. 2008; Francois *et al*. 2010). Indeed, even under relatively simple scenarios, such as the admixture between two ancestral groups, admixed individuals can be incorrectly assigned to a third group that appears to be geographically intermediate (Novembre and Stephens 2008). However, when ancestral groups are known, coalescent estimates of admixture between distinct populations are mathematically straightforward. McVean (2009) showed that although PCA is a non-parametric analysis method, coordinates can be predicted from pairwise coalescence times between individuals. This allows a genealogical interpretation to principal component scores. The first principal component can be interpreted as the deepest coalescent event in a tree, and the projection of admixed individuals onto this axis can be used to estimate the proportion of mixture between two parental groups (McVean 2009). As a test case, we were able to correctly reconstruct the known ancestry of recent taurine–indicine hybrid breeds created for agricultural purposes: Santa Gertrudis [group 45], a Brahman-Shorthorn cross developed in 1918, and Beefmaster [group 46], a cross between Hereford, Shorthorn and Brahman cattle

developed in 1954 (the "Hybrid" groups shown in Fig. 1.1 and Fig. 1.2). In addition, we were able to recover the taurine–indicine hybridization cline across Africa along the first principal component (PC1) shown in Figure 1.1.

The second principal component (PC2) shown in Figure 1.1 separates Eurasian from African cattle, indicating a distinctive genomic component in African breeds. Our samples of African cattle breeds all appear to have admixed taurine–indicine ancestry, based on the intermediate position of African cattle on PC1 (Fig. 1.1) and the STRUCTURE analyses when $K = 2$ (Fig. 1.2). However, the distinctiveness of northern African breeds on PC2 (Fig. 1.1), as well as in the STRUCTURE analyses when $K = 3$, indicates additional genomic differentiation in northern African cattle. If African breeds are derived entirely from a mixture of European and Asian cattle, this differentiation must have occurred after the importation of domestic cattle to Africa. Alternatively, this unique African component may be derived from additional domestication events involving north-African aurochsen, as has been suggested previously (Grigson 1991; Bradley *et al*. 1996; Hanotte *et al*. 2002).

Both the principal components analysis (Fig. 1.1 and 1.4) as well as the model-based STRUCTURE analyses (Fig. 1.2) support a hybrid ancestry for New World cattle, although the patterns of hybridization are distinct from the recently constructed hybrid breeds. New World cattle are largely of taurine descent, but they exhibit an average of 11% average indicine ancestry (as estimated from the STRUCTURE analyses of the 1.8k data, with $K = 2$). In this regard, New World cattle are much like some modern breeds

from southern Europe. However, when $K = 3$ in the STRUCTURE analyses (Fig. 1.2), much of this "indicine" component in southern European and New World cattle appears to be more specifically associated with cattle from northern Africa. The principal components analysis is also consistent with the hypothesis that New World cattle (as well as modern breeds from southern Europe) are influenced by ancestral gene flow from northern Africa, based on the placement of these breeds at intermediate positions along PC1 and PC2 in Figure 1.1.

The pattern of African admixture in southern Europe is consistent with movement of cattle across the Straits of Gibraltar during the Moorish invasion and occupation of the Iberian peninsula in the 8th to 13th centuries CE (Loftus *et al.* 1994; Davis 2008; Decker *et al.* 2009). However, sequencing of Bronze Age cattle mtDNA from Spain suggests that earlier African introgression into Iberia may also have occurred (Anderung *et al.* 2005). The elevated disease resistance of Texas Longhorn cattle (compared to northern European cattle breeds that have been imported to southwestern North America) may be partially related to the portions of their genomes that stem from this African ancestry. African N'Dama cattle also exhibit some substantial types of disease resistance (Murray *et al.* 1984), consistent with this hypothesis.

Using model-based clustering analyses we found Spanish-derived New World cattle breeds — Texas Longhorns [group 39], Corriente [group 41], and Romosinuano [group 42] — did not differ significantly in levels of indicine introgression (Fig. 1.3, Table 1.2). The Brazilian breeds Nelore [group 54] and Guzerat [group 56] are recently

developed breeds from indicine stock, which is reflected in our estimates of their ancestry. All sampled New World cattle that are descended from old Spanish imports (114 Texas Longhorns, 5 Corriente, and 8 Romosinuano) show indicine ancestry (estimated by STRUCTURE, with $K = 2$). As well, these breeds group together in our principal components analyses in a position consistent with African introgression. This suggests that introgression from African cattle occurred prior to the introduction of these cattle to the New World. This conclusion is supported by the STRUCTURE analysis of the breeds sampled from southern Europe, particularly Italy, which also show indicine and African ancestry. In fact, among the breeds that we sampled, and using the coarse geographical resolution of 'country,' we found a significant correlation between latitude in Europe and degree of indicine introgression as estimated from STRUCTRE at $K = 2$.

The signal of "indicine" introgression in southern Europe may be somewhat misleading, however, depending on the complexity of domestication history in African cattle. In our STRUCTURE analyses at $K = 3$, the variation captured by the African-like group was not a subset of either of the groups distinguished at $K = 2$, as would be expected from a strictly bifurcating evolutionary process. This suggests that the African subdivision at $K = 3$ is at least partly composed of hybrid taurine–indicine genotypes. But if African cattle are partly derived from a third domestication event involving aurochsen from northern Africa, this deep divergence may be a more important driver of the differentiation between European and African cattle than is indicine introgression. In that case, the 'indicine' component of African and European lineages at $K = 2$ may reflect

African diversity, rather than true indicine ancestry.

Additional analyses, including a more thorough sampling of African and Iberian cattle, are needed for a conclusive determination of the number of independent domestication events in cattle. Although our results cannot exclude the possibility of an independent domestication of aurochsen in northern Africa, the relatively low level of variation captured by the second principal component (2 – 3%) is more consistent with European and African taurine cattle both being derived primarily from a single domestication in the Middle East, with the likely continued but occasional incorporation of genetic material from wild aurochsen in both areas. However, our results do suggest that if there was a third distinct domestication event, it took place in Africa.

Achilli et al. (2009) found a novel haplogroup in Italian cattle (Cabannina, not sampled here), for which the timing of divergence was consistent with introgression from European aurochsen. Although continued introgression of aurochs derived genetic material after the original domestication events probably led to greater diversity in European taurine cattle populations, this diversity is not expected to have been indicine-like, and therefore is not the likely explanation for the indicine genetic component observed in southern European cattle.

There are at least two alternatives to our interpretation of introgression from Africa into Europe prior to the introduction of cattle to the New World: (a) relatively recent hybridization with indicine cattle in the New World (within the last 150 years); and (b) direct importation of cattle from Africa to the Caribbean early in Spanish

colonization. Although we cannot completely rule out the possibility of either of these alternatives, neither of these hypotheses explains the signal of shared ancestry between southern Europe and the Americas. Moreover, the first explanation is inconsistent with the clear African-like genomic component in the New World breeds. Finally, the admixture of genomes across chromosomes indicates ancient, rather than recent, introgression. Thus, the simplest explanation is that introgression of genetic material from African cattle occurred before the importation of cattle to the New World by Spanish colonists. The genetic diversity captured by this hybridization likely provided variation for selection when the ancestors of these animals were transported to North America in the late 1400s to early 1500s. However, there is individual variation among Texas Longhorn cattle, with some individuals showing elevated levels of indicine introgression (Fig. 1.2). This suggests that additional, more recent introgression with indicine cattle may also have occurred in some Texas Longhorn herds.

Our analyses made use of SNP data from across the genome. SNP-chip data have the advantage of being easily replicable, and data reuse across labs is straightforward allowing results to be readily comparable. Furthermore, informative sequence data can inexpensively be generated, allowing investigators to sample many individual cattle. However, it is important to keep in mind the limits of these analyses. As the SNPs selected for the chip were chosen by re-sequencing individuals on an ascertainment panel, genetic diversity represented in that panel is expected to be over-represented in future samples (Albrechtsen *et al.* 2010). In the case of the cattle 50k SNP chip, the

ascertainment panel consisted mostly of taurine cattle. The SNPs were selected to be common polymorphisms in these animals (Matukumalli *et al.* 2009), and therefore diversity estimates based on these data will overestimate diversity in taurine lineages and underestimate diversity in indicine lineages. Average heterozygosity within groups sampled in this study is consistent with this bias. Because of this bias, we did not attempt to estimate diversity metrics such as F_{ST} (Nielsen 2004). In addition, the bias towards polymorphisms found in European taurine breeds as well as for alleles with high minor allele frequencies make these data inappropriate for identifying selective sweeps in New World cattle (Matukumalli *et al.* 2009). Although mathematical methods have been developed to correct for ascertainment biases in some cases (Kuhner *et al*. 2000; Wang and Nielsen 2012), we did not have the appropriate data regarding the ascertainment process to do so in this case. Nonetheless, McVean (2009) showed that although ascertainment bias has an effect on principal component projections, it does not affect the relative placing of samples. Therefore ancestry estimation by this method is robust to this source of bias.

Our results are complementary to previous work on the relationships and genetic diversity among cattle breeds (The Bovine HapMap Consortium 2009; Decker *et al.* 2009). Our conclusions match those of the Bovine HapMap Consortium (2009) for the breeds that were sampled in both studies.

Our findings of introgression in New World Cattle breeds suggest that European–African admixture (which results in greater apparent divergence) may have driven the apparent

sister-group relationship between Texas Longhorns and all other European taurine cattle in some analyses presented by Decker et al. (2009). Our results also suggest that finding may have resulted from imposing a tree-like structure on populations that arose through complex introgression events.

Although we have only a very sparse sampling of Asian cattle breeds from outside India, our results suggest that these animals are also of hybrid taurine–indicine origin. The possibility of introgression of genetic material from populations or species not sampled in our analysis limits our ability to make inferences about Asian cattle, but they promise to be an interesting area for future research. Although Kawahara-Miki and colleagues (2011) suggested that Japanese cattle are sister to all other domesticated cattle, their omission of an indicine breed in their analyses makes this conclusion difficult to test. In addition, introgression among taurine and indicine lines would produce a similar result in a tree-based analysis.

The recent publication of the first *Bos indicus* genome sequence (Canavez *et al.* 2012) will provide an opportunity to identify specific alleles of African or indicine origin that have contributed to the adaptation of New World cattle breeds. This is of particular interest given the rapidly changing global climate. New World cattle in general, and Texas Longhorns in particular, are reported to exhibit resilience to drought and harsh climatic conditions (Barragy 2003; Riely 2011). Previous work has shown that New World cattle are an important reservoir of genetic diversity (Giovambattista *et al.* 2000). As we show here, some of this diversity appears to derive from ancient introgression via

African cattle.

Materials and Methods

Sampling. We examined 1,495 cattle from 58 breeds, including 874 European individuals, 127 individuals from New World breeds, 209 primarily indicine individuals, 260 individuals of African or hybrid origin, and 17 individuals from Asia (Table 1.1). 1420 of these cattle were genotyped for 54,609 single nucleotide loci using the Illumina BovineSNP50 BeadChip (Van Tassell *et al.* 2008; Matukumalli *et al.* 2009). We refer to this as the 50k data set. These data were generated as described by Decker *et al.* (2009). We genotyped an additional 75 Texas Longhorn cattle on one of the Illumina 3K (25 individuals), or 6K (50 individuals) chips. These data were generated commercially at NeoGen/GeneSeek (Lincoln, NE). These individuals were not included in the 50k data set analysis as the amount of missing data would have greatly exceed the amount of genotype data. Across the 3k, 6k and 50k SNP chips are 1,814 shared SNPs, which we refer to as the 1.8k data set.

Filtering. We removed SNP loci from our analysis if they were missing from the SNP chip documentation and could not be decoded or identified, if average heterozygosity was > 0.5 in 10 or more breeds (which indicated paralogy or repeat regions), or if call rate was lower than 0.8 in 10 or more breeds (which indicated null alleles, or changes in flanking regions preventing DNA hybridization to the array). We also removed markers if they were not found in at least 30% of sampled individuals. We then removed

individuals with > 10% missing data across the markers on the 29 autosomes from our analyses, and subsequently removed markers which were missing in >10% of individuals. 1369 individuals (Table 1.1) and 47,506 markers (available on datadyrad.org, provisional DOI: doi:10.5061/dryad.42tr0) were included in the filtered 50k dataset. 1461 individuals (Table 1.1) and 1,814 markers were included in the filtered 1.8k dataset (available on datadyrad.org, provisional DOI: doi:10.5061/dryad.42tr0). The list of markers included in the 50k and 1.8k data sets are available with the data on datadryad.org.

To minimize the effects of possible recent hybridization (within the last 150 years), we considered shared genetic signal among Texas Longhorn (USA), Corriente (Mexico), and Romosinuano (Colombia) cattle. We excluded one Texas Longhorn individual from our analyses of New World, as high indicine introgression (~38%) and large unrecombined chromosomal blocks of indicine ancestry suggested that it was a recent indicine hybrid.

Breed was assigned based on information given by the owner when an individual was sampled. We removed from our analyses two 'Nelore' individuals that do not show any indicine ancestry, strongly suggesting that breed was incorrectly assigned. For all SNPs, we used physical map locations from the University of Maryland assembly of *B. taurus*, release 3 (Zimin *et al.* 2009). Geographic locations of breeds were treated at the centroid latitude and longitude of the country from which the breed was known to have originated.

Phasing. To impute missing data, we required phased haplotype data. Our SNP data were generated as genotype data, rather than as haplotype data. Therefore, if an individual was heterozygous at multiple loci, the phase relationship between alleles is not known. We divided our genotype data by chromosome and used a statistical method to phase our genotype data into haplotypes. Genotypes for all individuals in the 50k data set were phased, and missing data (mean 2%, Table 1.2), were imputed using fastPHASE (Scheet and Stephens 2006). We used the defaults of 20 random starts and 25 iterations of the EM algorithm. To avoid biasing haplotype imputation towards preconceived breed structure, we did not use subpopulation identifiers. We allowed fastPHASE to estimate the number of haplotype clusters via a cross-validation procedure described in (Scheet and Stephens 2006). Pei et al. (2008) found fastPHASE to be the most accurate among available genotype imputation software. Imputed genotype data were used only in the principal components analysis.

Principal Components Analysis. Principal components analysis requires complete data, and we therefore performed PCA on imputed, 50k and 1.8k genotype data. PCA was performed using *smartpca* in the software package EIGENSOFT (Patterson *et al.* 2006; Price *et al.* 2006; Price *et al.* 2009). The number of significant principal components was calculated using *twstats* in the *eigenstrat* package (Patterson *et al.* 2006). However, Tracy-Widom statistics are estimated based on the assumption of a random sampling of markers, and ascertainment bias in SNPs selected for inclusion on the utilized SNP chip likely violate this assumption (Tracy and Widom 1994; Patterson *et al.* 2006).

ANOVA. Analysis of variance (ANOVA) was performed to test for differences in indicine introgression across New World breeds (Corriente, Romosinuano, and Texas Longhorns), as estimated by the principal components analysis of the 50k and 1.8k data sets and by STRUCTURE. ANOVA was performed in R using *aov* in the stats package (R Core Team 2010).

Model-based clustering. Multi-locus model-based clustering, as well as the associated assignment of individuals to populations, was performed using STRUCTURE (Pritchard *et al.* 2000). The SNPs on all 29 autosomes were analyzed using the linkage model based on their UMD3.0 map positions. Recombination rate was treated as uniform. To test for convergence, and to aid in parallelization, analyses were repeated 5 times for each value of K, with a run time of 20,000 iterations and a burn-in of 1,000 iterations. We tested values of K from 2 to 9. In simulations, Evanno et al. (2005) found that run lengths above 10,000 iterations were not additionally beneficial, but that much longer runs still varied in likelihood. We used longer runs as our problem was more complex, and tested for convergence across runs after 5 runs were completed using Structure Harvester (Earl and Vonholdt 2012). STRUCTURE analyses were only conducted on the full unphased 1.8k data set.

We selected the optimum number of ancestral populations (K) from our STRUCTURE analyses using Evanno et al.'s (2005) method, implemented in Structure Harvester (Earl and Vonholdt 2012). This method avoids overfitting by selecting the value of K for which there is the largest increase in likelihood from $K-1$ to K.

We did not calculate F_{ST} values between breeds because the ascertainment in SNP discovery and assay design was strongly biased towards loci common in taurine cattle, which leads to the overestimation of diversity within these breeds.

Table 1.1. Breeds included in the analysis

Figure legend	Name	Region of origin	sample size 50k (1.8k)
1	Shorthorn	Great Britain	99
2	Maine Anjou	Southern Europe	5
3	White Park	Great Britain	4
4	Kerry	Great Britain	3
5	Angus	Great Britain	90
6	Devon	Great Britain	4
7	Hereford	Great Britain	98
8	Simmental	Northern Europe	77 (78)
9	Red Angus	Great Britain	15
10	Tarentaise	Southern Europe	5
11	Belgian Blue	Northern Europe	4
12	South Devon	Great Britain	3
13	Murray Grey	Australia (via Great Britain)	4
14	English Longhorn	Great Britain	3
15	Red Poll	Great Britain	5
16	Limousin	Southern Europe	100
17	Dexter	Great Britain	4
18	Finnish Ayrshire	Northern Europe	10
19	Guernsey	Channel Islands	10
20	Welsh Black	Great Britain	2
21	Norwegian Red	Northern Europe	21
22	Gelbvieh	Northern Europe	8
23	Scottish Highland	Great Britain	8
24	Pinzgauer	Northern Europe	5
25	Salers	Southern Europe	5
26	Montbeliard	Southern Europe	5
27	Blonde d'Aquitaine	Southern Europe	5
28	Galloway	Great Britain	4
29	Holstien	Northern Europe	85 (100)
30	Sussex	Great Britain	4
31	Charolais	Southern Europe	53

Table 1.1 continued

Figure legend	Name	Region of origin	sample size 50k (1.8k)
32	Belted Galloway	Great Britain	4
33	Brown Swiss	Northern Europe	10
34	Piedmontese	Southern Europe	29
35	Jersey	Channel Islands	10
36	Romagnola	Southern Europe	29
37	Chianina	Southern Europe	7
38	Marchigiana	Southern Europe	2 (4)
39	Texas Longhorn	Southwestern USA	40 (114)
40	Texas Longhorn cross	Southwestern USA	5
41	Corriente	Mexico	5
42	Romosinuano	Colombia	8
43	Hanwoo Korean	Asia	7
44	Japanese Black	Asia	10
45	Santa Gertrudis	Indicine-Taurine Hybrid (USA)	24
46	Beefmaster	Indicine-Taurine Hybrid (USA)	24
47	Senepol	Africa	36 (37)
48	N'Dama	Africa	59
49	Tuli	Africa	4 (5)
50	Ankole-Watusi	Africa	5
51	N'DamaXBoran	Africa	42 (41)
52	Sheko	Africa	20
53	Boran	Africa	44
54	Nelore	Brazil (via India)	58 (60)
55	Brahman	United States (via India)	98
56	Guzerat	Brazil (via India)	3
57	Sahiwal	India/Pakistan	10
58	Gir	India	25

Column “Figure legend” shows label number for figures 1.2 and 1.3. Sample sizes show the number of individuals included in the analysis after filtering the 50k and 1.8k data sets. Sample sizes for the 1.8k data set were identical to the 50k except where noted.

Table 1.2. Detailed information for breeds included in the analysis

Figure legend	Name	Country	Lat.	Long.	avg % taur	s.d.	Avg. het.	s.d.
1	Shorthorn	England	52.00	0.75	1.00	0.003	0.25	0.02
2	Maine Anjou	France	46.00	2.00	1.00	0.001	0.28	0.00
3	White Park	England	52.00	0.75	0.99	0.002	0.21	0.02
4	Kerry	Ireland	53.00	-8.00	0.98	0.011	0.29	0.00
5	Angus	Scotland	55.95	-3.20	0.99	0.006	0.27	0.02
6	Devon	England	52.00	0.75	0.98	0.009	0.26	0.01
7	Hereford	England	52.00	0.75	0.99	0.010	0.29	0.03
8	Simmental	Switzerland	47.00	8.00	0.99	0.008	0.28	0.01
9	Red Angus	Scotland	55.95	-3.20	0.99	0.002	0.28	0.01
10	Tarentaise	France	46.00	2.00	0.99	0.007	0.29	0.01
11	Belgian Blue	Belgium	50.83	4.00	0.99	0.007	0.30	0.00
12	South Devon	England	52.00	0.75	0.97	0.018	0.27	0.00
13	Murray Grey	Australia	-27.00	133.00	0.98	0.017	0.28	0.01
14	English Longhorn	England	52.00	0.75	1.00	0.001	0.20	0.01
15	Red Poll	England	52.00	0.75	0.98	0.020	0.26	0.01
16	Limousin	France	46.00	2.00	0.98	0.019	0.29	0.01
17	Dexter	Ireland	53.00	-8.00	0.97	0.024	0.22	0.04
18	Finnish Ayrshire	Finland	64.00	26.00	0.97	0.013	0.28	0.00
19	Guernsey	Channel Islands	49.47	-2.58	0.98	0.013	0.25	0.01
20	Welsh Black	Wales	51.50	-3.22	0.99	0.001	0.29	0.00
21	Norwegian Red	Norway	62.00	10.00	0.98	0.013	0.29	0.01
22	Gelbvieh	Germany	51.00	9.00	0.96	0.020	0.29	0.01
23	Scottish Highland	Scotland	55.95	-3.20	0.99	0.010	0.25	0.01
24	Pinzgauer	Austria	48.12	16.12	0.98	0.008	0.29	0.01
25	Salers	France	46.00	2.00	0.95	0.018	0.27	0.04
26	Montbeliard	France	46.00	2.00	0.96	0.004	0.28	0.01
27	Blonde d'Aquitaine	France	46.00	2.00	0.96	0.023	0.29	0.01
28	Galloway	Scotland	55.95	-3.20	0.97	0.011	0.26	0.01
29	Holstien	Netherlands	52.50	5.75	0.92	0.029	0.30	0.01

Table 1.2 continued

Figure legend	Name	Country	Lat.	Long.	avg % taur	s.d.	Avg. het.	s.d.
30	Sussex	England	52.00	0.75	0.94	0.028	0.25	0.02
31	Charolais	France	46.00	2.00	0.96	0.031	0.30	0.01
32	Belted Galloway	England	52.00	0.75	0.95	0.032	0.26	0.01
33	Brown Swiss	Switzerland	47.00	8.00	0.95	0.018	0.26	0.01
34	Piedmontese	Italy	42.83	12.83	0.94	0.032	0.29	0.01
35	Jersey	Channel Islands	49.25	-2.17	0.92	0.022	0.24	0.01
36	Romagnola	Italy	42.83	12.83	0.87	0.027	0.27	0.01
37	Chianina	Italy	42.83	12.83	0.89	0.045	0.28	0.02
38	Marchigiana	Italy	42.83	12.83	0.84	0.015	0.27	0.01
39	Texas Longhorn	USA	32.00	100.00	0.89	0.068	0.28	0.05
40	Texas Longhorn cross	USA	32.00	100.00	0.86	0.082	n/a	n/a
41	Corriente	Mexico	23.00	-102.00	0.91	0.034	0.29	0.02
42	Romosinuano	Colombia	4.00	-72.00	0.86	0.029	0.27	0.01
43	Hanwoo Korean	Korea	38.50	127.00	0.86	0.036	0.27	0.00
44	Japanese Black	Japan	36.00	138.00	0.89	0.021	0.23	0.03
45	Santa Gertrudis	Brahman/ Shorthorn	n/a	n/a	0.68	0.037	0.29	0.01
46	Beefmaster	Brahman/ Hereford/ Shorthorn	n/a	n/a	0.67	0.035	0.31	0.01
47	Senepol	Caribbean (RedPoll/ N'Dama)	14.00	-14.00	0.83	0.068	0.28	0.01
48	N'Dama	Guinea	11.00	-10.00	0.71	0.075	0.21	0.02
49	Tuli	Zimbabawe	-20.00	30.00	0.64	0.022	0.26	0.01
50	Ankole-Watusi	Uganda	1.00	32.00	0.45	0.032	0.22	0.01
51	N'DamaXBoran	Ndama/ Boran	n/a	n/a	0.37	0.024	0.26	0.01
52	Sheko	Ethiopia	8.00	38.00	0.35	0.030	0.23	0.00
53	Boran	Kenya	1.00	38.00	0.18	0.022	0.22	0.01
54	Nelore	Brazil	-10.00	-55.00	0.00	0.004	0.15	0.01
55	Brahman	India	20.00	77.00	0.03	0.028	0.18	0.01
56	Guzerat	Brazil	-10.00	-55.00	0.01	0.006	0.15	0.01

Table 1.2 continued

Figure legend	Name	Country	Lat.	Long.	avg % taur	s.d.	Avg. het.	s.d.
57	Sahiwal	Pakistan	30.00	70.00	0.00	0.002	0.15	0.01
58	Gir	India	20.00	77.00	0.00	0.003	0.15	0.01

Column labeled “Figure legend” shows label number for figures 1.2 and 1.3. Sample sizes show the number of individuals included in the analysis after filtering the 50k and 1.8k data sets. “Country” is the country of breed origin, or describes the cross for known hybrids. Latitude and longitude coordinates are from CIA World Factbook (CIA 2008) for the breed’s country of origin. Average % taurine (avg % taur) was estimated for the 1.8k data set using STRUCTURE (Pritchard *et al.* 2000) at K=2. Average heterozygosity (Avg. het.) was calculated for the 50k dataset.

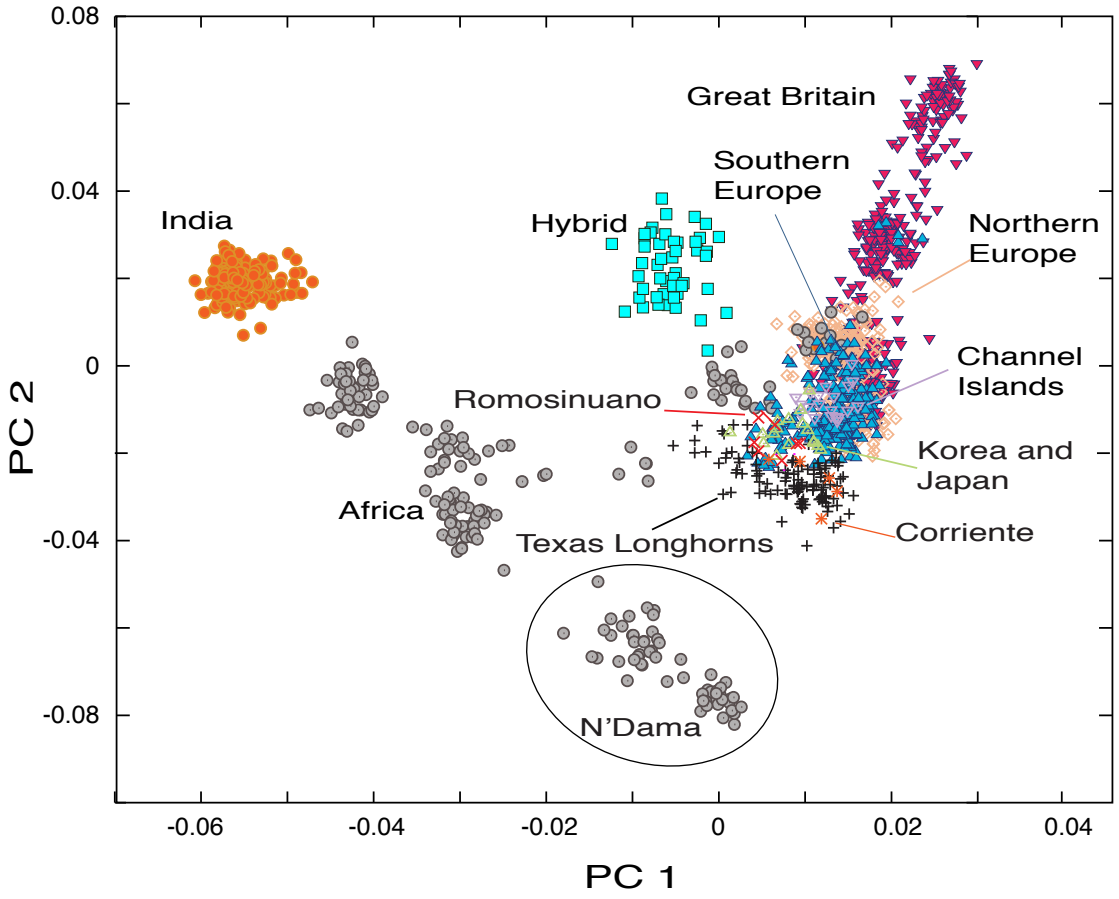


Figure 1.1. A statistical summary of genetic variation in 1,461 cattle individuals genotyped at 1,814 SNP loci.

Individuals are grouped by the region from which their breed originated, as described in Table 1.1. Principal component 1 (PC1) captures the split between indicine and taurine domestications. The position of individuals along this axis can be interpreted as the proportion of admixture between these two groups. PC2 captures the European–African split within taurine cattle.

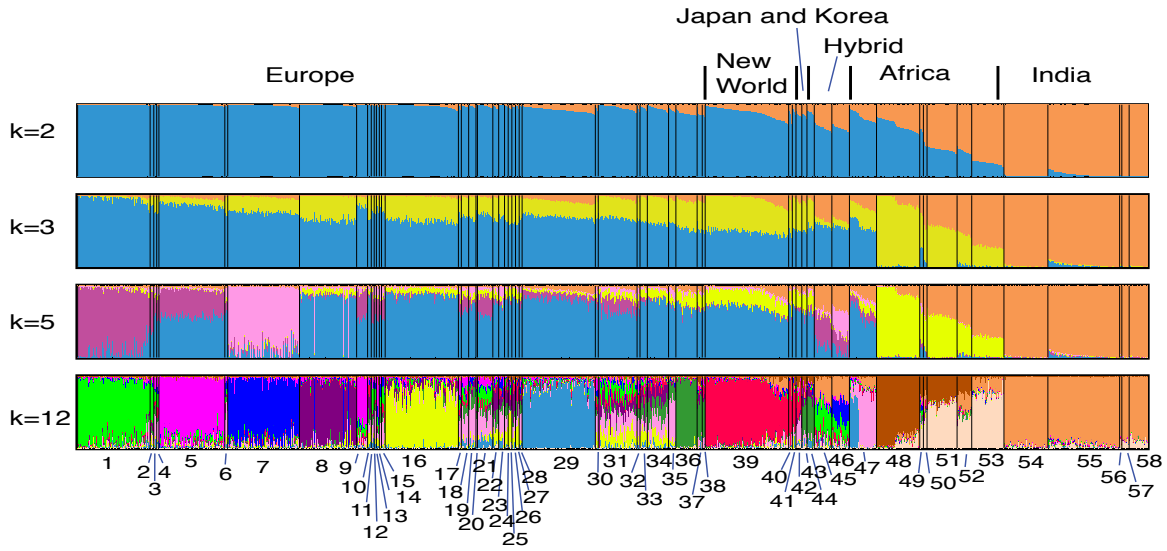


Figure 1.2. Model-based population assignment for 1461 individuals based on 1,814 SNP markers.

Estimated using STRUCTURE (Pritchard *et al*. 2000) and plotted using Distruct (Rosenberg 2003). Individuals are represented as thin vertical lines, with the proportion of different colors representing their estimated ancestry deriving from different populations. Individuals are grouped by breed as named when sampled; breeds are arranged by regions, and are individually labeled by numbers at the bottom. Breed name associated with each number is listed in the "Figure legend" column in Table 1.1. The best-supported number of ancestral populations was two (*K*=2). This split captures the known indicine–taurine split. 'Hybrid' labels refer to Santa Gertrudis [group 45] and Beefmaster [group 46] cattle breeds developed from indicine–taurine crosses within the past 100 years. At *K*=3, population groupings were not consistent across runs, but generally followed the division between indicine, European taurine, and African taurine cattle. At higher values of *K* individual breed structure predominated, although some breeds (e.g., Jersey, group 35) consistently showed complex ancestry. *K*=5 and *K*=12 were selected to demonstrate these patterns.

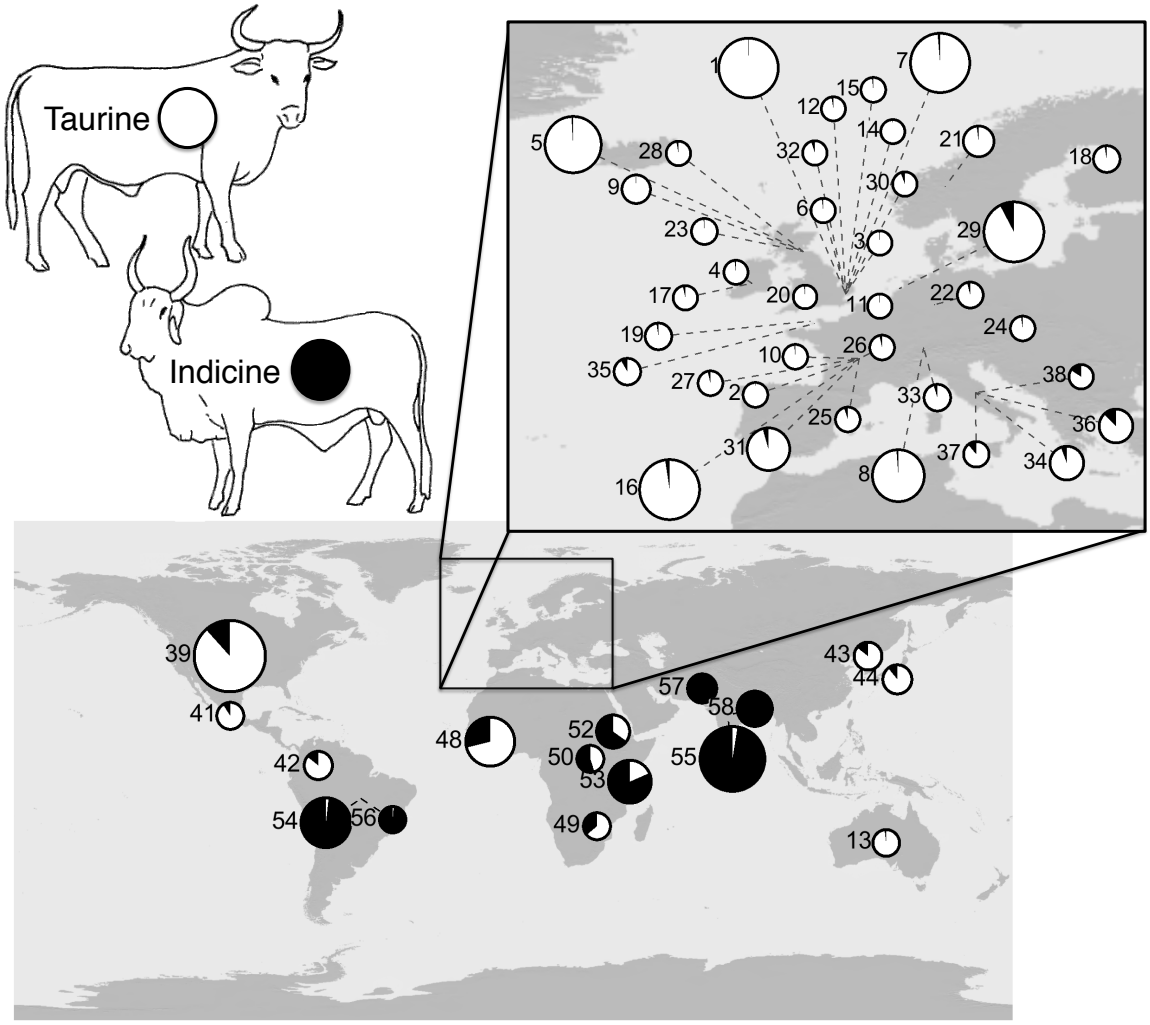


Figure 1.3. Geographic structure of breed ancestry.

Estimated at *K*=2 on the 1.8k dataset using STRUCTURE (Pritchard *et al*. 2000). Taurine ancestry is indicated in white and indicine ancestry in black in the pie diagrams. Breed name associated with each number is listed in the "Figure legend" column in Table 1.1. Note higher levels of indicine introgression in southern Europe, particularly for the Italian breeds Romagnola [group 36], Piedmontese [group 34], Chianina [group 37] and Marchigiana [group 38]. The Brazilian breeds Nelore [group 54] and Guzerat [group 56] are recently developed breeds from indicine stock. Pie chart size is scaled to sample size. Breed location is based on the latitude/longitude coordinates from CIA World Factbook

(CIA 2008) of the breed's country of origin. Silhouettes of cattle are reproduced from (Grigson 1991). This figure was created using the software package GenGIS (Parks *et al.* 2009).

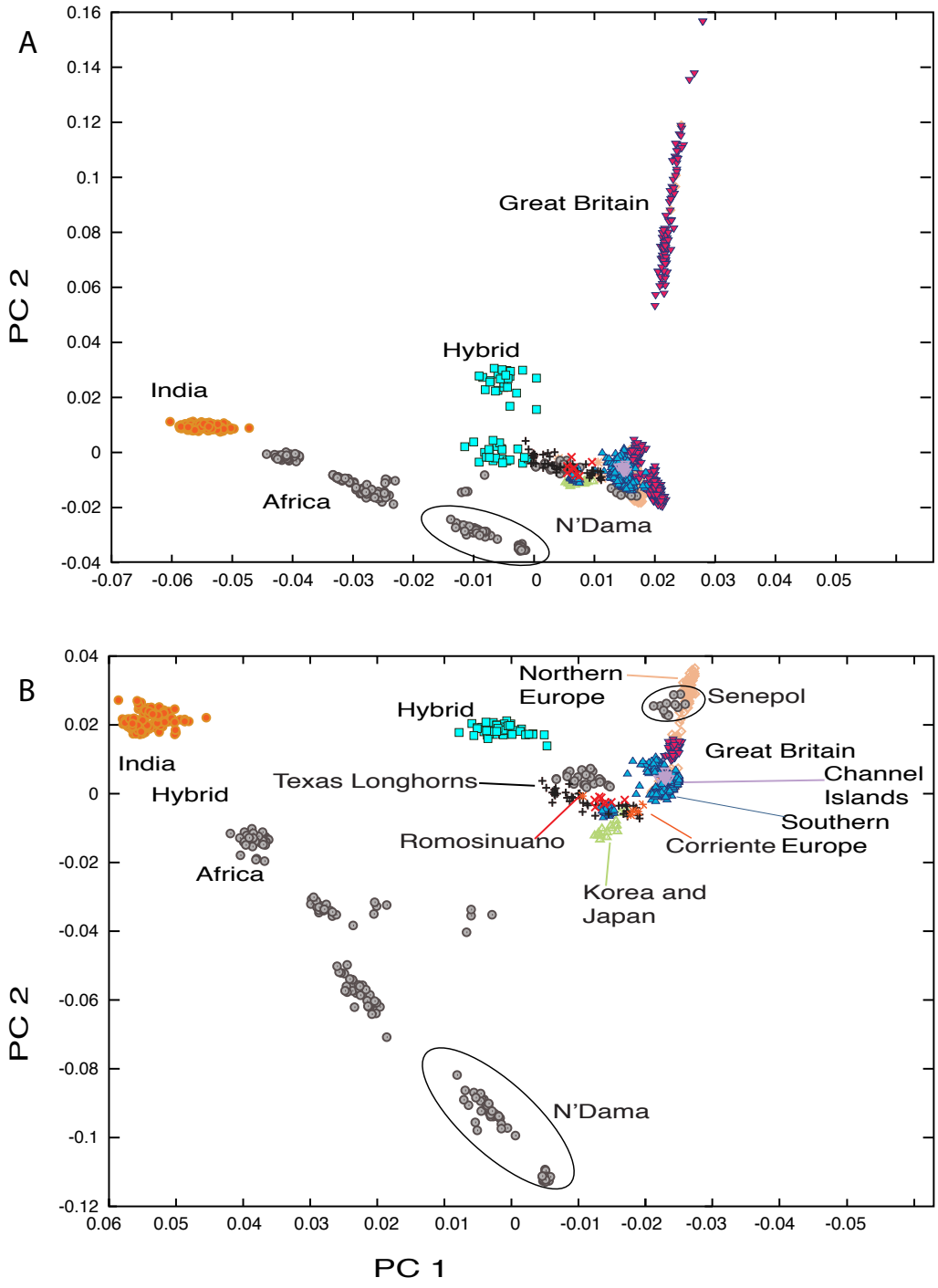


Figure 1.4. Principal components analysis of genetic variation for 47,506 SNP loci

(*A*) A principal components analysis of genetic variation in 1,369 cattle individuals genotyped at 47,506 SNP loci. Individuals are grouped by the region of breed origin, as described in Table 1.1. Ascertainment bias in the SNPs selected for the chip is reflected in greater apparent genomic diversity of British cattle breeds (especially Hereford). (*B*) A principal components analysis based on a subset of individuals shown in figure 1.4A; all individuals with a PC2 score > 0.04 in the full analysis were removed prior to performing the PCA to clarify relationships among the remaining groups.

Chapter 2: A genomic approach for distinguishing between recent and ancient admixture

Geographically widespread species often exhibit considerable genetic diversity across populations. Estimating the timing and extent of divergence and gene flow among such populations is important for understanding the current structure and differentiation of individual genomes. Genomic data provide opportunities to capture the complexity of the evolutionary history of populations and reconstruct even rare historical events. Although many studies have used mitochondrial DNA (mtDNA) to study geographic variation and gene flow, the clonal maternal inheritance of mtDNA limits its usefulness (Edwards *et al*. 2005). Many independently segregating loci are required to capture the multiple coalescent histories that comprise a genome with hybrid ancestry (Edwards and Bensch 2009). For example, the conclusion that most humans of non-African descent have some Neanderthal ancestry (Green *et al*. 2010; Reich *et al*. 2010) would not have been possible without sufficient genomic data to capture coalescent histories that involve less than 4% of the genome. In this study we developed a method for analyzing the structure of individual genomes to simultaneously capture information about the timing and character of admixture between groups of interacting populations.

Migration is an important evolutionary force. Admixture between populations can provide the genetic variation for selection to act on, or swamp local adaptation (Slatkin 1987). To make sense of the evolutionary history of populations it is necessary to understand patterns of gene flow. Here we explore an approach for reconstructing gene

flow using genomic data that explicitly models recombination and admixture through time. Using this approach, we can capture complex population histories and gain fine-scale information about the timing of admixture events. In addition, we can assess whether regions of the genome differ in their evolutionary history from the patterns expected as a result of lineage sorting and coalescence.

Lawson *et al*. (2012) developed and implemented a chromosome painting model for estimating the ancestry of regions of the genome. This model has been applied to estimating gene flow among chimpanzee populations for conservation purposes (Bowden *et al*. 2012), as well as to reconstructing fine scale human population structure associated with cultural differentiation (Haber *et al*. 2012). We extend the applications of this model to comparing timing of admixture between populations by comparing the unrecombined chromosomal fragment size inherited from each parent population against reference individuals for which timing of admixture is known. Inferences about timing of admixture can distinguish between alternate phylogeographic hypotheses (Vila *et al*. 2005). In addition, conservation biologists can use admixture information to select appropriate candidates for conservation (Allendorf *et al*. 2001)

We applied this technique to estimating timing of admixture to cattle populations. There is a considerable database of genomic and genetic information of cattle as a result of their economic and environmental importance (Womack 2005). This makes cattle ideal for studying the relationships between genome architecture and hybridization. There are at least two major groups of domesticated cattle, which were independently domesticated

from geographically disjunct populations of the wild aurochs (*Bos primigenius*) around 10,000 years ago (Loftus *et al*.1994). The descendants of the cattle domesticated in the Middle East are designated *Bos taurus*, whereas those domesticated on the Indian subcontinent are *Bos indicus*. The genome of *Bos taurus* was the first assembled domesticated species (The Bovine Genome Sequencing and Analysis Consortium *et al*. 2009; Zimin *et al*. 2009). A full genome sequence of *Bos indicus* has also been reported and aligned to *Bos taurus* genome (Canavez *et al*. 2012). These two groups of cattle are more divergent than their domestication dates would suggest—a result of pre-existing spatial genetic variation in the ancestral aurochsen. Estimates of the age of the most recent common ancestor of all domesticated cattle range from 200,000–2,000,000 years ago (Loftus *et al*. 1994). Nonetheless, these two lineages interbreed readily (Demeke *et al*. 2003). They are variously by different authors treated as species (*B. taurus* and *B. indicus*) or as subspecies (*B. t. taurus* and *B. t. indicus*). For simplicity and clarity, we refer to these two lineages as taurine cattle and indicine cattle, respectively.

Taurine and indicine cattle have some important phenotypic differences. Indicine cattle have a fatty hump at the withers, as well as a dewlap (Grigson 1991). They also have increased heat tolerance compared to taurine cattle, and an ability to digest lower quality forage (Cartwright 1980). Although indicine cattle are more common worldwide (Cartwright 1980), taurine cattle have been subject to more extensive artificial selection in Europe. As a result of this intense artificial selection for a number of agriculturally desirable traits (such as high meat and milk production), taurine breeds account for the

vast majority of beef and dairy production, based on the numbers of registered progeny in the United States (Heaton *et al*. 2001).

In this study we compare patterns of admixture among four groups with hybrid ancestry between taurine and indicine cattle: (1) a group comprised of two breeds of known recent admixed ancestry dating to the early 1900s (Beefmaster and Santa Gertrudis); (2) Spanish-derived New World cattle and two ancient hybrid lineages from Africa; (3) a predominantly taurine western African breed (N'Dama); and (4) a predominantly indicine eastern African breed (Boran).

The Santa Gertrudis breed was developed from a cross of Brahman and Shorthorn cattle in 1918 (Rhoad 1949; Warwick 1958). Beefmaster was developed from a cross of Brahman, Shorthorn, and Hereford cattle beginning in 1908 (Warwick 1958). Previous work (McTavish *et al*. 2013) has shown that Santa Gertrudis cattle have 32% ± 4 standard deviation (s.d.) indicine ancestry, and Beefmaster cattle have 33% ± 4 (s.d.) indicine ancestry. Given estimates of effective generation time in cattle of between 2 and 5 years (Kidd and Cavalli-Sforza 1974; Chikhi *et al*. 2004), these two recent hybrid breeds reflect admixture within the past 20–50 generations.

African cattle have a complex history. Taurine cattle have been present in North Africa since at least 4,000 BP and indicine cattle were introduced to eastern Africa by c. 2000-3000 BP (Clutton-Brock 1999) and were present in western Africa by 1000 BP (Freeman *et al*. 2004). These independent introductions of taurine and indicine cattle to Africa set up an historic cline of hybridization across Africa. This cline is marked by

cattle of predominantly indicine ancestry in the east and cattle of predominantly taurine ancestry in the west, and may be reinforced by geographically variable selection for trypanosome resistance (Loftus *et al*. 1994; Freeman *et al*. 2004). In this study we were particularly interested in two African breeds: N'Dama cattle and Boran cattle from western and eastern Africa, respectively. About 32 % ± 2 (s.d.) of N'Dama genomes appear to be derived from indicine origins, as are 82% ± 2(s.d.) of Boran cattle genomes (McTavish *et al*. 2013). Some of this admixed ancestry extends into southern Europe, likely as a result of transport of cattle across the Straits of Gibraltar (Cymbron *et al*. 1999; Anderung *et al*. 2005).

New World cattle, represented here by Texas Longhorns, Corriente, and Romosinuano breeds, are the descendants of cattle brought to the New World by Spanish colonists approximately 500 years ago. These cattle also exhibit genomic signatures of admixed ancestry between African hybrid cattle and European cattle, consistent with their southern European origins (McTavish *et al*. 2013). Another possibility, however, is that some or all of the indicine genomic component found in New World breeds (11% ± 6 s.d.) may be a result of recent introgression with indicine cattle in the New World, rather than ancient admixture (Martínez *et al*. 2012; McTavish *et al*. 2013). Based on variation among 19 microsatellite loci, Martínez *et al*. (2012) found that indicine ancestry was present in all 27 sampled New World cattle populations, but that this signal of indicine ancestry were absent in 39 cattle breeds sampled from the Iberian peninsula. Gautier and Naves (2011) also found evidence of excess African ancestry in New World cattle

relative to European cattle. This pattern of African and indicine ancestry across all New World cattle may be explained by importation of admixed African cattle into the Canary Islands off of western Africa; Spanish colonists used these islands as cattle depositories (Rouse 1977; Gautier and Naves 2011). These admixed cattle from the Canary Islands may have been included with Iberian cattle in the first introductions to the New World.

Here we contrast the patterns of admixture seen in cattle of ancient hybrid origin (as described above) with the patterns seen in recent taurine–indicine hybrid breeds of known origin (Santa Gertrudis and Beefmaster), and use these differences to assess timing of admixture in New World cattle.

The independent domestication events that led to taurine and indicine cattle captured divergent genetic information. By examining repeated instances of admixture between the two genomes at a range of time scales, we here examine which ancestor's alleles have been maintained through time. In addition, we examine whether or not the genomic architecture of introgression is similar between independent origins of hybrid lineages. We also use patterns of recombination and sizes of linkage blocks to compare the ages of admixture events, and assess the evidence for recent versus ancient admixture. The *SBS* metric we developed can be applied to assessing the timing of admixture in other species.

Methods

We analyzed 1369 individuals of 58 breeds genotyped at 54,001 SNP (single nucleotide polymorphism) loci using an Illumina 55K chip (Matukumalli *et al*. 2009). We performed analyses on all breeds, but we focused on the four groups of seven breeds that were of particular interest to our questions, as described above.

Filtering and Phasing

We removed SNP loci from our analysis if (1) they were missing from the manifest and could not be decoded; (2) if average heterozygosity was > 0.5 in 10 or more breeds (an indication of paralogy or repeat regions); (3) if call rate was lower than 0.8 in 10 or more breeds (an indication of null alleles); or (4) if data from a given locus was missing in at least 70% of sampled individuals. We then removed individuals with > 10% missing data across the loci on the 29 autosomes and the X chromosome, and subsequently removed loci that were missing in >10% of individuals. 1369 individuals and 47,506 autosomal markers remained after filtering. The list of loci is available with the data at doi:10.5061/dryad.42tr0. For the X chromosome, we also excluded the estimated pseudoautosomal region (PAR) based on the UMD3.1 genome assembly (physical map locations >=137109768 bp; Zimin *et al*. 2009). After removal of the PAR, 872 X-linked loci remained in our analyses.

We phased the SNP loci into haplotypes, and imputed missing data simultaneously using fastPHASE (Scheet and Stephens 2006). We used fastPHASE to estimate the number of haplotype clusters via a cross-validation procedure described in (Scheet and Stephens 2006). Pei *et al*. (2008) found fastPHASE to be the most accurate

among available genotype imputation software. We conducted all analyses on phased data.

Sexing

Because gender was not recorded for some samples from previously collected datasets, we estimated gender from polymorphisms at markers thought to be on the X chromosome. As males only have one X chromosome, they are not expected to be polymorphic at X-linked loci. We excluded the PAR region of the X, as described in above in *Filtering*. Based on samples of known gender, as well as the bimodality observed in plotting polymorphism on the X chromosomes across all individuals, we assigned individuals with less than 1% polymorphism at X-linked loci as males. We used the 1% threshold to account for possible genotyping error. We recoded the <1% of called heterozygous alleles in males as missing data. By this assignment, we had a total of 352 females and 1017 males.

Model-Based Clustering

We performed model-based clustering analysis for each chromosome using Bayesian parametric analysis, based on a fit to the Hardy-Weinberg equilibrium model, as implemented in the software STRUCTURE (Pritchard *et al*. 2000). In order to differentiate histories across chromosomes, we independently analyzed each of the 29 autosomes and the X chromosome. The SNPs from each chromosome were analyzed using the linkage model based on their UMD3.1 map positions (Zimin *et al*. 2009).

Recombination rate was treated as uniform. For X-linked loci in males, we used hemizygous genotypes. We ran 5 independent Markov chain Monte Carlo runs.

Significance Testing

We used a bootstrap resampling approach (Efron 1981) to test for significant departures from median admixture proportions of individual chromosomes within breeds. As distributions of proportions are not normally distributed, we could not use methods that assume normality for these tests. We tested for significant differences across chromosomes in the median and the variation of admixture proportions, compared to the expected distributions assuming uniform admixture across chromosomes within breeds. For these tests, we first calculated the median taurine ancestry for each chromosome grouping each breed. We created a distribution of values of taurine ancestry consisting of all the proportions for all the individuals of each breed. We then drew bootstrap samples of new chromosomes by sampling from this distribution. We then compared the actual median introgression of each chromosome in the original data to the expected distribution (if admixture were uniform across chromosomes). We performed 50,000 resampling replicates to generate the expected distribution, and used a Bonferroni corrected α-value of 0.0002 (two tailed-test). This value was calculated by taking a p value of 0.025 for a two-sided test, and dividing by 120 (30*4) to account for multiple tests of 30 chromosomes across 4 different groups.

To test for significant deviations in variability across chromosomes, we calculated the absolute difference from the group median for each individual for each chromosome,

and performed an ANOVA on these values (Levene 1960). As the deviations from the mean were not normally distributed, we created an expected distribution of *F*-statistics by resampling from this pool and performing an ANOVA on the distributions of the randomized deviations from the median (Boos and Brownie 2004). We performed 5,000 resampling replicates in this test. All ANOVAs were performed in Python using the *scipy.F_oneway* function (Jones *et al*. 2001).

Chromosome Painting

We used Li and Stephen's (2003) copying model, as implemented in ChromoPainter (Lawson *et al*. 2012), to estimate regions of ancestry across the chromosome. This model relates the patterns of linkage disequilibrium (LD) across chromosomes to the underlying recombination process and avoids the assumption that LD must be block-like by computing LD across all sites simultaneously. This method uses a Hidden Markov Model to reconstruct a sampled haplotype as it would be generated by an imperfect copying process from all other haplotypes in the population. Ancestry of regions can be inferred by estimating copying probabilities from two or more donor populations for chromosomal regions of admixed individuals. An estimate of 'copying' from a population is equivalent to inferring that a particular region of a haplotype coalesced with an individual from the identified population more recently than with an individual of another population. Using this approach, we were able to assign ancestry of regions along chromosomes, even when there were no fixed differences between populations, because the method takes into account the physical position of loci

and makes estimates based on all sites simultaneously. We used an estimated effective population size for all breeds together of 4000, as estimated from the ChromoPainter software. This estimate is consistent with the low estimates (in the 100s) of effective population sizes for most European breeds of cattle (The Bovine HapMap Consortium 2009).

The two donor populations (taurine and indicine) were based on individuals that were estimated to have < 2% of introgressed ancestry (McTavish *et al*. 2013). The donor populations are used to represent the taurine and indicine lineages (Lawson *et al*. 2012). These donor populations consisted of 502 taurine individuals and 151 indicine individuals. Because we were interested in admixed groups, we *a priori* set equal probabilities of copying chromosomal regions from either of these donor populations. Because the likelihood estimate is dependent on the order in which individual haplotypes are considered, we used the averaged estimates across five random runs of the expectation maximization algorithm.

Timing of Admixture

Baird (1995) showed that following admixture, the breakdown of linkage among alleles from parental population occurs slowly and may be used to estimate time of contact. Theoretical expectations for breakdown of linkage through time are mathematically straightforward, and were described by Fisher (1954). However, genetic details such as differences in recombination rate across chromosomal regions present obstacles for making empirical estimates of time from admixture data. To obtain a metric

of timing for introgression events, we calculated the scaled median introgressed block size, which we refer to as *SBS* (scaled block size). *SBS* is calculated using the less common ancestor as the "introgressed" genome. For each individual and chromosome, we calculated median block size as a proportion of the chromosome (range from 0 –0.5). We used medians rather than means because distributions were skewed. We scaled block sizes by the total proportion of that chromosome inherited from the introgressed ancestor. If only one recombination event occurred since admixture, the introgressed region would be expected to lie in a single segment, and the scaled average block size would be 1. However, as further recombination and backcrossing occurs, the introgressed material is divided up across the genome, and the block size decreases. This introgressed block size is expected to be strongly correlated with time since introgression (Baird 1995; Rieseberg *et al*. 2000) For each individual we averaged values of *SBS* across all autosomal haplotypes.

Results

Model-Based Clustering

We reconstructed the distributions of ancestry across chromosomes for individuals in each of the four study groups (Figure 2.1). We averaged admixture proportions for each individual for each chromosome across runs. All runs converged on highly congruent estimates. The maximum range of ancestry estimates for an individual across all 5 runs was 3 percentage points. We found that several chromosomes exhibited

significant differences in median introgression levels compared to expectations under a model of equal introgression across chromosomes (Table 2.1; Figure 2.2). Although no particular chromosome showed extreme patterns of introgression in all four groups, the X chromosome had reduced indicine ancestry in recent hybrid cattle, New World cattle, and N'Dama cattle (Table 2.1). This pattern was not shared with eastern African Boran cattle.

Chromosome Painting

We reconstructed the ancestry of chromosomal regions through chromosome painting (Figure 2.3). This analysis indicated differences in structure of ancestry both within and between populations. As expected, large non-recombined tracts of DNA from each ancestral linage were apparent in recent hybrid breeds, such as Beefmaster. The analysis also indicates differences among groups within breeds. N'Dama cattle showed breed substructure associated with time of sample collection and herd of origin (Figure 2.3).

Quantitative Comparisons

Estimates of *SBS* differed across groups (one way ANOVA; $P<0.00001$). New World cattle and both African groups each showed older admixed ancestry compared to recent hybrid breeds, as reflected in smaller introgressed fragment sizes (Figure 2.4). We found that recent hybrid cattle have larger non-recombined blocks of introgressed genetic material, as measured by the *SBS* metric, compared to New World cattle, N'Dama cattle, or Boran cattle (Figure 2.4). *SBS* can differentiate timing of introgression even among individuals with the same overall proportion of introgression (Figure 2.5). Each New

World cattle, N'Dama cattle, and Boran cattle had modal SBS close to 0.05, whereas in recent hybrid cattle it was approximately 0.11. The minimum *SBS* value for an individual of known recent hybrid cattle was 0.09. Using this value as a cutoff for admixture within the last 100 years, we found a few individuals within both N'Dama and New World cattle breeds that showed evidence of relatively recent indicine introgression. These bins are shown in orange in Figure 2.4. An individual of New World origin with an *SBS* value of 0.076 also had a large non-recombined block of indicine origin on the X chromosome (marked by an '*' in Figure 2.3), strongly suggesting recent admixture.

Discussion

The similarity between scaled indicine fragment sizes between African cattle and New World Spanish-derived cattle suggests that the admixture observed between taurine and indicine lineages in New World cattle predated or was concurrent with their introduction to the New World. This pattern is consistent with the hypothesis of crossing between admixed African lineages and taurine lineages from the Iberian Peninsula in the Canary Islands (the source for at least some of the Spanish cattle imports into the New World; Rouse 1977).

Introgression becomes progressively harder to reconstruct with time. Denser genomic sampling is required to reconstruct smaller blocks of linkage disequilibrium (Villa-Angulo *et al*. 2009). However, if populations are not subject to gene flow following admixture, eventually introgressed blocks will become fixed in the population

(Ungerer *et al* 1998; Rieseberg 2000). After introgressed regions in a population are fixed, no further information about timing of admixture can be gleaned from introgressed block size.

In addition to differences in timing of admixture among groups, we also found differences among individuals within groups. Individual *SBS* values were unimodal and close to symmetric in Boran and recent hybrid cattle, which is consistent with a uniform admixture history within those groups. In contrast, the distributions of scaled fragments sizes appear skewed to the right in both N'Dama and New World cattle (Figure 2.4). The smaller peaks at higher levels of introgression in these groups are consistent with those individuals having undergone more recent admixture. We used the lowest *SBS* score of known recently admixed cattle as a lower cutoff to distinguish individuals of likely recent admixture. However, the *SBS* metric relies on scaling sizes of introgressed fragments by the overall introgressed proportion of each respective chromosome. This scaling may limit the usefulness of this approach at low levels of introgression. This metric can be applied to estimate timing of admixture in other species for which at least some known hybrid individuals have been sampled.

Applying the 'ChromoPainter' chromosome painting model to our SNP data (Li and Stephens 2003; Lawson *et al*. 2012) has several advantages. Because of bias in the selection of loci used on the SNP-chip (Matukumalli *et al*. 2009), each SNP has high minor allele frequencies and is highly polymorphic even within groups. Therefore, although our analysis included many loci, each individual locus provides limited ancestry

information. The high minor allele frequencies reduce the power for methods that rely on pairwise allele sharing to estimate LD and timing of admixture, such as *rolloff* (Moorjani *et al*. 2011; Patterson *et al*. 2012). But by co-estimating across all loci and using linkage information to inform our model of genomic regions of ancestry using Chromo Painter, we were able to integrate information from many sites to estimate recombination break points since admixture. For these analyses we used physical map distances from the UMD3.1 assembly of the taurine (*Bos taurus*) genome (Zimin *et al*. 2009). Ideally we would use genetic map distances for our chromosome painting analyses. Previous linkage maps have found concordance between physical map and genetic map locations (Arias *et al*. 2009), but there is not currently a full linkage map for the SNP loci we analyzed. In addition, although the *Bos indicus* genome has been sequenced, it was assembled through alignment to the *Bos taurus* genome. Thus, some synteny changes may have been missed. Synteny changes would impact recombination rates between these genomes, and could bias estimates of absolute dates of admixture. We mitigated this bias by using comparisons among groups derived from recombination between these same two ancestral lineages. By comparing among groups, we can standardize for variation that results from changes in recombination rate across regions between these two taxa.

In all groups sampled, we found at least one chromosome that was not consistent with a uniform distribution of introgressed ancestry across chromosomes. However, with the exception of the X chromosome, these differences were not consistent across groups.

The observed variation across groups in the distribution of ancestry across chromosomes may result from differences in the natural and artificial selective regimens that these populations have experienced. Alternatively, if these breeds underwent strong bottlenecks following admixture, chromosomes with highly biased ancestry could have become more common in those populations as a result of drift. The variation across groups in which chromosomes have biased introgression suggests that differences are not due to chromosomal rearrangements or other barriers to recombination.

In contrast, the X chromosome was the most extreme outlier in three groups: recent hybrid cattle, N'Dama cattle, and New World cattle. Indicine ancestry was reduced on X as compared to the autosomes in all three of these groups.

Several genetic characteristics differentiate the X chromosome from autosomes. The population size of the X chromosome is reduced compared to that of autosomes, since males only have one X chromosome. In addition, apart from the pseudoautosomal region, the X chromosome only undergoes recombination in females. The combination of these two facts makes drift a stronger force on the X chromosome than in the autosomes, and could result in differences in apparent admixture among chromosomes. Although the Y chromosome is acrocentric in indicine cattle and submetacentric in taurine cattle, there are no obvious karyotypic differences between the X chromosomes in the two groups (Frisch *et al*. 1997).

Sex-biased introgression may also explain the reduced indicine component on the X chromosomes of the various admixed groups. If admixed males from an F_1 generation

were preferentially used in backcrosses to one parental line, this practice would decrease the contribution of introgression on the X chromosome relative to the autosomes. As standard breeding practices tend to preserve female offspring in preference to male offspring, this scenario seems unlikely.

Rapid evolution of sex chromosomes has been shown to lead to reproductive isolation among populations (Kitano *et al*. 2009). However, the lack of biased introgression on the X chromosome in Boran cattle suggests that X chromosome–autosomal incompatibilities between taurine and indicine cattle are not responsible for the reduced levels of apparent indicine introgression seen in the X chromosomes of other admixed breeds.

Evidence of indicine ancestry is nearly absent on the X chromosome in New World cattle, with the exception of the recent hybrid individual marked in Figure 2.3. This absence of X-linked indicine loci is consistent with the hypothesis that New World cattle are derived from crossing taurine Iberian cattle with admixed western African cattle. This cross would decrease the already reduced introgression on the X chromosome in western African cattle. The near complete absence of indicine ancestry makes the X chromosome sequences useful for detecting recent indicine introgression in New World cattle.

Table 2.1. Chromosomes that fall outside of the expectations for distribution of taurine ancestry, assuming ancestry proportions are uniform across chromosomes.

Medians and ranges are shown for chromosomes with more extreme values than expected based on bootstrap samples, as described in text (Bonferroni corrected p-value of 0.0002). Asterisks (*) indicate significant deviations below the lower significant cutoff value or above the upper significant cutoff value. The bootstrap distributions and median values for outlying chromosomes are shown in Figure 2.2.

	Significant chromosomes	Median proportion taurine	Range	Lower cutoff at α=0.0002	Upper cutoff at α=0.0002	*P*-value
Recent Hybrids	*Bootstrap sample*	*0.69*	*0.558-0.813*	*0.59*	*0.79*	
	5	0.50	0.152-0.720	*		<0.00002
	8	0.58	0.320-0.857	*		<0.00004
	18	0.79	0.416-0.988		*	<0.00014
	X	0.79	0.356-0.999		*	<0.00014
New World cattle	*Bootstrap sample*	*0.91*	*0.647-0.995*	*0.71*	*0.99*	
	X	1.00	0.718-0.999		*	<0.00002
N'Dama	*Bootstrap sample*	*0.71*	*0.632-0.771*	*0.65*	*0.76*	
	2	0.78	0.549-0.866		*	<0.00002
	5	0.79	0.527-0.882		*	<0.00002
	9	0.64	0.435-0.726	*		<0.00008
	19	0.78	0.469-0.898		*	<0.00002
	21	0.64	0.493-0.744	*		<0.00008
	X	0.88	0.522-0.993		*	<0.00002
Boran	*Bootstrap sample*	*0.19*	*0.134-0.251*	*0.14*	*0.23*	
	4	0.14	0.042-0.378	*		<0.00004
	7	0.13	0.066-0.251	*		<0.00002
	10	0.23	0.141-0.435		*	<0.0001
	11	0.14	0.064-0.373	*		<0.00004
	14	0.37	0.132-0.550		*	<0.00002
	29	0.26	0.128-0.485		*	<0.00002

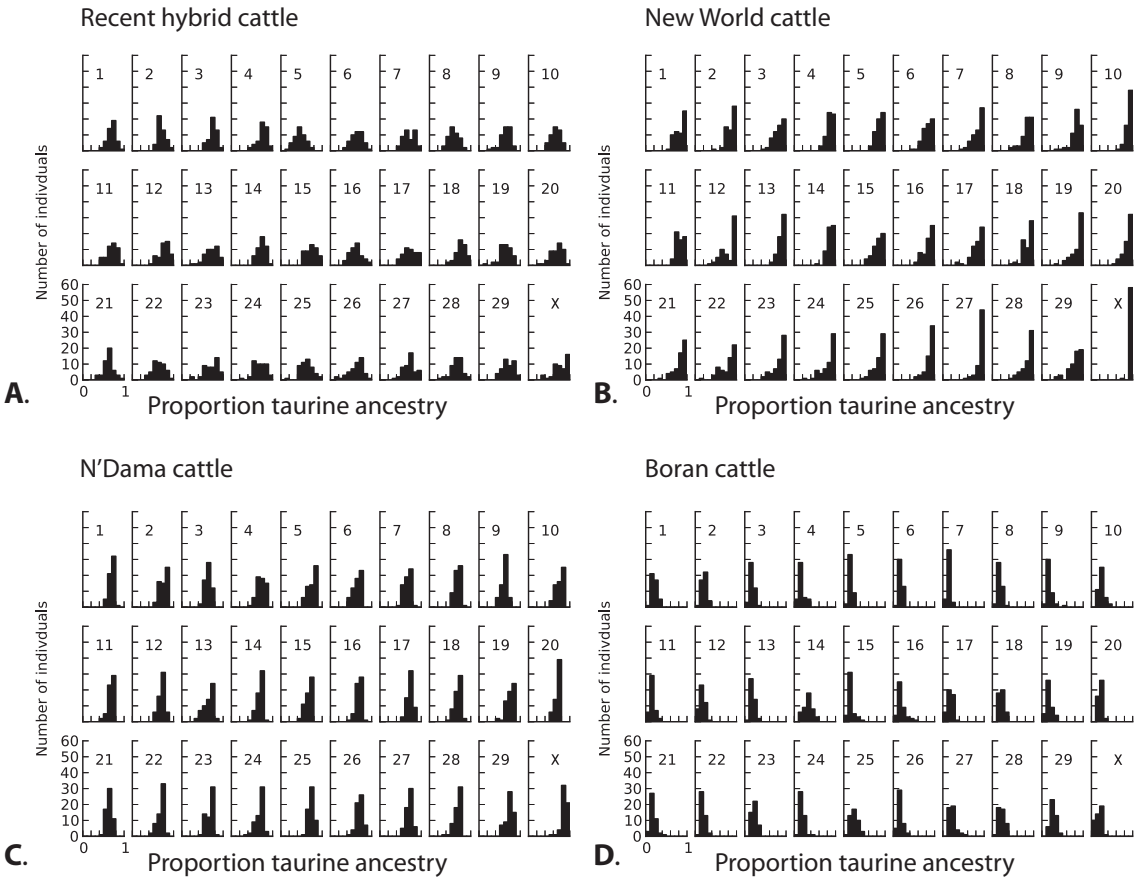


Figure 2.1. Histograms showing estimated proportion of taurine ancestry for individuals on each chromosome.

The X axis indicates estimated taurine ancestry as calculated using STRUCTURE. Y axes are scaled to percentages of sampled individuals. Panels show: **A.** Recent hybrid cattle (Beefmaster and Santa Gertrudis). **B.** New World cattle (Texas Longhorns, Corriente, and Romosinuano). **C.** Western African cattle (N'Dama), **D.** Eastern African cattle (Boran). Note near complete absence of admixture on the X chromosome in New World cattle.

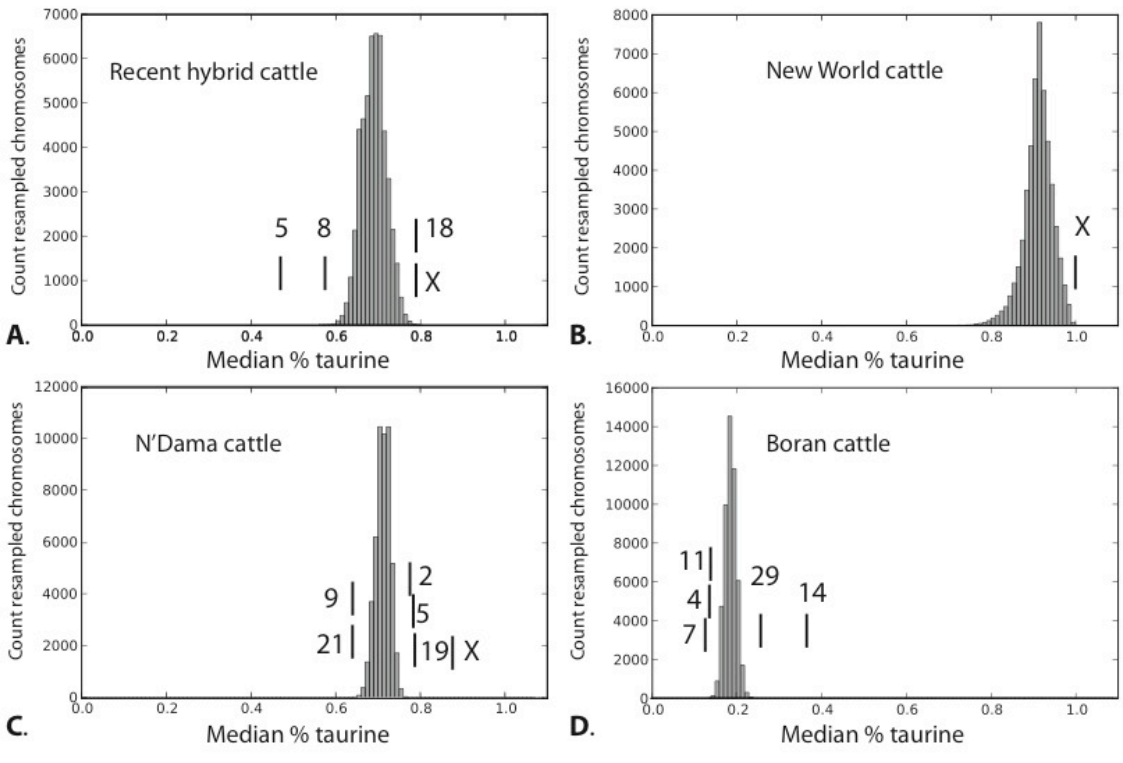
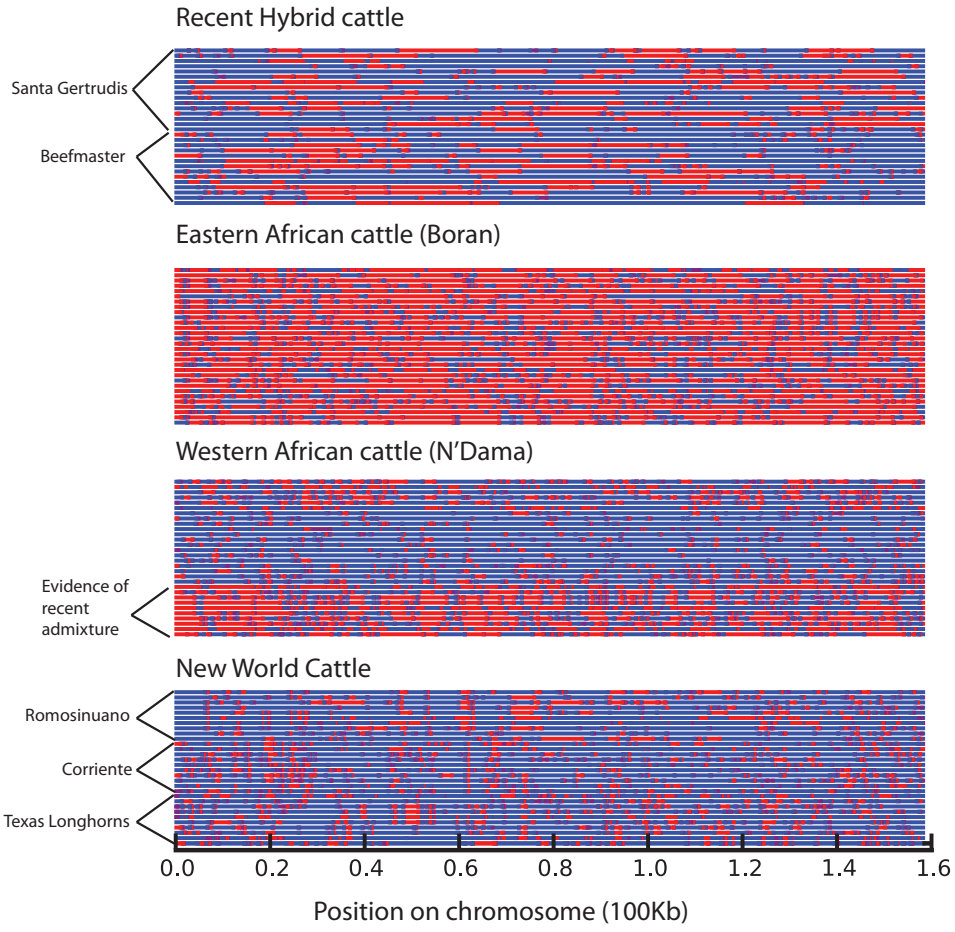
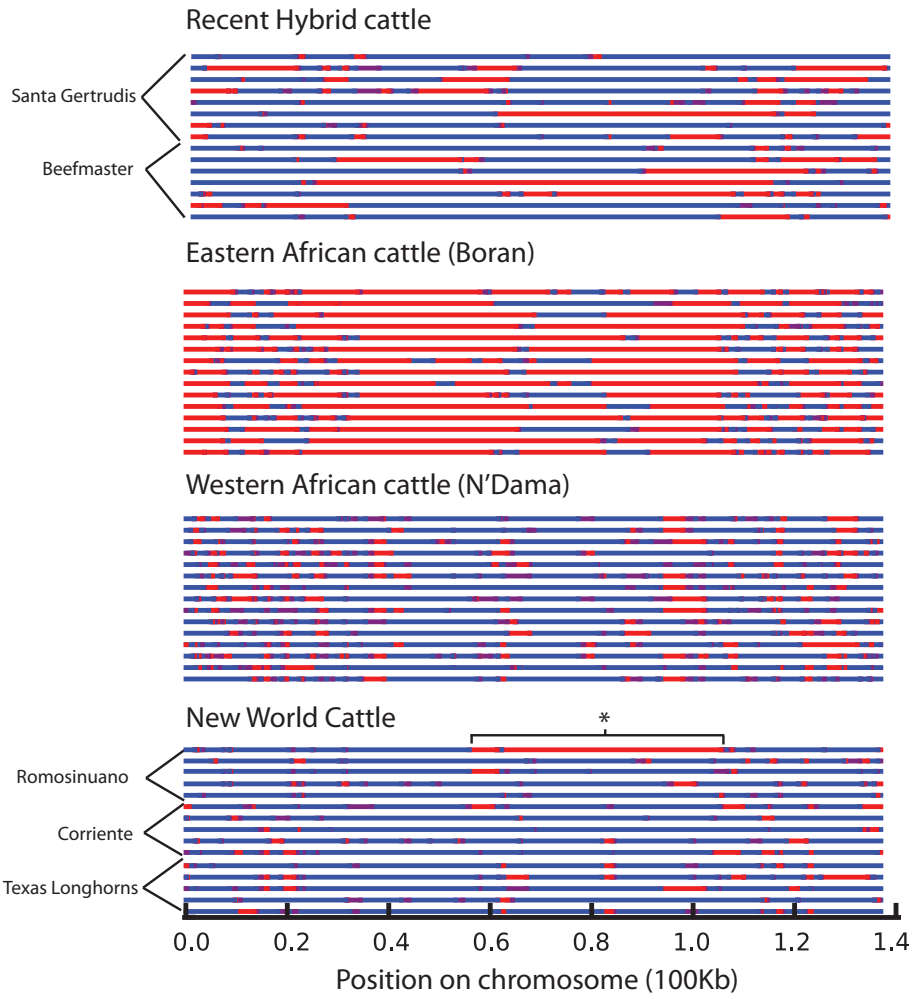


Figure 2.2. Monte-Carlo resampling of median ancestry across chromosomes.

Distributions show the expected median values of ancestry across chromosomes (assuming introgression is randomly allocated), and vertical lines represent the actual median values of introgression for each chromosome that is significantly different from the expected distribution (see Table 2.1). **A.** Recent hybrid cattle (Beefmaster and Santa Gertrudis). **B.** New World cattle (Texas Longhorns, Corriente, and Romosinuano). **C.** Western African cattle (N'Dama). **D.** Eastern African cattle (Boran).


Recent Hybrid cattle
Santa Gertrudis
Beefmaster
Eastern African cattle (Boran)
Western African cattle (N'Dama)
Evidence of recent admixture
New World Cattle
Romosinuano
Corriente
Texas Longhorns
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Position on chromosome (100Kb)

A.



B.

Figure 2.3. Admixed ancestry across chromosomes.

Ancestry of chromosomal regions estimated by ChromoPainter (Lawson *et al*. 2012) **A.** Chromosome 1 inferred from 3,150 SNP markers **B.** X chromosome with pseudoautosomal region excluded, inferred from 872 SNP markers. Each horizontal line represents a haplotype (two from each individual on chromosome 1, single haplotypes displayed for the X) and the colors represent estimated ancestry of each chromosomal region (blue indicates > 75% probability taurine; red indicates > 75% probability indicine; purple indicates intermediate probabilities). The two donor populations (taurine and indicine) were based on individuals that were estimated to have < 2% of introgressed ancestry. The figure illustrates 15 representative individuals from each of four groups of

interest (New World cattle, N'Dama, Boran, and recent hybrids). The asterisks (*) marks evidence of recent introgression in a Romosinuano individual.

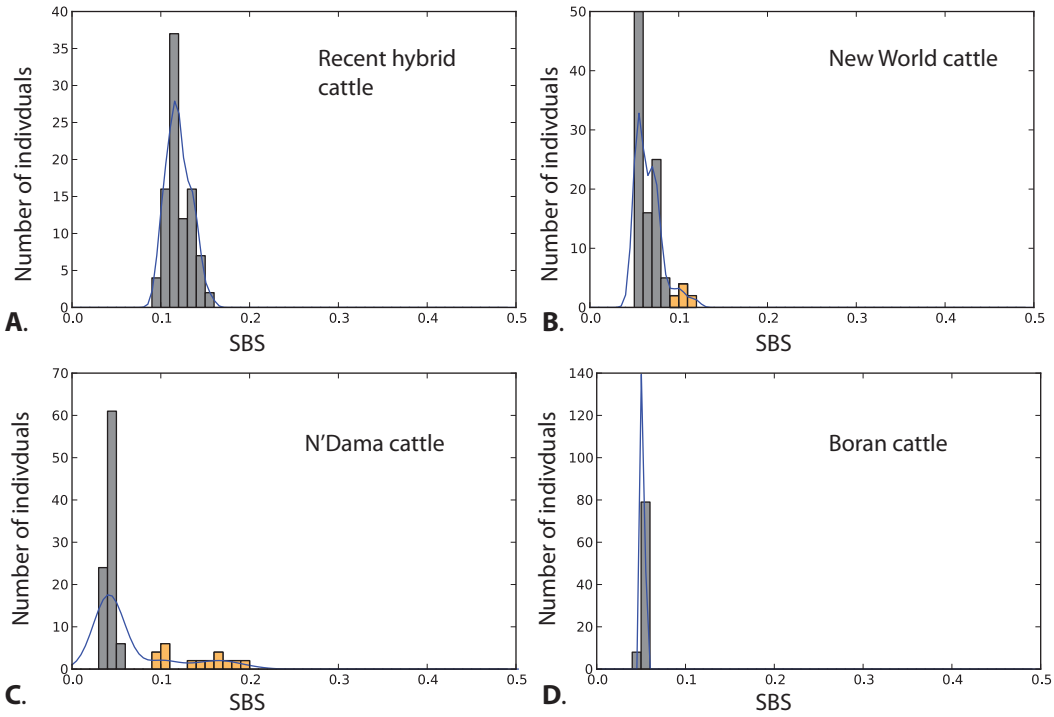


Figure 2.4. The distribution of scaled average introgressed block sizes (*SBS*) of the less common genome.

Values above 0.09 overlap with values for known recent admixed individuals, and are colored in orange. **A**. Recent hybrid cattle (Beefmaster and Santa Gertrudis). **B**. New World cattle (Texas Longhorns, Corriente, and Romosinuano). **C**. Western African cattle (N'Dama). **D**. Eastern African cattle (Boran).

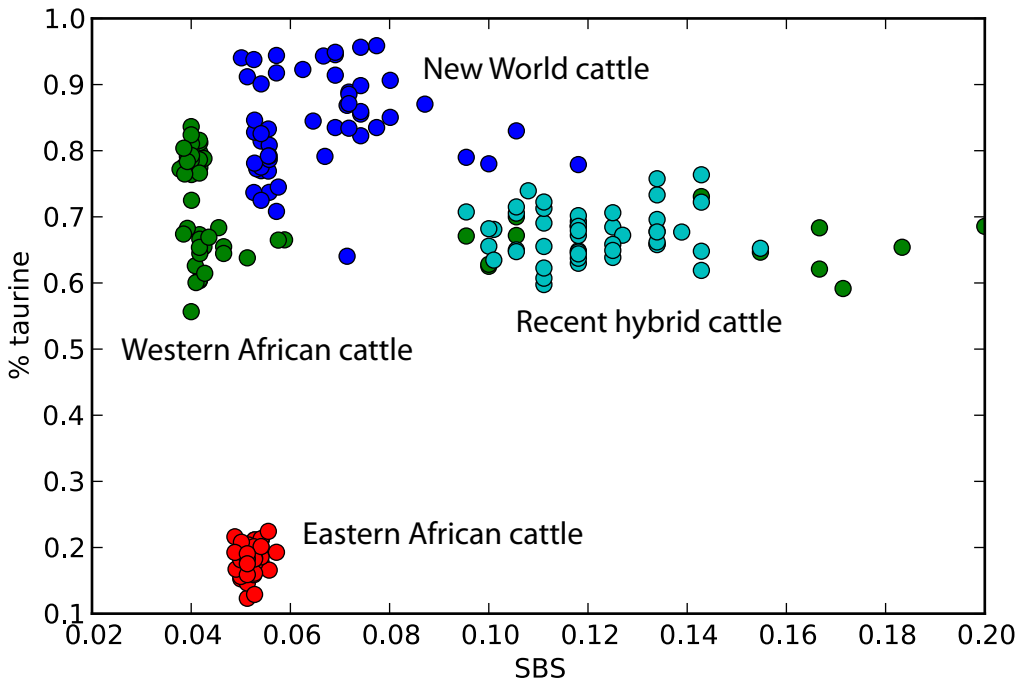


Figure 2.5. Proportion of taurine ancestry vs. *SBS* score.

Shown for individuals included in figure 2.4. Proportion taurine ancestry from McTavish *et al*. (2013)

Chapter 3: How does ascertainment bias in SNP analyses affect inferences about population history?

The availability of genomic data provides new opportunities to address population genetic and phylogeographic questions. Sequencing many loci makes it possible to understand complex biological histories. In sexual organisms, each independent locus reflects a realization of the coalescent process (Wakeley 2009). Therefore, population-level information can be inferred from the genotype of even a single individual (Green *et al*. 2010).

Although next-generation sequencing has made genomic sequence data readily available even in non-model organisms, analyzing these data requires overcoming assembly and alignment problems. Over the past few years several methods have been developed to maximize the information content of sequenced data by focusing sequencing efforts on loci likely to be informative for population genetic and phylogenetic analyses. Some methods, such as restriction-site associated DNA (RAD) sequencing (Baird *et al*. 2008) and exome sequencing (Bi *et al*. 2012), increase coverage of certain regions of the genome, which ideally results in more easily aligned sequences (Davey *et al*. 2011). In contrast, single nucleotide polymorphism (SNP) panel methods rely on resequencing a subset of the population of interest and then using this information to select polymorphic loci for additional genotyping among in a much larger pool of

individuals, often using chip-based genotyping. All these methods produce SNP data that can be used to study population genetics (Brumfield *et al*. 2003; Brito & Edwards 2009).

SNP-panel methods have the advantage of focusing sequencing effort on informative loci. By standardizing a SNP panel on a chip-based genotyping array, many loci can be sequenced inexpensively. Alignment and assembly problems are also avoided. Standardizing SNP panels, as was done for the Human Hap-Map project (International HapMap Consortium 2003), makes it straightforward for research groups to combine data and collaborate. SNP-panel analyses have been used extensively for disease research (reviewed in Manolio *et al*. 2008), Commercial direct-to-consumer applications of SNP-panel genotyping allow individuals to trace their ancestry and test for disease-associated SNPs (Ng *et al*. 2009). Novembre *et al*. (2008) used SNP loci genotyped for the POPRES project (Li *et al*. 2008) to demonstrate the genetic spatial structure of human populations in Europe. Chip-based SNP sequencing is also available for several plants and animals of scientific or agricultural importance, including dogs, mice, cattle, chickens, horses, pigs, sheep, and corn (GeneSeek 2013). Analyses of SNP genotypes are used to predict and select for trait values in animal breeding (Hayes *et al*. 2009). Chip-based SNP analyses have been used to resolve evolutionary relationships in extinct ruminants (Decker *et al*. 2009), and to understand global patterns of population structure in cattle and dogs (Vonholdt *et al*. 2010; McKay *et al*. 2008; McTavish *et al*. 2013). SNP sets are also being developed for conservation applications (Seeb *et al*. 2011) and have used to test for

hybridization between common and endangered species (e.g. Schwenke *et al*. 2006; Finger *et al*. 2009, Hohenlohe *et al*. 2011).

To discover variable SNP loci for inclusion in a SNP panel, a sample of individuals representing the taxon of interest is sequenced. This sample is called the "ascertainment group". The ascertainment group's size and composition of individuals is determined by the researchers depending on the aims of the study at hand. A panel of SNPs is then selected from the resequencing data of the ascertainment group. The selection of individuals to included in the ascertainment group can bias which SNPs are discovered and which SNP loci are included in later genotyping analyses. Ascertainment bias is of course not unique to SNP analyses. For example, in morphological analyses, variable traits are often preferentially selected over fixed traits for analysis. Furthermore, in gene sequencing studies, genes are often chosen for sequencing based on their levels of variability within a group of interest. Arnold *et al*. (2013) recently demonstrated that RAD sequencing introduces genealogical biases due to nonrandom haplotype sampling. All of these forms of ascertainment bias influence the variability of the sampled data relative to the expectations for data sampled at random from the genome.

There are two main forms of ascertainment bias associated with SNP-panel analyses: minor allele frequency (MAF) bias and subpopulation bias. MAF bias results in the over-representation of polymorphisms with high minor allele frequencies and the under-representation of polymorphisms with low minor allele frequencies. The number of individuals in the ascertainment group will influence the lower frequency limits of

SNPs included on the SNP panel. Mutations that are less common than $1/n$, where n is the number of alleles in the panel, are unlikely to be observed in the ascertainment group. Also, to minimize the impact of sequencing error, very rare mutations are often ignored. For both these reasons, the bias in minor allele frequency results in recent, and therefore rare, mutations not being included on SNP panels. If the ascertainment panel is chosen from individuals from a subpopulation or geographic region, variability in that group will be over-represented (Rosenblum & Novembre 2007). Much recent research has been devoted to describing and mitigating the impacts of minor allele frequency cut-offs in the generation of SNP panels (Nielsen 2004; Clark *et al*. 2005; Albrechtsen *et al*. 2010; McGill *et al*. 2013). Wang and Nielsen (2012) addressed phylogenetic aspects of ascertainment bias in an outgroup of the taxon of interest. However, the role of subpopulation bias in the composition of the ascertainment group, while recognized (Albrechtsen *et al*. 2010; McGill *et al*. 2013), has not been fully explored.

This study focuses on the impact of subpopulation ascertainment bias on population inference using F_{ST} values and principal components analysis (PCA). F_{ST} is a frequently used measure of population differentiation that summarizes differentiation between groups (Holsinger and Weir 2009). PCA is a statistical method for reducing the dimensionality of data. PCA was one of the earliest methods for inferring population structure from genetic data (Cavalli-Sforza 1966; Jombart *et al*. 2009). Following Novembre *et al*.'s (2008) demonstration that the first two principal component (PC) axes of human SNP data correlated strongly with spatial coordinates, PCA has been widely

applied to inferring spatial genetic structure using SNP data in humans (Reich *et al*. 2009; Bryc *et al* 2010 among others) as well as other species (e.g. cattle : McTavish *et al*. 2013; and dogs : Vonholdt *et al*. 2010). McVean (2009) described a genealogical interpretation of the principal component axes for SNP data, where the first PC axis is expected to capture the deepest coalescent split in a tree. In addition, relative PC components can be used to infer admixture between ancestral populations (McVean 2009).

To test the impacts of subpopulation biased ascertainment on inference of population histories, we simulated data based on demographic models of cattle evolution (Murray *et al*. 2010; Teasdale & Bradley 2012). We then investigated the impact of biased ascertainment of SNP loci on estimates of population genetic parameters. We compared data simulated under three demographic models to empirical data collected using a 50K marker bovine SNP chip (Matukumalli *et al*. 2009). Cattle are a useful system to investigate the impacts of ascertainment bias because there exist well-parameterized demographic models based on sequence data, which allow us to simulate large unbiased data sets.

Domesticated cattle are comprised of lineages derived from two independent domestication events, which resulted in taurine and indicine lines. Indicine cattle are common in the Indian subcontinent and taurine cattle are common in Europe; hybrids between these lines exist in Africa. These lineages share a most recent common ancestor 200,000 or more years ago (Loftus *et al*. 1994; Murray *et al*. 2010). There is a several-thousand-year history of admixture between these lineages in Africa (Freeman *et al*.

2004). The 50K SNP panel was generated by a complex ascertainment scheme including taurine, indicine and hybrid African breeds, but it is biased towards capturing polymorphism segregating in European breeds, as well as polymorphisms segregating in both taurine and indicine cattle (Matukumalli *et al*. 2009). We tested the impacts of SNP ascertainment bias on F_{ST} values and PCA. By exploring the impacts that these biases have on these methods for population genetic inference, we can better use SNP data to understand population history.

Methods

The term 'SNP' is commonly used to mean "variable site" between samples irrespective of whether a given 'SNP' is polymorphic within a population. Although Wakeley et al. (2001) coined the more accurate term 'SNP discovered locus' (SDL) to describe these single nucleotide differences that may or may not be segregating within sampled groups, this terminology has is not widely used. Here, we use 'SNP' in the broad sense of "variable site."

Empirical data

Our empirical data set was a subset of the cattle SNP data published in McTavish *et al*. (2013). We used genotypes for 25 individuals from each of three breeds representative of the three major geographic clusters of cattle: Indian (Gir), African (N'Dama), and European (Shorthorn). We included all 25 Gir samples from the published

data set. The 25 Shorthorn individuals included were a random subset of the total set of Shorthorn samples (n = 99). The 25 N'Dama individuals included were a random subset of the N'Dama samples excluding 13 individuals estimated to have admixed ancestry within the last 100 years (unpublished McTavish and Hillis, in prep; n = 46). The loci examined consisted of 47,506 SNPs genotyped using the bovine 50K SNP chip (Matukumalli *et al*. 2009). The data set had been filtered and missing data imputed as described in McTavish *et al*. (2013).

Demographic model

We simulated data under a demographic model for population structure in domesticated cattle and their wild ancestor, the aurochs (Fig. 3.1). In this model a split in the aurochs population between the ancestors of the taurine and indicine lineages occurred 280,000 years ago (Loftus *et al*. 1994; Murray *et al*. 2010). The ancestral population size (N_a) was set at 15,000 individuals (rounded from 14,127 in Murray *et al*. 2010) in our simulations. Following this split, these populations did not exchange migrants. A bottleneck reducing the population size to 150 individuals ($0.01*N_a$) occurred in the taurine lineage from 40-36 kya, followed by a population expansion to 19,212 ($1.36*N_a$; parameters from Murray *et al*. 2010). No bottleneck occurred in the indicine lineage (Teasdale & Bradley 2012). In the taurine lineage, we further simulated a population division 5,000 years ago to represent the division between European and African taurine cattle (Freeman *et al*. 2004). We used a generation time of 5 years for both aurochs and domesticated cattle (Chikhi *et al*. 2004; Murray *et al*. 2010).

We simulated data with this demographic model under three different migration conditions: *a*) no migration; *b*) low levels of symmetric gene flow between indicine and taurine lineages equivalent to 1 migrant every 10 generations (50 years) from the time of domestication, 10 kya, to present; and *c*) migration as described in *b* plus moderate levels of gene flow (1 migrant every 4 generations) from indicine lineages into the African taurine population from 3 kya to present.

Simulation software

We simulated demographic histories using the software *ms* (Hudson 2002). The *ms* program is a backwards-in-time coalescent simulator that generates samples according to a Wright-Fisher neutral model. To match our simulated data to the empirically generated data set, we simulated samples of 50 haplotypes at 47,506 SNP loci for each of the groups of European, Indian, and African cattle. We paired consecutive haplotypes to create diploid genotypes. The software *ms* uses θ ($4N_0\mu$) where N_0 is the diploid population size, and μ is the neutral mutation rate for the locus. As we were interested only in variable sites, we used a high neutral mutation rate ($3x10^{-6}$). The infinite sites assumption of the model prevents multiple mutations at the same site from occurring. The commands we used are listed in the supplemental information. We repeated the simulations 5 times to calculate confidence estimates for parameter values.

Ascertainment schemes

We subjected each of these simulated migration conditions to three SNP ascertainment treatments. We selected 1,000 SNPs under each of the following

ascertainment scenarios: (I) *Random*: SNPs were selected at random without replacement; (II) *Geographically-biased:* SNPs were selected from loci that were polymorphic in Europe, regardless of polymorphism in other groups; and (III) *Polymorphism-biased:* SNPs were selected from SNPs that were polymorphic in more than one group. SNPs that were polymorphic in all three groups were four times as likely to be selected as those only polymorphic in two groups. We performed the analyses described below on each of these subsampled data sets, and compared the parameter values to those calculated from 1,000 SNP subsamples of the empirical data set.

Population genetic parameters

We calculated number of polymorphic sites in each group in each of the full data sets. We calculated pairwise F_{ST} among all pairs of populations for the subsampled data using Weir and Cockerham's (1984) method implemented in Genepop 4.2 (Rousset 2008). We calculated the mean and standard deviation of the F_{ST} values across the 5 simulation runs.

Principal components analysis

We performed principal components analysis on each simulated data set using *smartpca* in the EIGENSTRAT software package (Patterson *et al*. 2006). We calculated the average proportion of variation explained by each PC1 and PC2 under each condition across the 5 simulation runs. We compared the major axes of variation in the PCA and the proportion of variation explained by each PC axis between data sets generated under each of these ascertainment schemes.

Results

We generated 47,506 polymorphic loci for 150 sampled chromosomes under three migration scenarios: (a) no migration; (b) low symmetric taurine-indicine gene flow since domestication; and (c) low taurine-indicine gene flow since domestication, combined with higher recent indicine to Africa gene flow (Figure 3.1, Table 3.1). We also sampled 100 gene trees under each of these demographic scenarios (Figure 3.2). The distributions of these polymorphisms across groups were very different in the simulated data and the observed data, and are compared in Figure 3.3. F_{ST} values were calculated for each pair of populations under each scenario and are reported in Table 3.2.

Principal components analysis

Projections of the first two principal components of the data under each migration scenario (a, b, c as described above) and ascertainment scheme (I, II, and III as described above) are shown in Figure 3.4. The proportion of variation accounted for by the first two principal component axes are reported in Figure 3.4 and Table 3.3. In all principal components analyses, the major axis of variation, (PC1) differentiated taurine and indicine genotypes. Although differences in migration between simulations had a minor effect on the composition of PC2, a stronger effect resulted from the type of ascertainment bias. When SNP selection was unbiased, the greater diversity in indicine cattle than taurine produced the predominant signal in PC2. Although variation within indicine lineages was the major secondary signal when SNPs were selected at random (I), under either of the ascertainment schemes selecting for polymorphism (II and III), dif-

ferentiation between European and African lineages drive variation in PC2. The proportion of variation captured by PC1, which represents the taurine–indicine split, was much greater under unbiased ascertainment than under biased ascertainment schemes (Table 3.4). ANOVA tests also indicate that differences in ascertainment scheme affect the relative PC1 score of admixed African lineages, under all migration treatments: (a) F= 39.05, P=<0.0001; (b) F=36.08, P=<0.0001; and (c) F=148.89, P=<0.0001).

Discussion

Impact of subpopulation ascertainment bias

We found that subpopulation bias in the selection of SNP loci can strongly affect inferences of population history. The type of ascertainment bias affected both the direction and extent of deviation in estimates of both F_{ST} and the population structure revealed by PCA.

As described in Albrechtsen et al (2010), selection of loci that are polymorphic within populations decreases the estimates of F_{ST} between populations. This decrease in measured F_{ST} suggests lower differentiation between populations than would be estimated from unbiased data. Across our simulated data sets, F_{ST} values were more strongly affected by differences in ascertainment scheme than by differences in migration. These results suggest that ascertainment bias may obscure information about actual population differentiation as estimated by F_{ST} values in empirical SNP data.

The impact of ascertainment bias on PCA was more surprising. The genealogical

interpretation of PCA on SNP data usually assumes that the first principal component (PC) axis captures the deepest coalescent split in the tree, and subsequent axes capture later splits (McVean 2009). Admixed populations should fall between their two ancestral populations, and the portion of ancestry inherited form each can be estimated linearly (McVean 2009). This interpretation assumes that SNP ascertainment will have a simple and predictable effect on PC projections with little influence on the relative placing of samples, except in the most extreme cases. However, in our analysis, the ascertainment scheme did impact the relative placing of simulated samples. In particular, the position of the African samples with respect to the Indian and European samples was strongly affected by ascertainment scheme (Fig. 3.4). The change in relative PC1 score is important for population genetic inference, because differences in the PC1 coordinates of the African samples have been interpreted as the difference in their proportion of indicine–taurine ancestry (McTavish *et al*. 2013). Across all migration scenarios, using SNPs preferentially selected for polymorphism in multiple groups (III) slightly overestimated indicine ancestry of African cattle in migration scenarios *a* and *c* , and selection for polymorphism in Europe (II) strongly overestimated indicine ancestry of African cattle in comparison to using randomly selected SNPs (I).These results show that care must be taken in interpreting PCA analyses of SNP data that contain ascertainment biases. Although recent analyses of human SNP data have made an effort to select polymorphisms within their population of interest (e.g. Rasmussen *et al*. 2010), subpopulation ascertainment bias is likely to be a concern as SNP panels are developed in

other species (Seeb *et al*. 2011). Subsets of SNPs that are informative about population structure within subpopulations may not be informative when applied to larger geographic samples (Paschou *et al*. 2007). The impacts of bias are likely to be even stronger when SNP panels are applied across species. Furthermore, SNPs that have been selected to differentiate between two species may mislead about relationships among populations within other species.

Application to inference of cattle population history

Murray *et al*. (2010) estimated the demographic parameters we applied to simulation using 37 kb of autosomal DNA sequenced in cattle from Europe, Africa, and the Indian subcontinent (Murray *et al*. 2010). Although these loci were selected based on their variability, this data set lacks the strong ascertainment bias of the SNP data set. The SNP panel captures many sites that are polymorphic in both taurine and indicine cattle. Figure 3.3 demonstrates that if our demographic simulations are accurate, the 50K bovine SNP panel data greatly over-represents both European and African polymorphism and shared polymorphism among groups. This SNP panel also greatly underestimates indicine diversity.

There are surprisingly high levels of shared polymorphisms maintained between indicine and taurine lineages across 280 kya of divergence. This prevalence of deep coalescence events is particularly surprising given the estimates from mtDNA of extremely narrow bottlenecks associated with domestication (Bollongino *et al*. 2012). Using 50K SNP data, MacEahern *et al*. (2009a) found that approximately 10% of all

polymorphisms that segregate in two taurine breeds (Angus and Holstein) also segregate in at least one of Bison, Yak, or Banteng. Matukumalli *et al*. (2009) also found that 1–5% of SNPs in the 50K panel were polymorphic in other *Bos* species, and some were variable in multiple outgroup species. Taken together, these results suggest that these SNP data may be capturing sites with unusual evolutionary histories, such as loci that reflect long-term balancing selection. Indeed, ascertainment bias in the SNP data likely had a strong affect on MacEahern *et al*.'s (2009a) estimate for the effective population size of the ancestor of cattle, the aurochs, at 90,000 individuals. Using autosomal data, which are far less likely to be biased towards maintained polymorphisms, Murray *et al*. (2010) estimated an effective population size for aurochsen of around 15,000 individuals. This estimate is much more consistent with estimates of N_e in aurochsen based on ancient mtDNA (~2,000-8,000; Mona *et al*. 2010). Nonetheless, even in autosomal data, there are sufficient shared polymorphisms among taurine and indicine lineages that the best-fit model requires gene flow between the lineages at low levels, strong balancing selection on segregating sites, very large population sizes, or some combination of these factors (MacEachern *et al*. 2009b; Murray *et al*. 2010).

By comparing the simulation results with the estimates based on empirical data from cattle, we can assess the effects of types of ascertainment bias on estimates of population history. We found that estimates of F_{ST} between European and Indian cattle were similar between empirical data (0.42, Table 3.2) and simulated ascertainment schemes II and III (0.37-0.64, Table 3.2). However, estimates of F_{ST} between African and

Indian cattle were lower in empirical data (0.40, Table 3.2) than in any of our simulations (0.47-0.78). In addition, the estimate of F_{ST} between European and African cattle in the empirical data (0.28) was 3–4 times higher than that in simulated data (0.06–0.08). Taken together, these results suggest that indicine gene flow into Africa occurred at a higher rate than we assumed in our demographic model. Comparing the PCAs (Fig. 3.4) for the empirical data and simulated data, the possible role of ascertainment bias is further apparent. Selection for European polymorphism under migration scenario *c* resulted in PC estimates very similar to those observed in empirical data. However, even given the biases observed in estimation of admixture in African cattle, this comparison also suggests higher indicine gene flow into Africa than has been estimated from prior studies or than was assumed in our demographic model.

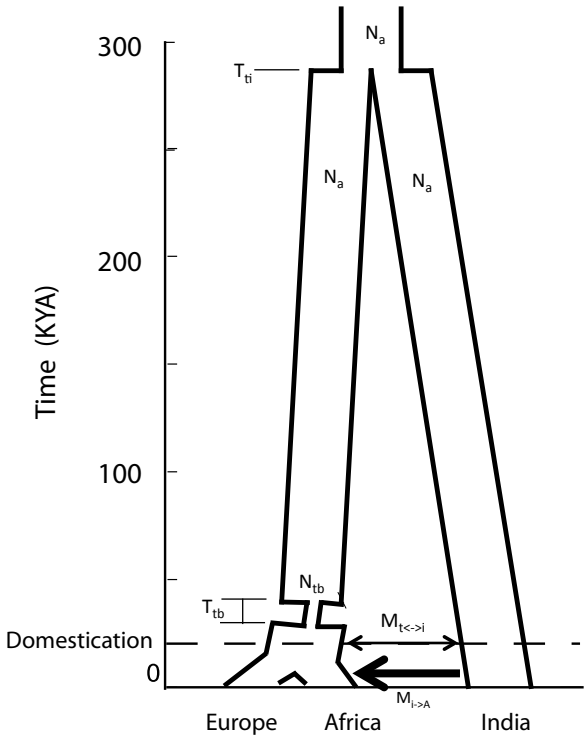


Figure 3.1. Demographic model for simulations.

Parameter values are described in Table 3.1.

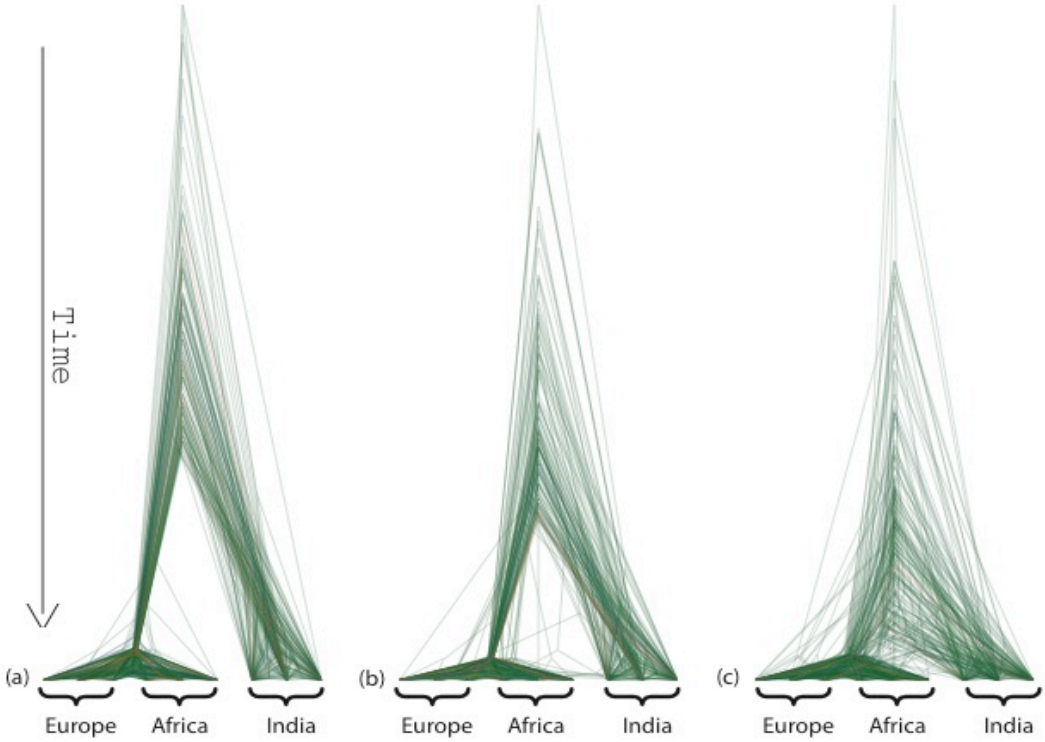


Figure 3.2. Gene trees generated according to the demographic models under each of three migration scenarios.

Gene trees are plotted atop one another so that patterns of variation among loci are visible. (a) no migration; (b) low taurine-indicine gene flow since domestication; (c) low taurine-indicine gene flow since domestication, combined with higher recent indicine to African gene flow. Figure created using Densitree (Bouckaert 2010).

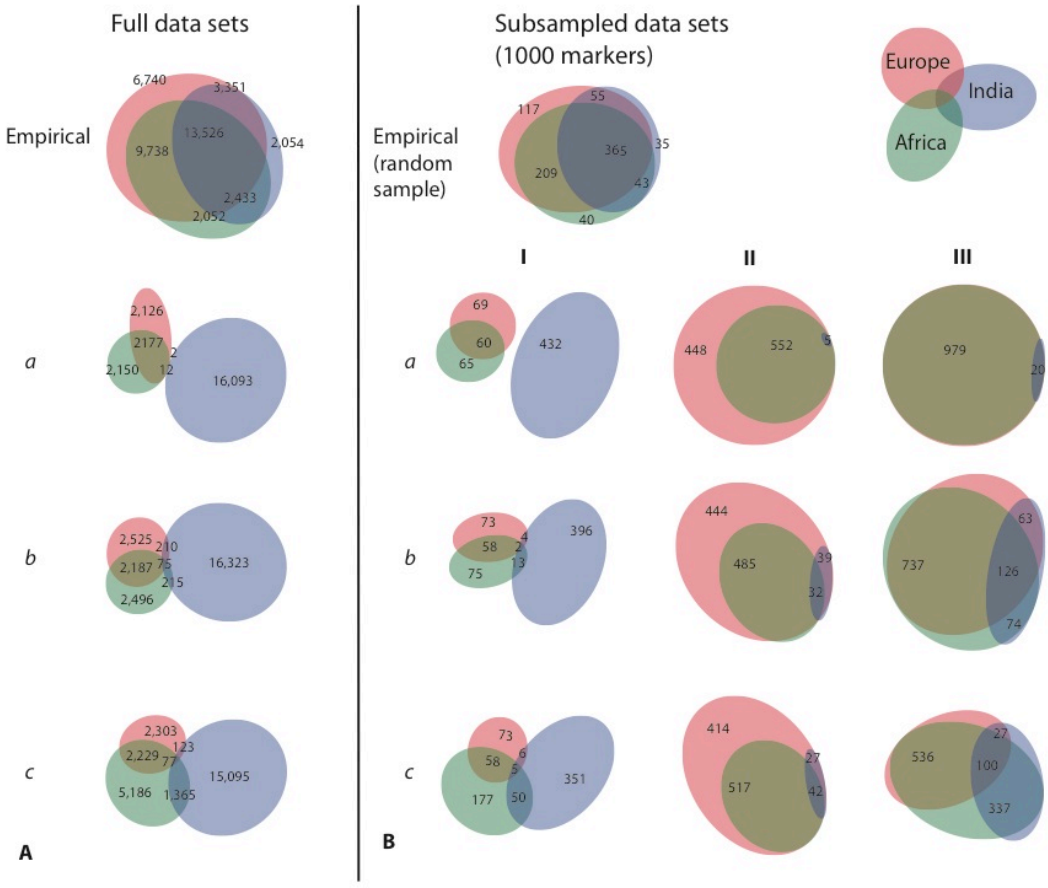


Figure 3.3. Venn diagrams demonstrating the counts of polymorphisms segregating within each continental group.

Area of ellipses and areas of overlap are approximately proportional to number of sites in those categories. Number of sites in each category is labeled if there were sites in that category. **A)** Full data sets for the empirical data and the three simulated data sets. All data sets consisted of 47,506 polymorphic sites but some sites were fixed differences among populations and are not shown here. (*a*) no migration (*b*) low taurine-indicine gene flow since domestication (*c*) low taurine-indicine gene flow since domestication, plus higher recent indicine-> Africa gene flow **B**) 1,000 marker subsets of the empirical data set and the simulated data sets; migration condition *a*, *b, c,* as described above, and ascertainment bias conditions (I) random samples (II) sampled loci that were polymorphic within Europe (III) sampled loci that were polymorphic in two or more subpopulations. Figure made using EulerAPE (Micallef and Rodgers, 2012).

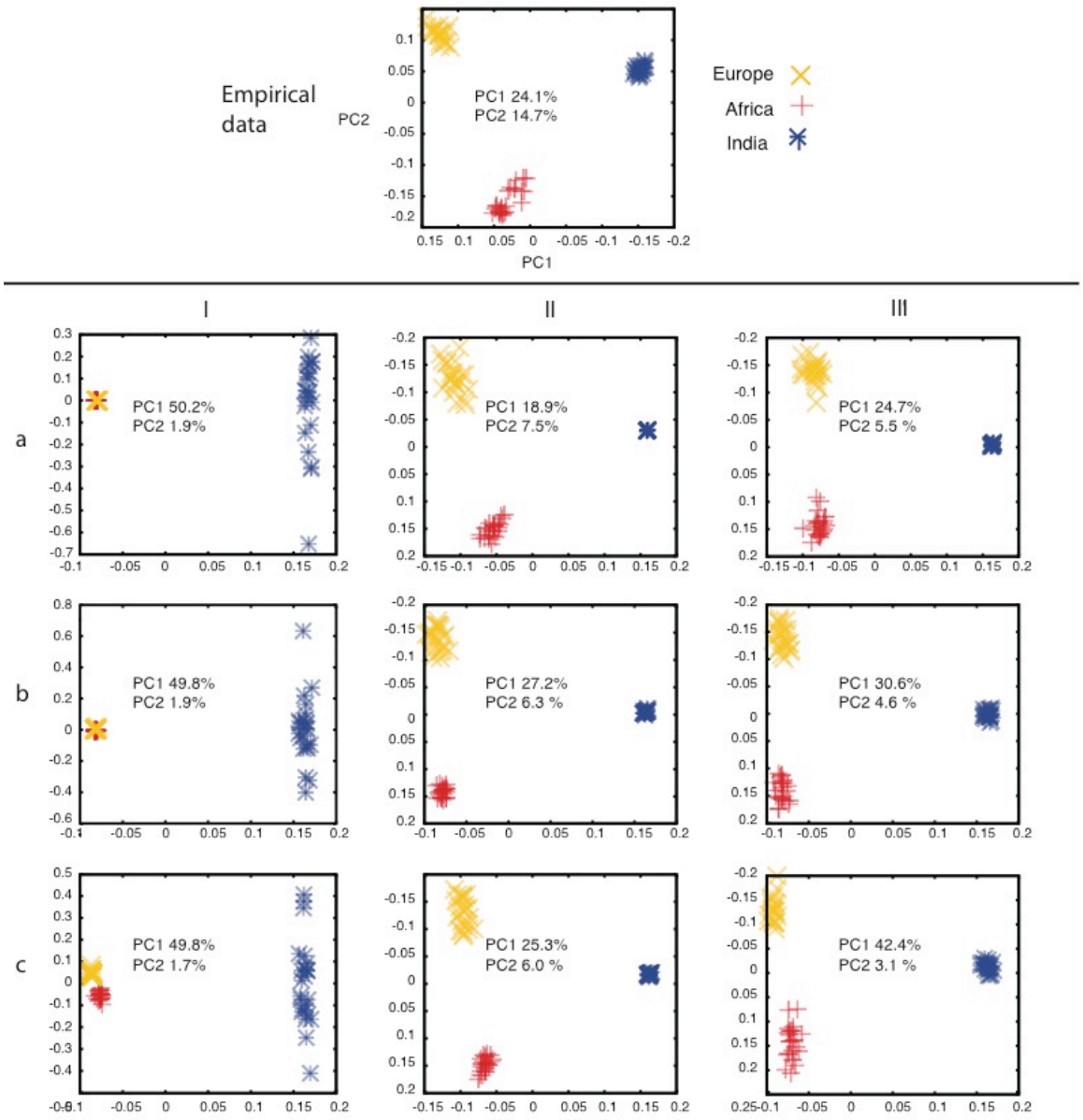


Figure 3.4. Principal components analysis performed on 1,000 marker subsets of simulated data under 3 migrations schemes and three ascertainment bias conditions, and the empirical data.

(a) No migration; (b) low taurine-indicine gene flow since domestication; and (c) low taurine–indicine gene flow since domestication, combined with higher recent indicine to Africa gene flow. Ascertainment bias conditions: (I) random samples; (II) preferential for

loci that were polymorphic within Europe; and (III) preferential for loci that were polymorphic in two or more subpopulations. Proportion of variation in the data accounted for by the first two PC axes labeled on figure.

Table 3.1. Parameter values for the three demographic models simulated.

Parameter values adapted from Murray *et al*. (2010). Values for simulations (b) and (c) were the same as for (a) unless specified.

Variable	Description	a	b	c
	Generation time	5 years	-	-
N_a	Ancestral population size	15,000	-	-
N_{tE}	Current European taurine population size	7,500	-	-
N_{tA}	Current African taurine population size	7,500	-	-
N_a	Current indicine population size (= ancestral population size)	15,000	-	-
T_{AE}	Time of African-European divergence	5 kya (1,000 generations)	-	-
T_{tb}	Timing of bottleneck in taurine cattle	40-36 kya	-	-
N_{tb}	Size of bottleneck in taurine cattle	150 (0.01*Na)	-	-
T_{ti}	Time of indicine-taurine divergence	280 kya (56,000 generations)	-	-
$M_{t<->i}$	Number of migrants between taurine-indicine lineages per generation since domestication 10 kya (symmetric).	0	0.1	0.1
$M_{i->A}$	Number of migrants from indicine lineages into Africa per generation for the past 3 kya (asymmetric)	0	0	1

Table 3.2. Mean multilocus F_{ST} values (± standard deviation) calculated for each pair of populations under each ascertainment scheme and migration scenario using *Genepop* (Rousset 2008).

(a) No migration; (b) low taurine-indicine gene flow since domestication; (c) low taurine-indicine gene flow since domestication, combined with higher recent indicine to Africa gene flow. Ascertainment schemes: (I) random; (II) biased towards polymorphism in Europe; and (III) biased towards polymorphism in multiple lineages.

		Eur	Afr
Empirical data	Afr	0.28 ± (0.015)	
	Indi	0.42 ± (0.007)	0.40 ± (0.011)

80		I		II		III	
		Eur	Afr	Eur	Afr	Eur	Afr
a	Afr	0.08 ± (0.008)		0.08 ± (0.005)		0.07 ± (0.007)	
	Indi	0.78 ± (0.015)	0.78 ± (0.015)	0.37 ± (0.012)	0.47 ± (0.014)	0.44 ± (0.004)	0.44 ± (0.007)
b	Afr	0.07 ± (0.014)		0.08 ± (0.005)		0.07 ± (0.001)	
	Indi	0.77 ± (0.010)	0.77 ± (0.010)	0.56 ± (0.038)	0.68 ± (0.038)	0.53 ± (0.033)	0.54 ± (0.036)
c	Afr	0.08 ± (0.007)		0.07 ± (0.005)		0.06 ± (0.004)	
	Indi	0.78 ± (0.009)	0.74 ± (0.009)	0.52 ± (0.028)	0.62 ± (0.026)	0.64 ± (0.027)	0.58 ± (0.030)

Table 3.3. Commands used for simulations in ms

a ./ms 150 100000 -t 0.18 -I 3 50.0 50.0 50.0 0 -en 0 1 0.5 -en 0 2 0.5 -en 0 3 1 -ej 0.023333 1 2 -ej 0.933333 3 2 -en 0.12 2 0.01 -en 0.133333 2 1.0

b ./ms 150 100000 -t 0.18 -I 3 50.0 50.0 50.0 0.4 -en 0 1 0.5 -en 0 2 0.5 -en 0 3 1 -ej 0.023333 1 2 -ej 0.933333 3 2 -en 0.12 2 0.01 -en 0.133333 2 1.0 -eM 0.033333 0

c ./ms 150 100000 -t 0.18 -I 3 50.0 50.0 50.0 0 -en 0 1 0.5 -en 0 2 0.5 -en 0 3 1 -ej 0.023333 1 2 -ej 0.933333 3 2 -en 0.12 2 0.01 -en 0.133333 2 1.0 -em 0 2 3 4 -eM 0.01 0.4 -eM 0.033333 0

Table 3.4. Mean proportion of variation captured by PC1 and PC2 ± (standard deviation).

Values calculated for 1,000 SNP data subsets for empirical data and under each migration scenario (a) no migration (b) low taurine-indicine gene flow since domestication (c) low taurine-indicine gene flow since domestication, plus higher recent indicine to Africa gene flow, and each ascertainment scheme (I) random, (II) biased towards polymorphism in Europe, (III) biased towards polymorphism in multiple lineages.

Empirical data	PC1	0.25 ± (0.005)		
	PC2	0.15 ± (0.005)		
		I	**II**	**III**
a	PC1	0.53 ± (0.018)	0.17 ± (0.012)	0.24 ± (0.003)
	PC2	0.02 ± (0.001)	0.07 ± (0.005)	0.06 ± (0.003)
b	PC1	0.52 ± (0.015)	0.32 ± (0.025)	0.36 ± (0.028)
	PC2	0.02 ± (0.001)	0.06 ± (0.003)	0.04 ± (0.003)
c	PC1	0.52 ± (0.015)	0.27 ± (0.011)	0.48 ± (0.029)
	PC2	0.02 ± (0.001)	0.06 ± (0.001)	0.03 ± (0.003)

References

Achilli A, Bonfiglio S, Olivieri A, Malusà A, Pala M *et al.* (2009) The multifaceted origin of taurine cattle reflected by the mitochondrial genome. *PLoS ONE* 4:e5753.

Achilli A, Olivieri A, Soares P, Lancioni H, Kashani BH *et al.* (2012) Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. *Proceedings of the National Academy of Sciences of the United States of America* 109:2449–2454.

Albrechtsen A, Nielsen F, Nielsen R (2010) Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution* 27:2534–2547

Allendorf FW, Leary RF, Spruell P, Wenburg JK (2001) The problems with hybrids: setting conservation guidelines. *Trends in Ecology and Evolution* 16:613–622.

Anderung C, Bouwman A, Persson P, Carretero JM, Ortega AI *et al.* (2005) Prehistoric contacts over the Straits of Gibraltar indicated by genetic analysis of Iberian Bronze Age cattle. *Proceedings of the National Academy of Sciences of the United States of America* 102:8431–8435.

Arias JA, Keehan M, Fisher P, Coppieters W, Spelman R (2009) A high density linkage map of the bovine genome. *BMC Genetics* 10:18.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology* Early View

Bailey JF, Richards MB, Macaulay VA, Colson IB, James, IT *et al.* (1996) Ancient DNA suggests a recent expansion of European cattle from a diverse wild progenitor species. *Proceedings of the Royal Society B: Biological Sciences* 1376:1467–1473.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3:e3376.

Baird S (1995) A simulation study of multilocus clines. *Evolution* 49:1038–1045.

Barragy TJ (2003) Gathering Texas Gold (Cayo Del Grullo Press, Cayo del Grullo, TX).

Beja-Pereira A, Caramelli D, Lalueza-Fox C, Vernesi C, Ferrand N *et al.* (2006) The origin of European cattle: evidence from modern and ancient DNA. *Proceedings of the National Academy of Sciences of the United States of America* 103:8113–8118.

Bi K, Vanderpool D, Singhal S, Linderoth T, Moritz C, Good JM (2012) Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics* 13:403

Bollongino R, Burger J, Powell A, Mashkour M, Vigne JD, Thomas, MG (2012) Modern taurine cattle descended from small number of near-eastern founders. *Molecular Biology and Evolution,* 29:2101–2104.

Boos DD, Brownie C (2004) Comparing variances and other measures of dispersion. *Statistical Science* 19:571–578.
Bouckaert RR (2010) DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* 26:1372–1373.
The Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC, (2009) The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science* 324:522–528.
The Bovine HapMap Consortium (2009) Genomewide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324:528–532.
Bowden R, MacFie TS, Myers S, Hellenthal G, Nerrienet E *et al.* (2012) Genomic tools for evolution and conservation in the chimpanzee: *Pan troglodytes ellioti* is a genetically distinct population. *PLoS Genetics* 8:e1002504.
Bradley DG, MacHugh DE, Cunningham P, Loftus RT (1996) Mitochondrial diversity and the origins of African and European cattle. *Proceedings of the National Academy of Sciences of the United States of America* 93:5131–5135.
Brito PH, Edwards SV (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica* 135:439–455.
Brumfield R, Beerli P, Nickerson D (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology and Evolution* 18:249–256.
Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL *et al.* (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences of the United States of America* 107:786–791.
Canavez FC, Luche DD, Stothard P, Leite KRM, Sousa-Canavez JM, *et al.* (2012) Genome sequence and assembly of Bos indicus. *Journal of Heredity* 103:342–348.
Cartwright TC (1980) Prognosis of zebu cattle: Research and application. *Journal of Animal Science* 50:1221–1226.
Cavalli-Sforza LL (1966) Population structure and human evolution. *Proceedings of the Royal Society B: Biological Sciences* 164, 362–379.
Central Intelligence Agency (2008) *The World Factbook*. (Central Intelligence Agency, Washington). [Available at https://www.cia.gov/library/publications/the-world-factbook/. Accessed October 5, 2011.]
Chikhi L, Goossens B, Treanor A, Bruford, MW (2004) Population genetic structure of and inbreeding in an insular cattle breed, the Jersey, and its implications for genetic resource management. *Heredity* 92:396–401.
Clark A, Hubisz M, Bustamante C, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Research, 15, 1496–1502.
Clutton-Brock J (1999) *A natural history of domesticated mammals* (Cambridge Univ Pr, Cambridge, UK).

Cymbron T, Loftus R, Malheiro M, Bradley D (1999) Mitochondrial sequence variation suggests an African influence in Portuguese cattle. *Proceedings of the Royal Society B: Biological Sciences* 63:1467-1473.
Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* 12:499–510.
Davis SJM (2008) Zooarchaeological evidence for Moslem and Christian improvements of sheep and cattle in Portugal. *Journal of Archaeological Science* 35:991–1010.
Decker JE, Pires JC, Conant GC, McKay SD, Heaton MP e*t al.* (2009) Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proceedings of the National Academy of Sciences of the United States of America* 106:18644–18649.
Demeke S, Neser F, Schoeman SJ (2003) Early growth performance of *Bos taurus* x *Bos indicus* cattle crosses in Ethiopia: evaluation of different crossbreeding models. You have full text access to this content. *Journal of Animal Breeding and Genetics* 120:39–50.
Dobie JF (1941) *The Longhorns* (University of Texas Press, Austin, TX). 1980. Reprint.
Earl DA, Vonholdt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4:359-361.
Edwards S, Bensch S (2009) Looking forwards or looking backwards in avian phylogeography? A comment on Zink and Barrowclough 2008. *Molecular Ecology* 18:2930–2933.
Edwards SV, Kingan SB, Calkins JD, Balakrishnan CN, Jennings WB, *et al*. (2005) Speciation in birds: Genes, geography, and sexual selection. *Proceedings of the National Academy of Sciences of the United States of America* 102:6550–6557.
Efron B (1981) Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika* 68:589–599.
Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* 14:2611–2620.
Figueroa J, Chieves L, Johnson G, Buening G (1992) Detection of Babesia bigemina-infected carriers by polymerase chain reaction amplification. *Journal of Clinical Microbiology* 30:2576.
Finger AJ, Stephens MR, Clipperton NW, May B (2009) Six diagnostic single nucleotide polymorphism markers for detecting introgression between cutthroat and rainbow trouts. *Molecular Ecology Resources* 9:759–763.
Fisher RA (1954) A fuller theory of "junctions" in inbreeding. *Heredity* 8:187–197.
Francois O, Currat M, Ray N, Han E, Excoffier L, Novembre J (2010) Principal component analysis under population genetic models of range expansion and admixture. *Molecular Biology and Evolution* 27:1257–1268.

Freeman A, Meghen C, MacHugh D, Loftus R, Achukwi M, *et al.* (2004) Admixture and diversity in West African cattle populations. *Molecular Ecology* 13:3477–3487.
Frisch J, Drinkwater R, Harrison B (1997) Classification of the southern African sanga and East African shorthorned zebu. *Animal Genetics* 28:77–83.
Frisch J, Vercoe J (1977) Food intake, eating rate, weight gains, metabolic rate and efficiency of feed utilization in *Bos taurus* and *Bos indicus* crossbred cattle. *Animal Production* 25:343–358.
Gautier M, Naves M (2011) Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Molecular Ecology* 20:3128–3143.
GeneSeek Genotyping Services (2013) Illumina Genotyping Services. (accessed March 25, 2013) [http://www.neogen.com/geneseek/SNP_Illumina.html].
George JE, Davey RB, Pound JM (2002) Introduced ticks and tick-borne diseases: the threat and approaches to eradication. *Veterinary Clinics of North America: Food Animal Practice* 18:401-416.
Ginja C, M. C. T. Penedo MCT, L. Melucci L, J. Quiroz J, Martínez López OR, Revidatti MA *et al.* (2010) Origins and genetic diversity of New World Creole cattle: inferences from mitochondrial and Y chromosome polymorphisms. *Animal Genetics* 41:128–141.
Giovambattista G, Ripoli MV, De Luca JC, Mirol PM, Lirón JP, Dulout FN (2000) Male-mediated introgression of *Bos indicus* genes into Argentine and Bolivian Creole cattle breeds. *Animal Genetics* 31:302–305.
Green RE, Krause J, Briggs AW, Maricic T, Stenzel U *et al.* (2010) A draft sequence of the Neandertal genome. *Science* 328:710–722.
Grigson C (1991) An African origin for African cattle?—some archaeological evidence. *African Archaeological Review* 9:119–144.
Haber M, Gauguier D, Youhanna S, Patterson N, Moorjani P *et al*. (2013) Genome-wide diversity in the Levant reveals recent structuring by culture. *PLoS Genetics* 9:e1003316.
Hanotte O, Bradley D, Ochieng J, Verjee Y, Hill E (2002) African pastoralism: genetic imprints of origins and migrations. *Science* 296:336-339.
Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science* 92:433–443.
Heaton MP, Chitko-McKown CG, Grosse WM, Keele JW, Keen JE, Laegreid WW (2001) Interleukin-8 haplotype structure from nucleotide sequence variation in commercial populations of U.S. beef cattle. *Mammalian Genome* 12:219–226.
Hiendleder S, Lewalski H, Janke A (2008) Complete mitochondrial genomes of *Bos taurus* and *Bos indicus* provide new insights into intra-species variation, taxonomy and domestication. *Cytogenetic and Genome Research* 120:150–156.
Hohenlohe PA, Amish SJ, Catchen JM, Allendorf RW, Luikart, G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing

hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources* 11:117–122.
Holsinger KE, Weir BS (2009) Fundamental concepts in genetics: Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics* 10:639–650.
Hoyt AM (1982) History of Texas Longhorns. *Texas Longhorn Journal* 1982:1–48. Available at: http://doublehelixranch.com/History.html [Accessed February 9, 2012].
Hudson R (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337.
International HapMap Consortium (2003) The International HapMap Project. *Nature* 426:789–796.
Jombart T, Pontier D, Dufour A-B (2009) Genetic markers in the playground of multivariate analysis. *Heredity* 102:330–341.
Jones E, Oliphant T, Peterson P *et al.* (2001) SciPy: Open Source Scientific Tools for Python. [http://www.scipy.org].
Kantanen J, Olsaker, I, Holm, LE, Lien S, Vilkki J, *et al.* (1999) Temporal changes in genetic variation of North European cattle breeds. *Animal Genetics* 30:16–28.
Kawahara-Miki R, Tsuda K, Shiwa Y, Arai-Kichise Y, Matsumoto T *et al.* (2011) Whole-genome resequencing shows numerous genes with nonsynonymous SNPs in the Japanese native cattle Kuchinoshima-Ushi. *BMC Genomics* 12:103.
Kidd K, Cavalli-Sforza L (1974) The role of genetic drift in the differentiation of Icelandic and Norwegian cattle. *Evolution* 28:381–395.
Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR *et al.* (2012) Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biology* 10:e1001258.
Kitano J, Ross JA, Mori S, Kume M, Jones FC *et al.* (2009) A role for a neo-sex chromosome in stickleback speciation. *Nature* 461:1079–1083.
Kuehn LA, Keele JW, Bennett GL, McDaneld TG, Smith TPL *et al.* (2011) Predicting breed composition using breed frequencies of 50,000 markers from the US Meat Animal Research Center 2,000 Bull Project. *Journal of Animal Science* 89:1742–1750.
Kuhner M, Beerli P, Yamato J, Felsenstein J (2000) Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* 156: 439–447.
Larson G, Karlsson EL, Perri A, Webster MT, Hoe SYW *et al.* (2012) Rethinking dog domestication by integrating genetics, archeology, and biogeography. *Proceedings of the National Academy of Sciences of the United States of America* 109:8878–8883.
Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *PLoS Genetics* 8:e1002453.
Levene H (1960) Robust tests for equality of variances. In *Contributions to Probability and Statistics* (I. Olkin, ed.) 278–292. Stanford Univ. Press, Stanford, CA.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165: 2213–2233.

Loftus R, MacHugh D, Bradley D, Sharp P, Cunningham P (1994) Evidence for two independent domestications of cattle. *Proceedings of the National Academy of Sciences of the United States of America* 91:2757.

Loftus RT, Ertugrul O, Harba AH, El-Barody MAA, Machugh DE *et al.* (1999) A microsatellite survey of cattle from a centre of origin: the Near East. *Molecular Ecology* 8:2015–2022.

MacEachern S, Hayes B, McEwan J, Goddard M (2009a) An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (*Bos taurus*) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. *BMC Genomics* 10:181.

MacEachern S, McEwan J, Goddard M (2009b) Phylogenetic reconstruction and the identification of ancient polymorphism in the Bovini tribe (Bovidae, Bovinae). *BMC Genomics* 10:177.

MacHugh DE, Shriver MD, Loftus RT, Cunningham P, Bradley DG (1997) Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics* 146:1071–1086.

Magee D, Meghen, C Harrison S, Troy CS, Cymbron T *et al.* (2002) A partial African ancestry for the Creole cattle populations of the Caribbean. *Journal of Heredity* 93:429-432.

Manolio TA, Brooks LD, Collins FS (2008) A HapMap harvest of insights into the genetics of common disease. *Journal of Clinical Investigation* 118:1590–1605.

Martínez AM, Gama LT, Cañón J, Ginja C, Delgado JV *et al.* (2012) Genetic footprints of Iberian cattle in America 500 years after the arrival of Columbus. *PLoS ONE* 7:e49066.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF *et al*. (2009) Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE* 4:e5350.

McGill JR, Walkup EA, Kuhner MK (2013) Correcting coalescent analyses for panel-based SNP ascertainment. *Genetics* Early Online.

McKay SD, Schnabel R, Murdoch B, Matukumalli LK, Aerts J *et al.* (2008) An assessment of population structure in eight breeds of cattle using a whole genome SNP panel. *BMC Genetics* 9:37.

McTavish EJ, Decker JE, Schnabel RD, Taylor JF, Hillis DM (2013) New World cattle show ancestry from multiple independent domestication events. *Proceedings of*

the National Academy of Sciences of the United States of America 110:E1398-E1406.
McVean G (2009) A genealogical interpretation of principal components analysis. *PLoS Genetics* 5:e1000686.
Micallef L, Rodgers, P (2012) Poster: Drawing area-proportional Venn-3 diagrams using ellipses. In: 12th Annual Grace Hopper Celebration of Women in Computing, ACM Student Research Competition and Poster Session, Baltimore, MD, USA. [Software at http://www.eulerdiagrams.org/eulerAPE].
Mirol PM, Giovambattista G, Lirón JP, Dulout FN (2003) African and European mitochondrial haplotypes in South American Creole cattle. *Heredity* 91:248–254.
Mona S, Catalano G, Lari M, Larson G, Boscato P *et al.* (2010) Population dynamic of the extinct European aurochs: genetic evidence of a north-south differentiation pattern and no evidence of post-glacial expansion. *BMC Evolutionary Biology* 10:83.
Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L *et al.* (2011) The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genetics* 7:e1001373.
Murray C, Huerta-Sanchez E, Casey F, Bradley DG (2010) Cattle demographic history modelled from autosomal sequence variation. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365:2531–2539.
Murray M, Trail J, Davis C, Black S (1984) Genetic resistance to African trypanosomiasis. *The Journal of Infectious Diseases* 149:311–319.
Ng PC, Murray SS, Levy S, Venter JC (2009) An agenda for personalized medicine. *Nature* 461:724–726.
Nielsen R (2004) Population genetic analysis of ascertained SNP data. *Human Genomics* 1:218–224.
Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR *et al*. (2008) Genes mirror geography within Europe. *Nature* 456:98–101.
Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics* 40:646–649.
Parks DH, Porter M, Churcher S, Wang S, Blouin C *et al.* (2009) GenGIS: A geospatial information system for genomic data. *Genome Research* 19:1896–1904.
Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W *et al.* (2007) PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genetics* 3:e160.
Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N *et al.* (2012) Ancient admixture in human history. *Genetics* 192:1065–1093.
Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics* 2:e190.
Pei Y-F, Li J, Zhang L, Papasian CJ, Deng H-W (2008) Analyses and comparison of accuracy of different genotype imputation methods. *PLoS ONE* 3:e3551.

Perkins D (1969) Fauna of Çatal Hüyük: evidence for early cattle domestication in Anatolia. *Science* 164:177.
Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38:904–909.
Price AL, Tandon A, Patterson NJ, Barnes KC, Rafaels N *et al.* (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics* 5:e1000519.
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.
R Core Team (2010) R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria).
Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen S *et al.* (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* 463:757–762.
Reich D, Green RE, Kircher M, Krause J, Patterson N *et al.* (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468:1053–1060.
Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. *Nature* 461:489–494.
Rhoad AO (1949) The Santa Gertrudis breed: The genesis and the genetics of a new breed of beef cattle. *Journal of Heredity* 40:115.
Riely A (2011) The grass-fed cattle-ranching niche in Texas. *Geography review* 101:261–268.
Rieseberg LH, Baird SJ, Gardner KA (2000) Hybridization, introgression, and linkage evolution. *Plant Molecular Biology* 42:205–224.
Rosenberg NA (2003) distruct: a program for the graphical display of population structure. *Molecular Ecology Notes* 4:137–138.
Rosenblum EB, Novembre J (2007) Ascertainment bias in spatially structured populations: A case study in the Eastern Fence Lizard. *Journal of Heredity* 98:331–336.
Rouse JE (1977) *The Criollo: Spanish Cattle in the Americas* (University of Oklahoma Press, Norman, OK).
Rousset F (2008) Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* 8:103–106.
Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* 78:629–644.
Schwenke PL, Rhydderch JG, Ford MJ, Marshall AR, Park LK (2006) Forensic identification of endangered Chinook Salmon (*Oncorhynchus tshawytscha*) using a multilocus SNP assay. *Conservation Genetics* 7:983–989.
Seeb JE, Carvalho G, Hauser L, Naish K, Roberts S, Seeb LW (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources* 11:1–8.

Slatkin M (1987) Gene flow and the geographic structure of natural populations. *Science* 236:787–792.
Teasdale MD, Bradley DG (2012) The origins of cattle. In *Bovine Genomics.* (ed. J Womack) Ames: Wiley-Blackwell, pp 1–10.
Tracy CA, Widom H (1994) Level-spacing distributions and the Airy kernel. *Communications* in Mathematical Physics 159:151–174.
Ungerer MC, Baird SJ, Pan J, Rieseberg LH (1998) Rapid hybrid speciation in wild sunflowers. *Proceedings of the National Academy of Sciences of the United States of America* 95:11757–11762.
Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD *et al.* (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods* 5:247–252.
Vila M, Vidal-Romaní JR, Björklund M (2005) The importance of time scale and multiple refugia: Incipient speciation and admixture of lineages in the butterfly *Erebia triaria* (Nymphalidae). *Molecular Phylogenetics and Evolution* 36:249–260.
Villa-Angulo R, Matukumalli LK, Gill CA, Choi J, Van Tassell CP, Grefenstette JJ (2009) High-resolution haplotype block structure in the cattle genome. *BMC Genetics* 10:19.
Vonholdt BM, Pollinger JP, Lohmueller KE, Han E, Parker HG *et al.* (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* 464:898–902.
Wakeley J (2009) *Coalescent Theory*. Roberts & Company, Greenwood Village, Colorado.
Wakeley J, Nielsen R, Liu-Cordero S, Ardlie K (2001) The discovery of single-nucleotide polymorphisms--and inferences about human demographic history. *The American Journal of Human Genetics* 69:1332–1347.
Wang Y, Nielsen R (2012) Estimating population divergence time and phylogeny from single-nucleotide polymorphisms data with outgroup ascertainment bias. *Molecular Ecology* 21:974–986.
Warwick EJ (1958) Fifty years of progress in breeding beef cattle. *Journal of Animal Science* 17:922–943.
Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370.
Womack JE (2005) Advances in livestock genomics: opening the barn door. *Genome Research* 15:1699–1705.
Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC *et al.* (2009) A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biology* 10:R42.