Unravelling local adaptation in the genome of livestock

Daniel John Pitt

A thesis submitted to Cardiff University for the degree of Doctor of Philosophy

November 2021

SUMMARY

Livestock have been a cornerstone to human sustenance, wealth, culture, and production for over 10,000 years. Each contemporary livestock species is comprised of a vast assemblage of locally adapted breeds containing unique genetic variation. While most of this variation remains uncharacterised, understanding the genetic diversity is highly valuable in conserving and improving extant livestock breeds. Chapters two to four utilise SNP array data comprised of tens of thousands of loci distributed relatively uniformly across the genome. Firstly, the demographic history of taurine (*Bos taurus*) and indicine (*Bos indicus*) cattle was modelled with approximate Bayesian computation (ABC) to determine if domestication occurred across two or three independent events, with model rejection sampling indicating only two domestications took place. Additionally, characterising population structure, admixture, and gene flow to identify migratory events and potential sources adaptive introgression. Secondly, the 15th century colonisation of the Americas by taurine Creole cattle from their putative Iberian ancestors was modelled using ABC. Identifying a founding effective population size (N_e) of 84, followed by a large demographic expansion and subsequent contraction, with higher resolution of recent demographic fluctuations visualised with N_e Slope (NeS) analysis. Furthermore, signals of selective sweeps were scanned for, identifying loci important for tropical adaptation including suggesting a new candidate gene (*GDNF*) for the slick hair coat phenotype. Thirdly, population structure and selective signals were identified in Ryeland sheep (*Ovis aries*), which is a resilient and ancient British breed. Haplotype-based detection of selective sweeps and landscape genomics which incorporated environmental, land use, and topographical data identified prominent pathways involved in prion disease pathways and cancer regulation (specifically hepatocellular carcinoma); environmental associations indicate selection may be driven by liver fluke abundance. The final data chapter utilises whole-genome resequencing (WGRS) data to identify SNPs in the feral Chillingham cattle (*Bos taurus*) herd which has been isolated from gene flow for at least 300 years with minimal management. WGRS bypasses most ascertainment biases inherent to SNP arrays, allowing a more comprehensive characterisation of variation – identifying consistently long runs of homozygosity interspersed with islands of excess heterozygosity. Heterozygous peak windows were putatively maintained by balancing selection and were enriched with quantitative trait loci associated with fertility and milk. Unexpectedly, the major histocompatibility regions were almost absent of SNP variation in Chillingham.

TABLE OF CONTENTS

TABLE OF CONTENTS: FIGURES

TABLE OF CONTENTS: TABLES

TABLE OF CONTENTS: SUPPORTING INFORMATION

Table S4.3. Correlation between latitude and longitude with environmental variables used for landscape genomics in Samβada and loadings of environmental variables on the first 4 components from principal component analysis. Longitude and principal component 3 displayed a significant relationship with genotypes extracted from Samβada analysis.

Table S4.4. Significant functional annotation clusters and KEGG pathways derived from the XPEHH analysis against UK breeds and the median scores of nine global breeds.

Table S4.5. IBD relatedness between 60 Ryeland sheep, calculated with the --GENOME command in PLINK. Graphical representation in Figure 4.3.

Figure S5.1. Sequencing depth frequency histograms for Chillingham and Angus individuals, displaying only read depths less than 30. Vertical dashed line indicates mean read depth. Solid black curve indicates cumulative proportion of reads over increasing read depths shown on the right axis.

Figure S5.2. Runs of homozygosity (RoH) across cattle autosomes. Lowest (light blue) points indicate the position of variable sites across the dataset. Each horizontal level above displays the RoH for a unique individual – from bottom to top: Chillingham (PW_2 – PW_11); and Angus (AAN_1 – AAN_10).

Figure S5.3. Inbreeding coefficient estimation based on the proportion of the genome covered by runs of homozygosity (RoH). Visually displayed as a function of the length of RoH contributing to F_{RoH} for each individual.

Figure S5.4. Decay of linkage disequilibrium measured as r^2 between pairwise SNPs as a function of physical distance. Grouped into bins spanning 50 kb. Interquartile range displayed for each bin with median values displayed as a horizontal black line.

Figure S5.5. Difference between binned mean linkage disequilibrium (LD) measured as r^2 for HPW in each breed compared to the breed's respective genomic estimates. A positive value indicating greater r^2 across HPW with respect to genomic estimates.

Figure S5.6. Heterozygosity peak analysis for each autosome in Chillingham (n=10; bottom) cattle, each vertical blue line represents the median breed score of a 100 kb sliding window across the genome and the heterozygosity per base pair observed within. Windows with fewer than 10 SNPs were excluded. Grey regions show heterozygosity peak windows (HPW) – identified as outlying windows in at least 5 individuals for each given

breed. Quantitative trait loci (QTLs) that are significantly enriched within HPW detailed (top). QTLs displayed in descending order of enrichment magnitude and categorised as meat (blue), fertility (green), or milk (red).

Figure S5.7. Major histocompatibility (MHC) regions, continuation from Figure 5.6. Red regions denote the span of MHC genes described in Behl *et al.* (2012). Runs of homozygosity in Chillingham cattle genomes, each horizontal level represents a unique individual, displayed as the top half of the plots. Hardy-Weinberg equilibrium excess heterozygosity *p*-values for polymorphic SNPs within Chillingham (dark blue) and monomorphic SNPs within Chillingham that are polymorphic within Angus (light blue) displayed as the bottom half of the plots.

Figure S5.8. Normalised variant effect prediction and SIFT tolerances for exonic variable sites within heterozygosity peak windows (HPW) and genome-wide estimates for Chillingham (CIL) and Angus (AAN) *Bos taurus* breeds. SIFT scoring requires amino acid alterations so is only applicable to non-synonymous substitutions. For each category, counts were normalised by equating the sum of non-synonymous substitutions to 1, allowing a proportional comparison to synonymous sites and between SIFT predictions.

Table S5.1. Sequencing read depth statistics and quality control for Chillingham (n = 10) and Angus (n = 10) cattle.

Table S5.2. Breed genetic diversity, inbreeding, missingness across individuals and breeds derived for both whole-genome measures and restricted to heterozygosity peak windows.

Table S5.3. Heterozygosity peak analysis summaries. Individual-based initial window counts, heterozygosities and distribution. Chromosomal-based merged window counts, positions, and length.

Table S5.4. Significantly enriched quantitative trait loci in Chillingham heterozygosity peak windows.

Table S5.5. IBD relatedness between ten Chillingham individuals and ten Angus individuals, calculated with the --GENOME command in PLINK.

ACKNOWLEDGEMENTS

First of all, I'd like to thank BBSRC for funding the past four years of my life and giving me this opportunity as well as the incredible cohort I was able to be a part of. Thank you to my supervisors who have supported me over the last 4+ years. To Pablo, thank you for your guidance since my placement year, you've helped me grow as an independent researcher. You've been a great mentor and friend; it's been an enjoyable journey since we first spoke when you shouted at me for disrupting one of your lectures! To Mike, thank you for your support and your work has always been an inspiration to me since undergrad. To Mark, thank you for your amazing patience while trying to teach me about statistics, I learnt so much working with you. To Natalia, your excellent guidance and supervision got my post graduate career off to a fantastic start, thank you!

Thank you to everyone in the MolEcol group for making my time in Cardiff incredible. Jordan, I don't know I could've finished my PhD without you whispering weird Star Wars quotes in my ear or constantly supplying me with coffee made with the weirdest of flavours. Becca Hemsworth, thanks for being a great climbing buddy and you better take your desk with you. Lorna, thanks for being the less great climbing buddy and for always being someone I can bully in the office, I will miss the exchange of insults.

One of the highlights of my PhD was the time spent at The Donkey Sanctuary for my placement. Absolutely everybody was friendly, welcoming, and contagiously enthusiastic about their work – thank you everyone from the research, data, conservation, and veterinary teams. Thank you, Fiona, for allowing me to join your fantastic team and your effort to ensure my time spent with you was engaging and varied despite having substantial COVID disruptions! Thanks, Sarah, for making days in all sorts of weather measuring countless donkeys a thoroughly enjoyable time – apart from that one time with 'Twinkle Toes'… I learnt so much during my time with the conservation team, thank you so much for your patience and enthusiasm Daniel, Helen, Rob, and Ruth, I think my clothes are still soaked from the final ash survey with Helen.

A special thanks to Jody and Christina, there is no way I would've coped with the PhD without your love, rants, laughs and support. You two are by far the best thing to come out of this PhD for me, and I'll forever miss our little office corner.

Finally, thank you to the rest of my family and friends. Especially, my parents for their constant support and encouragement of my education. Matthew, your success has always been a motivational source of healthy sibling rivalry for me. Holly and Thea for always being there and making me laugh when I'm struggling most both inside and outside PhD life. Meg, binging climbing, takeaways, and house plants together was all that motivated my write up, thank you!

Chapter One

General Introduction

1.1 DOMESTICATION

For the vast majority of human existence (*Homo* spp.) food procurement occurred through hunting and gathering, enabling the support of small and often nomadic communities. At least 15,000 years ago (YA) the domestication of the wolf (*Canis lupus*) began, paving the way for domestication of a multitude of animal and plant species (MacHugh *et al.*, 2017). The subsequent cultivation of crops and livestock – agriculture – is regarded as a pivotal moment in human cultural development, allowing the relatively rapid transition from lower-density nomadic peoples to higher-density sedentary societies (Diamond, 2002). The greater abundance and reliability of food production supported specialisation of an individual's role within a community (including non-agricultural specialties), facilitating technological advancements, wealth accumulation, and eventually giving rise to modern civilisation, thus the period has been coined "the Neolithic revolution" (Weisdorf, 2005; Ajmone-Marsan *et al.*, 2010).

As agricultural practices were dispersed, so too were the domestic species. Progenitors of large livestock species such as cattle (*Bos primigenius*), sheep (*Ovis orientalis*), goats (*Capra aegargrus*), and pigs (*Sus scrofa*) had relatively small distributions (largely limited to regions within Eurasia, the Near East and north Africa), however, human migration has shaped the dispersal of livestock with contemporary domestics spread across most of the planet at high-density (Steinfeld *et al.*, 2006; Zeder, 2017). In the UK alone, there were 9.4 million cattle and calves and 21.8 million sheep and lambs as of December 2020 (DEFRA, 2021). Although outnumbered by chicken and poultry, cattle and sheep are the largest constituents of the British livestock output by biomass (DEFRA, 2021).

The utilisation of livestock surpassed that of just a meat source; providing milk, hides, wool, draught power, and even religious or culturally important ceremonies (Ajmone-Marsan *et al.*, 2010). Selection of traits associated with the production of primary and secondary products is common, for example increasing the size of smaller livestock to yield more meat, or the protein composition of milk having improved nutritional value for human consumption (Beja-Pereira *et al.*, 2003). Physiological and morphological traits less directly linked to production have also experienced selection, notably coat colour, reduced fear response and size of larger livestock species to improve tameness, shorter flight distance, and precocious sexual maturation (Mignon-Grasteau *et al.*, 2005). Contrastingly, some traits conferring an advantage in wild populations but less essential in domestic population experience relaxed selection, e.g. camouflage, predator detection and avoidance, sexual selection, and forage motivation (Alberto *et al.*, 2018). Domesticated species therefore offer an interesting model for understanding the genetic mechanisms underlying the trade-off of selection between opposing traits (Jensen, 2006).

The symbiotic relationship between humans and livestock has resulted in substantial cultural shifts and also facilitated co-evolution, e.g. milk production in cattle and lactose tolerance in Europeans (Beja-Pereira *et al.*, 2003). However, while agriculture may be considered a mutualistic relationship between the species directly involved, the practices have far reaching implications; it is a primary factor in transforming and degrading natural landscapes, reducing biodiversity, and altering atmospheric composition, thus contributing to the current climate crisis (Zeder, 2017; Rojas-Downing *et al.*, 2017). Nonetheless domesticated livestock are indispensable in producing protein from plants growing on uncultivable land and providing food security for the growing human population.

1.2 FOOD SECURITY AND CLIMATE CHANGE.

In 2019, the global human population reached 7.7 billion (UN, 2019). The past 30 years has seen an annual net gain of 82.5 million inhabitants, resulting in 2.5 billion more people. Forecasts show that by 2050 this will further increase to 9.7 billion (8.9 – 10.6 billion for low and high estimates), despite a declining growth rate from the present 1.08% annual increase to just 0.50% (UN, 2019). In the same time frame, worldwide agricultural production will have to increase by 70-110% to provide for the expanding population and the demand for higher living standards (Heinke *et al.*, 2020).

Presently, pastures and cropland occupy approximately 40% of the planet's terrestrial land (Foley *et al.*, 2011), yet the area covered has remained relatively constant since 1991 (O'Mara, 2012). Whilst this is partially reflective of technological improvements and intensification (Rojas-Downing *et al.*, 2017), the acquisition of additional cultivatable land, fisheries, and water will be increasingly challenging as more land is utilised to accommodate the growing human population (Myers *et al.*, 2017). This effect will be exacerbated as current systems are overgrazed, land is degraded through high intensity farming, and geopolitical tensions rise over increasingly limited resources (O'Mara, 2012).

Global meat production has tripled over the past 50 years, with around 340 million tonnes produced in 2018 (Heinke *et al.*, 2020). Similar expansion is observed with animal feed crops, where a third of cropland is now dedicated to supplement the diet of livestock (Steinfeld *et al.*, 2006). Livestock production demand will increase predominantly in less- and least-developed countries, not only because of rapidly rising populations, but also due to income per capita growing by 3.7% and 4.7%, respectively (FAO, 2009). Rising income often spurs higher dietary

spending and consumption of animal proteins, often through a shift to processed products (O'Mara, 2012). Meat and milk demand will increase so dramatically by 2050 – by 33% and 66%, respectively (Alexandratos and Bruinsma, 2012; Heinke *et al.*, 2020) – that it has been coined the "livestock revolution" (Thornton, 2010; Rojas-Downing *et al.*, 2017).

The diverse livestock industry can provide broad and numerous opportunities in tackling the UN's Sustainable Development Goals (UNHCR, 2018). It provides a direct livelihood for at least 600 million smallhold farmers in some of the poorest areas of the world (Thornton, 2010). This in turn supports a wider network of 1.3 billion jobs indirectly reliant on the sector (Thornton, 2010), producing an estimated 40% of the global agricultural gross domestic product (GDP; FAO, 2009). Reliance of livestock within the global population's diet is increasing, accounting for a third of the protein consumed and almost a fifth of calorific intake (Thornton, 2010). The greatest proportion of animal protein from livestock is provided by ruminants through both milk and meat production (FAO, 2011). Ruminants, such as sheep, goats, cattle, and buffalo, remain an integral part of agriculture in part due to their ability to digest hay, silage, and fibrous crop residue that is otherwise inedible to humans and convert it to usable calories (Eisler *et al.*, 2014). Furthermore, they can remain productive in environments with poor soil fertility such as mountainsides and low-lying wet grasslands, reserving more fertile farmland for crops consumed directly by humans (Eisler *et al.*, 2014; Pulina *et al.*, 2017).

Further expansion and intensification of livestock systems is also impeded by increasing concern for global climate change. The warming of the atmosphere is primarily caused by greenhouse gas (GHG) emissions (IPCC, 2013). While the livestock sector is responsible for approximately 2% of the global gross domestic product (FAO, 2009), it disproportionately contributes 14.5% of global GHG emissions (Gerber *et al.*, 2013); moreover, production of animal-based food (including animal feed) can be twice as GHG emission intensive as plant-based food (Xu *et al.*, 2021). Carbon dioxide (CO_2), methane (CH_4) and nitrous oxide (N_2O) are the predominant GHG associated with livestock. The latter two having a significantly higher global warming potential than CO_2, an estimated 28 and 265 times the carbon dioxide equivalent (CO_2e) effect on global warming for CH_4 and N_2O respectively (IPCC, 2006). It is noteworthy that the exact CO_2e values are somewhat debated in more recent literature and locally-sourced animal derived products can be more efficient both with respect to land use and GHG emissions (IPCC, 2013; Rojas-Downing *et al.*, 2017).

Direct contributions to GHG emissions from livestock include manure storage and excretion, enteric fermentation (in ruminants), and respiration (Jungbluth *et al.*, 2001). The majority is comprised of indirect emissions through feed crop production, transportation, farm

operations, manure application, product processing, as well as land use change and degradation (Mosier *et al.*, 1998). Enteric fermentation is the process in which carbohydrates are broken down by microorganisms in the fore-stomach (rumen) producing large quantities of methane and is the most emissive process in the livestock sector – accounting for 39.1% of GHG contextualised with CO_2e (Gerber *et al.*, 2013). Manure application, storage and excretion contributes 25.9%. Feed production contributes 21.1% to the overall sector, however, within monogastric species where enteric fermentation does not occur this may increase to 60-80%. (Sonesson *et al.*, 2009; Gerber *et al.*, 2013).

Utilising ruminants for production is a widespread and established system that provide substantial benefits and exploit important economic and environmental niches. It is therefore difficult to diversify away from such species and breeds solely based on high GHG emissions. Fortunately, there have been numerous studies investigating modifying aspects of the livestock sector to reduce the negative environmental impact, while minimising production loss. The high methane emissions from enteric fermentation can be reduced by marginally increasing dietary lipid content (Beauchemin *et al.*, 2008), improving the digestibility and quality of forage (Knapp *et al.*, 2014), supplements and feed additives (Boadi *et al.*, 2004; Beauchemin *et al.*, 2008; EPA, 2020) including an 80-99% reduction of enteric methane production by introducing red seaweed into the diet of beef bullocks (Roque and Duarte, 2020), or potentially chemicals and vaccines that inhibit the growth of methanogenic microorganisms in the rumen (EPA, 2020). Manure management can be improved by limiting storage time, introducing anaerobic digestors that convert methane to CO_2 (Gerber *et al.*, 2008) and separating solid from liquid excrement (Rojas-Downing *et al.*, 2017). Indirect sources of GHG emissions such as feed production could also be mitigated through regular soil testing, efficient and well timed fertiliser usage (Grossi *et al.*, 2019), and growing well-adapted higher-yield crops (Steinfeld *et al.*, 2006). While none of the aforementioned lists are exhaustive, they provide insight into some of the possible mechanisms may improve the sustainability of the livestock sector. Caution must be taken when implementing changes to avoid "pollution swapping" where one modification indirectly results in equal or greater emissions of GHG during a different phase of production (Hristov *et al.*, 2013). Although mitigation strategies should encompass as many aspects of improvement as possible, one approach with potentially high impact is the identification, management, and expression of the full genetic potential of livestock species (Grossi *et al.*, 2019).

1.3 IMPACT OF CLIMATE CHANGE ON LIVESTOCK

Negative feedback loops exist between climate change and the livestock sector; not only is climate change accelerated by the expanding livestock sector, but the production efficiency of the livestock sector is also inhibited by climate change. Anthropogenic global warming is likely to reach a 1.5°C increase above pre-industrial levels between 2030 and 2052 (IPCC, 2018) with an overall likely rise of between 0.3°C and 4.8°C by 2100 (IPCC, 2013). The negative effects of climate change will not be homogeneous, with a local mean temperature increase of 1 – 3°C in high altitude regions potentially boosting crop productivity (Thornton, 2010). Nonetheless, most lowlands in the northern hemisphere will face significant challenges, with North America, West-Central Asia, Northern Europe, and the Mediterranean basin experiencing the greatest adversity (Easterling *et al.*, 2007; Nardone *et al.*, 2010). Alongside increases in monthly and seasonal average temperatures, extreme climatic events such as coldwaves, heatwaves, droughts, and floods are increasing in both frequency and magnitude (Pasqui and Di Giuseppe, 2019). Importantly, the effects of climate change are expected to impact almost every step of the supply chain, including transport, storage, retail, and farm-level output with both direct and indirect influences on animal productivity (Figure 1.1). A few of the far reaching and complex factors influencing the production, welfare, and survival of livestock particularly relevant to work carried out in this thesis are *briefly* discussed below.

1.3.1 DISEASE

Management of diseases will experience complex disruption as a changing climate alters the reproduction, distribution, and development of disease hosts, vectors, and pathogens. Increased exposure to ultraviolet B radiation through the depletion of the ozone layer can reduce T helper 1 lymphocyte count, thus, inhibiting the mammalian cellular immune response to pathogens (Aucamp, 2003). Temperature increases may be associated with increased pathogen growth within the environment prior to host infection and ultimately increasing the pathogen burden (Abdela and Jilo, 2016). Naïve livestock populations could be exposed to agriculturally-relevant diseases such as malaria and trypanosomiasis through modified distributions of disease vector pests including mosquitos (*Anopheles* spp.) and tsetse flies (*Glossina* spp.) (Thornton, 2010). Fasciolosis, a debilitating and potentially lethal disease in sheep and cattle is caused by the parasitic liver fluke (*Fasciola hepatica*) (Machicado *et al.*, 2016). The distribution of *F. hepatica* is limited to regions with a temperature range between 10-25°C, sufficient moisture, and the presence of the intermediate molluscan host *Lymnaea truncatula*; the projected future changes to the UK climate, and in particular the mild winters, raises the risk of fasciolosis infection with

serious epidemics expected in Wales by 2050 (Fox *et al.*, 2011). This particular example is highly relevant to the Welsh sheep industry and is discussed further in **Chapter Four**.

Economic impact of unanticipated livestock diseases and epidemics can be immense but are often challenging to quantify due to the unique presentation and the diversity of sectors that are influenced. Common consequences of livestock disease range from impacting food demand and productivity, limiting international trade, damaging environmental resources, biodiversity, and landscape as well as burdening human health through drug resistance and direct medical costs (Thornton, 2010). Bovine spongiform encephalopathy (BSE) infections in the UK cost in excess of £4 billion in the late 1990's (Smith and Bradley, 2003). Additionally, the emergence of avian influenza (e.g. H5N1), swine flu (H1N1), MERS, and most recently, coronavirus (COVID-19) highlight the importance to also monitor potential reservoirs of zoonotic transmission to mitigate the risk of future outbreaks.

Fortunately, the incidents of livestock disease have reduced in recent decades due to advancements in diagnostics, vaccinations, and treatments (Thornton, 2010). However, management practices such as routine antibiotic usage, often used to encourage livestock growth or to substitute hygiene requirements, proliferate antimicrobial resistance (He *et al.*, 2020). The livestock sector is responsible for over half of all antibiotic consumption (Van Boeckel *et al.*, 2017). Not only does this underline the sector as a key source of antimicrobial resistance, but also indicates our current reliance on drug efficacy to reliably produce food.

1.3.2 WATER QUALITY AND QUANTITY

Water is increasingly becoming a scarce resource, with 64 percent of the human population expected to be living in water-stressed basins by 2025 (Rosegrant *et al.*, 2002). The agricultural sector is the largest contributor to freshwater consumption, accounting for 70% of global human usage, with a tenth of this going towards feed production for livestock alone (Thornton *et al.*, 2009; Legesse *et al.*, 2017). Livestock consumption is variable dependent on climatic conditions, with higher temperatures raising water requirements by up to three fold (Nardone *et al.*, 2010; Godde *et al.*, 2021). The water efficiency of crop and animal production will have to increase, or systems will have to be localised around water availability potentially further restricting food security in arid regions.

Raw water quality is also at risk from increasing temperatures, rising sea levels, and heavy rainfall which increase sediment suspension as well as nutrient and pollutant loading (Godde *et al.*, 2021). Salination of water sources will reduce livestock feed intake and production (Valente-

Campos *et al.*, 2019), while heavy metal and chemical contaminants may impact digestive, respiratory, cardiovascular, and nervous systems in exposed animals (Nardone *et al.*, 2010).

1.3.3 THERMAL STRESS

Each animal has a thermal comfort zone, dependant mostly on ambient temperature and humidity, in which physiological processes are most efficient (Rojas-Downing *et al.*, 2017). Susceptibility to deviations from this zone ranges dependent on numerous factors, including species, breed, body shape, size, life stage, nutrition, and genetic potential of the animal (Thornton, 2010). Before considering any effects on production rates or reliability, it is apparent that heat stress is an animal welfare concern indicated through increased cortisol secretion and negative behavioural responses including aggression, malaise, and thirst (Polsky and von Keyserlingk, 2017; Bagath *et al.*, 2019). Nonetheless, increased temperatures have short- and long-term implications on animal productivity, feed intake, fertility, health, and mortality (Rojas-Downing *et al.*, 2017). Furthermore, highly productive individuals are more vulnerable to heat stress, this for example, results in dairy breeds being impacted further by temperature and humidity when compared to beef cattle (Godde *et al.*, 2021). Contrastingly, some regions in high latitudes will experience an overall reduction in temperature or an increase in extreme cold events, thus extending growing seasons, reducing forage nutrition, reduced water availability, and require increased energy consumption to heat livestock housing (Godde *et al.*, 2021). Cold-stressed livestock require higher energy feed or greater feed intake to maintain equivalent production, however, this is not a linear effect as nutrient availability and enzyme activity are hindered by the reduced temperature, further decreasing efficiency (Guo *et al.*, 2021). Sheep are particularly vulnerable to cold stress; hypothermia in lambs is already the primary cause of economic loss within systems under risk of cold exposure (Bhimte *et al.*, 2018), the threat of which may increase under shifting climatic conditions.

1.3.4 DIET

Protein, minerals, vitamin, and energy requirements vary drastically between both livestock species and management systems, with some systems sustained mostly on forage and others exclusively on supplementary feed, making the effect of climate change on livestock diet variable and complex (Thornton *et al.*, 2009). For foraging livestock, changes in temperature or humidity is likely to reduce forage quality, quantity and extend the growing season (Polley *et al.*, 2013). The reduction in overall forage nutrition increases methane emissions (Benchaar *et al.*, 2001) and requires supplementation with other food sources such as grain, however, this both passes the demand onto another potentially vulnerable feed production system and requires the

dietary adaptation to handle more nutrient rich intake (Rojas-Downing *et al.*, 2017). Interestingly the biochemical structure of forage will also change, increased temperatures may increase lignin and cell well production in plants, requiring longer digestive periods for ruminants (Polley *et al.*, 2013). Additionally, the forage composition will change as species distributions shift, potentially exposing livestock to novel dietary items to which they are maladapted to digest (Thornton, 2010).

Many studies focus on the on production, where observations of reduced forage intake, milk yield, growth rates, and carcass quality all correlated with heat stress (Bernabucci *et al.*, 2010). Additionally, highly productive individuals are more vulnerable due to increased metabolic

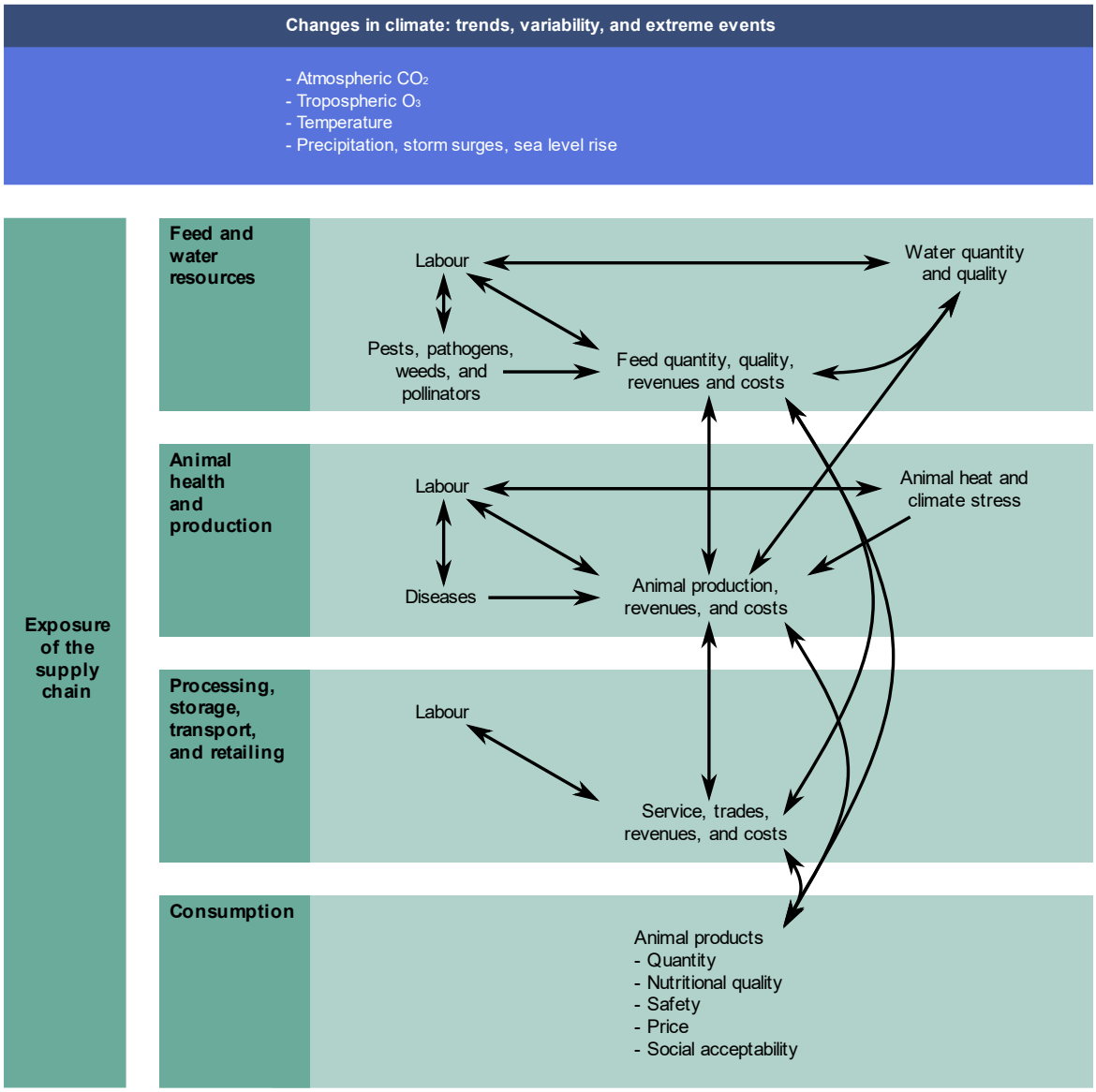


Figure 1.1. Potential climate-related impacts on the livestock food supply chain. Directional arrows indicate a negative interaction, all labelled impacts in the supply chain (green) are also directly affected by changes in climate (blue). Adapted from Godde *et al.* (2021).

heat generation, this for example results in dairy breeds being impacted further by temperature and humidity when compared to beef cattle (Godde *et al.*, 2021). Fertility declines across both sexes in both mammals and poultry, with reduced ovarian function, spermatozoa motility, pregnancy rates, and impaired embryo development (Rojas-Downing *et al.*, 2017; Godde *et al.*, 2021). Cortisol production during acute stress improves immune function, however, chronic secretion is associated with immunosuppression and increased vulnerability to pathogens (Bagath *et al.*, 2019).

Selection or local adaptation for thermotolerance is a possible solution, for example, *Bos indicus* cattle maintain lower respiration rate and rectal temperatures when compared to *Bos taurus* under the same heat stress conditions (Bernabucci *et al.*, 2010); however, there could be a potential production cost, heat-tolerant *B. indicus* generally produce meat which is less tender and fetches a lower price (O'Connor *et al.*, 1997). Alternatively, climate controlled livestock production systems provide conditions to maximise animal production and limit exposure to the external climate, however, the raw materials and energy expenditure required to sustain such systems are likely to ironically accelerate climate change (Hou *et al.*, 2021).

1.4 MOLECULAR APPROACHES

Historically, phenotype and pedigree data were primarily used to describe, identify, and analyse individual animals and breed compositions; however, rapidly advancing molecular techniques of the past three decades have allowed analysis of genotypes and genomic structure. Techniques include restriction fragment length polymorphism (RFLP), mitochondrial DNA sequencing (mtDNA), amplified length polymorphisms (AFLP), variable number tandem repeats (VNTR) which encompass mini- and microsatellites, and single nucleotide polymorphisms (Yaro *et al.*, 2017). The latter two have been particularly popular techniques in studies measuring genetic diversity of farm animal breeds in the past fifteen years (Olschewsky and Hinrichs, 2021).

1.4.1 SINGLE NUCLEOTIDE POLYMORPHISMS

A single nucleotide polymorphism (SNP) is variation at a specific single nucleotide in the genome; SNPs often become established through point mutations in the germline that are subsequently inherited by the offspring. While each individual SNP has relatively low power to discriminate between individuals, the scalability of the technique and development of SNP arrays provides extensive genome-wide information (McTavish and Hillis, 2015). SNP arrays are a type of DNA microarray that are comprised of immobilised allele-specific DNA probes for thousands

of biallelic loci of interest, with each locus allele fluorescently labelled. SNP arrays are read by an optical detection system capable of interpreting the release of the fluorescent labels once the DNA of a sample is hybridised to the array. There are commercial SNP arrays available for most major livestock species – including cattle, sheep, swine, chickens, goat, buffalo, and even salmon, trout, and shrimp – facilitating cost-effective genotyping of tens of thousands to hundreds of thousands of loci for a given individual (Georges *et al.*, 2019). The affordable extensive screening capabilities of SNP arrays allows population-wide insights into genetic variation, demographic history such as bottlenecks, expansions, domestication, and gene flow, as well as selection and the identification of selective sweeps and the genetic association of phenotypic traits.

Initial identification (calling) of SNPs for arrays, requires deep sequencing of a set of individuals, known as the 'ascertainment panel', that are deemed as representative of the taxon or group of interest. The ascertainment panel is often linked with the aims of the study for which the array is developed, potentially biasing inferences made with future applications of the array (Matukumalli *et al.*, 2009). Ascertainment bias in SNP arrays can arise from multiple steps in the design process (e.g. equal spacing between SNPs and several adjacent nucleotides are required to be invariable; Geibel *et al.*, 2021), however, the predominant biases that especially affect inter-breed comparisons arise from either: (i) a bias in minor allele frequencies (MAF) where there is an abundance of high MAF polymorphism and a lack of low MAF polymorphism, often resulting from an insufficient number of samples in the ascertainment panel and thus rare alleles are more easily missed; or (ii) a subpopulation bias, where SNPs identified as polymorphic in the ascertainment panel are not as variable or monomorphic in more distantly related groups, leading to an under-estimation of genetic variability (McTavish and Hillis, 2015; Orozco-terWengel *et al.*, 2015). Briefly, mitigation of ascertainment bias is possible through methodologies such as pre-analysis filtering such as pruning loci in linkage to reduce overestimation of heterozygosity in ascertainment populations (Malomane *et al.*, 2018), directly modelling the ascertainment scheme to account for the allelic variation missed in populations outside of the ascertainment panel (Nielsen, 2004), or by using haplotype-based analyses that are less vulnerable to single SNPs with outlying allele frequencies (Sabeti *et al.*, 2007).

The rapidly decreasing cost of sequencing has enabled the resequencing of thousands of livestock individuals (Georges *et al.*, 2019). The 1000 Bull Genomes Project is perhaps the most well-known and rapidly expanding whole-genome resequencing (WGRS) scheme in livestock, with 6,192 taurine and indicine animals sequenced in run 9 as of September 2021 (Daetwyler *et al.*, 2014; Hayes and Daetwyler, 2019; Daetwyler, 2021). WGRS effectively overcomes many of the issues of ascertainment bias observed in. SNP arrays by bypassing the non-random selection

of loci and attempting to capture all variants present in the genome. WGRS for the purpose of SNP capture is still relatively expensive with respect to microarray-based alternatives, especially considering commercial SNP arrays such as Illumina's Bovine High Density BeadChip assay 770k SNPs per individual. Furthermore, a high-quality reference genome is required to fully exploit the short read sequencing data and a large sample size is preferable to identify alleles with low MAF (Olschewsky and Hinrichs, 2021). Nonetheless, if these requirements are met, WGRS provides a powerful tool to capture the unique genetic variation in rare and overlooked breeds that are especially susceptible to the negative effects of SNP array ascertainment bias (Yaro *et al.*, 2017).

1.5 DEMOGRAPHY

Demography in livestock can be complicated to unravel. Firstly, it is commonplace for livestock to have undergone multiple, independent domestications, such as cattle (Loftus *et al.*, 1994; Bruford *et al.*, 2003; Ajmone-Marsan *et al.*, 2010), sheep (Kamalakkannan *et al.*, 2021), and pigs (Larson *et al.*, 2005). Nonetheless, signs from multiple wild progenitors do not necessarily confirm multiple domestications but may instead be indicative of introgression or restocking from divergent wild populations (Larson and Burger, 2013) as is observed in camels (Almathen *et al.*, 2016) and cattle (Upadhyay *et al.*, 2017). The incorporation of livestock species in Neolithic culture, economy, and trade inevitably tied animal demography to human movement. The slow dispersal of pastoral communities and practices out of the domestication centres resulted in serial founder effects and potential decline in genetic variation with increasing distance (Taberlet *et al.*, 2011). Additionally, the domestication process often modifies breeding behaviour, favouring systems with a high female to male ratio and non-random mating, introducing familial structure and reducing the effective population size (N_e).

1.5.1 POPULATION STRUCTURE

Familial relationships shape the genetic structure of populations and can lead to stratification within and between different sub-populations dependant on levels demographic isolation and gene flow. Understanding the population structure and relatedness between individuals can therefore give insight into the demographic history and shared ancestry. Algorithmic approaches such as principal component analysis (PCA) or multidimensional scaling (MDS) reduce the dimensionality of the input data by analysing the variance-covariance structure among genotypes (Duforet-Frebourg *et al.*, 2016). Complex, multilocus genetic data can therefore be simplified into a few synthetic variables that are representative of the major divisions of the

dataset and thus interpreted as population structure. Important advantages to such approaches is no prior knowledge of population structure is required and there are no strong underlying assumptions about the genetic model, such as allele frequencies or linkage (Jombart *et al.*, 2009). Model-based approaches such as ADMIXTURE (Alexander *et al.*, 2009) and STRUCTURE (Pritchard *et al.*, 2000) group individuals into a user-defined number of ancestral population clusters (K). A range of K is often analysed, both to identify the 'optimal' clustering solution based on likelihood and to examine population substructure. Due to the increased computational demands of these methods and the assumption that loci are independent, pruning markers that are non-randomly associated through linkage disequilibrium (LD) is recommended to improve computational speed and to align with model assumptions (Alexander *et al.*, 2009). Pairing the analysis with software capable of inferring directionality and strength of gene flow – such as TREEMIX (Pickrell and Pritchard, 2012) – minimises the risk of over-interpreting model-based clustering (Lawson *et al.*, 2018).

1.5.2 COMPLEX DEMOGRAPHIC INFERENCES

The high density of genetic data available allows exploration of complex demography including population divergences and changes in N_e over time, both of which are of key interest when investigating the development and health of livestock breeds (Yaro *et al.*, 2017). Notably, the majority of software capable of more complex demographic inferences often require heavier supervision than population structure algorithms; relevant parameter estimates are essential to improve model predictions (Schraiber and Akey, 2015). With the development of SNP arrays, a common strategy to estimate N_e is to exploit the relationship between N_e and LD across the physical distance of the genome, as is carried out by SNEP (Sved, 1971; Barbato *et al.*, 2015). LD is the non-random association between alleles at different loci; while alleles would be expected to be passed to offspring as part of a single haplotype, recombination may occur between loci. As recombination rate is function of the distance between loci and is scalable to a population level relative to N_e, LD can therefore be used to estimate N_e changes over time. Unfortunately, multiple confounding factors such as population structure and selection can disrupt LD and consequently effect N_e estimation (Orozco-terWengel and Bruford, 2014).

The sequentially Markov coalescent (SMC) algorithm approximates full coalescent with recombination of sequences and thus has been the focus for many programs estimating N_e with WGRS data (McVean and Cardin, 2005). Advancements have included inferences from a single diploid genome using pairwise SMC (PSMC; Li and Durbin, 2011) or inclusion of multiple genomes with multiple SMC (MSMC) which is able to estimate a cross-coalescent rate between populations, functioning as a proxy for migration (Schiffels and Durbin, 2014; Malaspinas *et al.*,

2016). Importantly, these haplotype-based approaches require substantial data processing steps and often require phased data, which can be challenging to produce depending on raw data quality and reference scaffolds. Fortunately, reference genomes are rapidly improving for many livestock species (Rosen *et al.*, 2018). Relatively affordable optical mapping and long-range sequencing allows for the construction of uninterrupted, high reliability reference scaffolds with deep read depth. In the case of the Bovine ARS-UCD 1.2 reference genome, long-read PacBio sequences at 80x genome coverage enabled *de novo* assembly of scaffolds spanning entire chromosomes (Rosen *et al.*, 2018). These improvements allow for greater base accuracy during the construction of highly repetitive genomic regions frequently found in Eukaryotes, such as the MHC which consists of numerous duplicated genes (He *et al.*, 2021)

A particularly adaptable and computationally faster approach is approximate Bayesian computation (ABC). The intuitive idea underlying ABC is that simulated predefined models with high posterior probabilities will produce summary statistics most similar to those calculated from the empirical data. ABC utilises approximate rejection sampling, in which posterior distributions are synthesised from the set of prior parameters that encode the simulations that yielded summary statistics closest to the observed summary statistics from real data (Wegmann and Excoffier, 2010). A major strength of ABC is that it allows for incredibly complex demographic histories to be explicitly programmed; however, increased complexity risks over-parameterisation and correct selection of descriptive, unbiased summary statistics is challenging while avoiding high dimensionality (Schraiber and Akey, 2015). Nonetheless, a well-defined model can produce precise posterior probabilities of any defined aspect of demography, including migration rate, demographic bottlenecks or expansions, domestication events and population divergences (Gray *et al.*, 2014). Self-contained software such as DIYABC (Cornuet *et al.*, 2008) improves the accessibility to the ABC approach, but is limited in application. Contrastingly, ABCTOOLBOX provides a highly flexible skeletal framework to incorporate independent simulation and summarising software or scripts dependent on the specific requirements of the study (Wegmann *et al.*, 2010).

1.6 SELECTION

As first postulated by Darwin and Wallace (1858) and coined as 'natural selection', the presence of beneficial character state within an individual can improve the likelihood of survival and reproduction, thus becoming more prevalent in subsequent generations. Particularly in domestic species, such as livestock, natural selection functions alongside selective forces applied

by humans, known as artificial selection (Gregory, 2008). Historically, artificial selection would have been largely unconscious with no ultimate end goal, such as a farmer breeding from more tame individuals simply due to the practicality of handling the animal (Jensen, 2006). Eventually, the mating of individual livestock began being controlled, restricting breeding largely to animals with the most desirable traits aiming to increase the frequency or strength of the trait's appearance in the offspring (Diamond, 2002). Traits considered 'desirable' still vary today, often dependant on the farmer's individual preferences but can range from body conformation (Deniskova *et al.*, 2018), behaviour (Jensen, 2006), reproduction (Beynon *et al.*, 2015), environmental adaptation (Huson *et al.*, 2014) as well as primary production traits such as milk, meat, but also secondary product production (Garforth, 2015). Initially selection occurred in a largely passive manner with slight preference to animals with greater adaptation to diseases, food types, and local adaptations (Russell, 2007), however, the intensity and methodical nature of artificial selection slowly increased, leading to the emergence of a more formal paradigm approximately 200 YA known as the "breed concept" (Taberlet *et al.*, 2011). Restricting the gene flow between populations and selectively breeding within populations has given rise to thousands of breeds with distinctive phenotypes across domestic species.

1.6.1 Patterns of selection

Regions with little or no effect on fitness (i.e. selectively neutral) are thought to represent most of the genetic variation observed both within and between populations (M. Kimura, 1968). The neutral theory forms a baseline as demographic processes such as migration, isolation, expansions, and bottlenecks are the predominant force in shaping the neutral allele frequency spectrum through the random sampling of alleles, otherwise known as genetic drift. Contrastingly, when considering loci with a fitness-altering affect, selective pressures also modify the allelic frequency. Therefore, it is possible to identify putative candidates for selection from substantial outliers to the neutral allele frequency spectrum distribution (Sabeti *et al.*, 2002).

Different types of selection leave behind complicated patterns or signatures in the genome; the most well studied form is directional (also referred to as positive) selection and negative (also referred to as purifying) selection (Vitti *et al.*, 2013). The nomenclature refers to fitness-altering effect of a given allelic variant, positive if the allele is beneficial and therefore favourably maintained, or negative if the allele is deleterious and therefore selectively removed. While neutral alleles by definition are not under selection, the variability at neutral sites can be influenced by selective pressures applied to nearby fitness-altering alleles (Sabeti *et al.*, 2002). This occurs through "hitchhiking" which relies upon the genetic linkage of nearby genomic regions. In the example of a beneficial allele under positive selection that is approaching fixation, nearby

neutral alleles will experience a loss or total elimination of variation in a process known as a selective sweep (Figure 1.2; Maynard Smith and Haigh, 1974). The recombination rate, N_e and intensity of selection influences the size of the genomic region influenced by a selective sweep, with neutral diversity increasingly less affected at greater genetic distances to the swept allele (Maynard Smith and Haigh, 1974; Kim and Stephan, 2002). A similar effect is observed for negative selection where an unfavourable or deleterious allele is removed and linked variation in nearby neutral sites is reduced, known as "background selection" (Charlesworth and Willis, 2009).

Numerous other forms of selection exist between the extreme forms of fixation or loss of variants through strong positive or negative selection, respectively. Weaker directional selective pressures may result in incomplete selective sweeps, where fixation is never fully achieved. Alternatively, the presence of standing genetic variation or recurrent identical (or very similar) mutations may occur resulting in a soft selective sweep which at the molecular level seems to have an unexpectedly rapid decay of LD (Figure 1.2). Contrasting strong directional selection, it is possible for multiple alleles to be selectively maintained at a single locus through balancing selection. Commonly, this occurs through a heterozygote advantage, where a heterozygous genotype confers a greater fitness than either homozygous genotypes (Allison, 1956; Hedrick, 2011). Balancing selection can also occur through mechanisms such as frequency dependent selection where whichever genotype is the lowest frequency has the highest fitness in the population at a given time (Wright, 1939). Additional selective processes include disruptive selection where the extreme phenotypes are selected for as they maintain an advantage over intermediate phenotypes, or stabilising selection where an intermediate phenotype is selected for as it is more advantageous than extreme phenotypes. Whilst mentioned as independent processes, these selective mechanisms are tightly linked at a molecular level, for example frequency dependent selection is often achieved through alternating positive and negative selection for an allele, whereas stabilising selection can be achieved through selection for heterozygous individuals (De Filippo *et al.*, 2016).

The literature has predominantly focussed positive selection, largely for two reasons. Firstly, positive selection is easier to detect when compared to both balancing selection which can produce allele frequencies similar to neutral expectations and negative selection which often acts in highly conserved regions resulting in little or no apparent change to targeted allelic frequencies after fixation has occurred (Vitti *et al.*, 2013). Secondly, positive selection has the greatest influence on adaptive evolution, this makes it easier to associate genomic signatures with phenotypic data or environmental variables (e.g. local adaptation), and it is of greater interest when investigating beneficial traits, particularly in the livestock sector (Utsunomiya *et al.*, 2013).

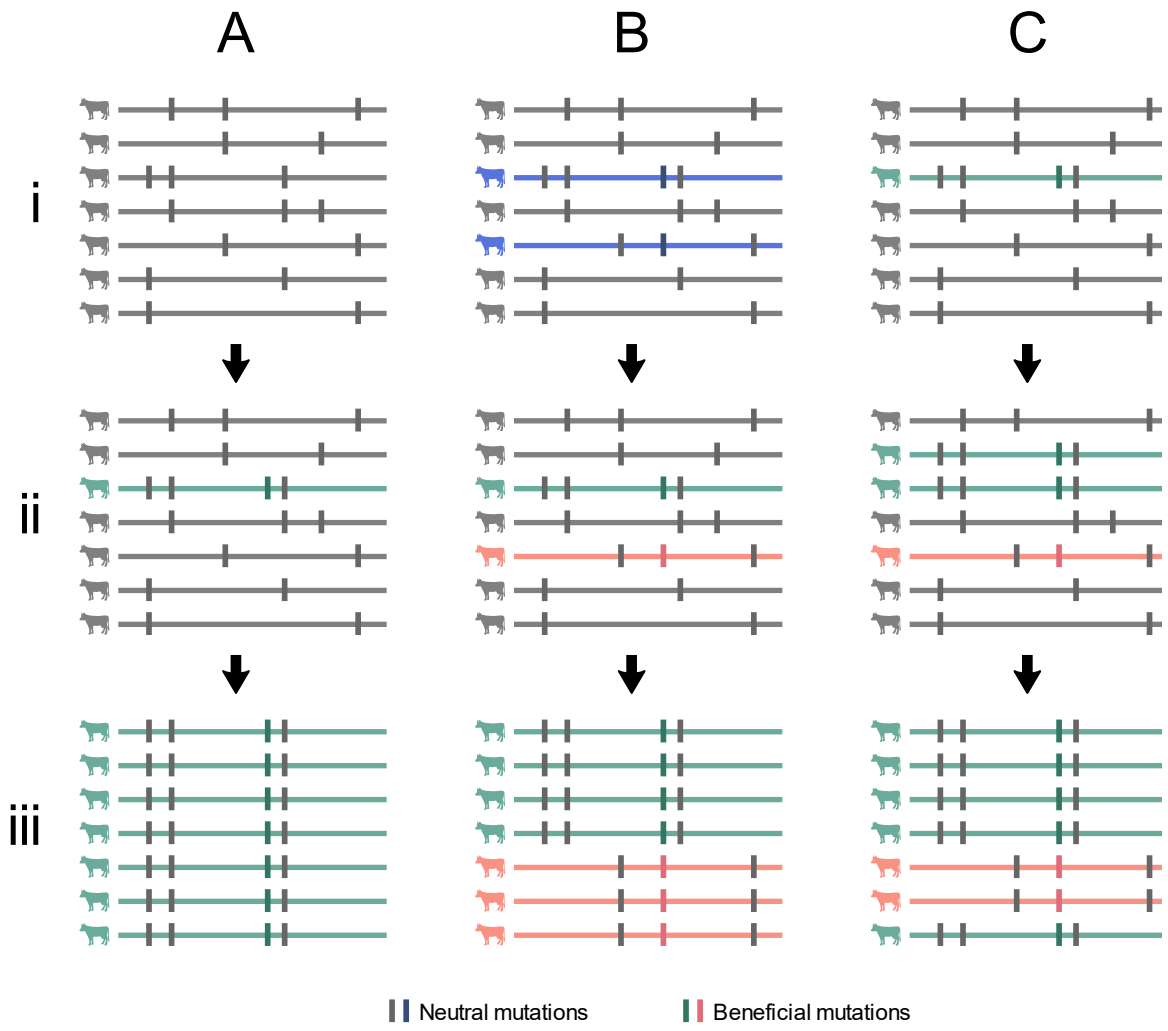


Figure 1.2. Types of selective sweep, without recombination. A) Classic/hard selective sweep: i) seven DNA sequences of a neutrally evolving region within a population, neutral alleles (grey) displayed as vertical bars; ii) a novel mutation arises in one genome that is associated with improved fitness of the individual (green); iii) over subsequent generations the beneficial mutation increases in frequency and eventually becomes fixed within the population. The genetic variation in the surrounding genome is reduced despite only containing neutral mutations (grey). B) Soft sweep from standing genetic variation: i) seven DNA sequences of a neutrally evolving region within a population, neutral alleles (grey/blue) displayed as vertical bars; ii) the neutral allele (blue) becomes beneficial (green/red) due to different environmental or genetic circumstances; iii) over subsequent generations the beneficial mutations increase in frequency and eventually become fixed within the population. The genetic variation in the surrounding genome is reduced, however, to a lesser extent due to the original standing genetic variation. C) Multiple origin soft sweep: i) seven DNA sequences, six are neutrally evolving with one showing a novel beneficial mutation (green); ii) the green mutation is increasing in frequency within the population while an identical or very similar beneficial mutation occurs at the same locus of a different individual (red); iii) both beneficial mutations increase in frequency causing genetic variation around the locus to decrease. Adapted from Saravanan *et al.* 2020.

1.6.2 DETECTING POSITIVE SELECTION

Some of the earliest indicators of positive selection were developed in the 1990s were derived from the measurements comparing substitution rates within genes across taxa, for example, McDonald Kreitman test (MKT) and d_N/d_S (McDonald and Kreitman, 1991; Goldman and Yang, 1994). Both of these metrics utilise the putative effect of a base pair substitution on a protein coding sequence and to detect selection on a macroevolutionary scale, i.e. between species (Goldman and Yang, 1994). Synonymous substitutions (d_S), where the altered codon sequence produces the same amino acid are assumed to be neutral, are compared to non-synonymous substitutions (d_N), where the amino acid coded by the codon sequence is changed by mutations thereby making the mutant more likely to be exposed to selective pressures (McDonald and Kreitman, 1991). Broadly, if the assessed region is neutrally evolving and free of selective pressures, d_N should be equivalent to d_S, or their ratio (d_N/d_S) should be equal to 1; however, under positive or negative selection, the ratio will be above or below one, respectively (Goldman and Yang, 1994). Both MKT and d_N/d_S ratios are more sensitive to macroevolutionary processes between species due to the underlying assumption that diverged positively selected variants are swiftly brought to fixation and do not appear as polymorphic within a species (Messer and Petrov, 2013). Similarly, it is assumed deleterious mutation are quickly lost and do not contribute to either polymorphic or divergent sites (Messer and Petrov, 2013). More recent implementations of MKT can consider multi-locus effects allowing within-species assessments of diversity and selection across genes (Egea *et al.*, 2008), as well as considering allele frequencies of polymorphic non-divergent sites that capture the presence of weakly deleterious alleles at low frequency and beneficial alleles that have not reached fixation (Haller and Messer, 2017).

The development and affordability of high density genome-wide markers, such as SNP arrays, has allowed a transition to methods to assess fine scale patterns of selection and given insight into microevolutionary processes within species (Vitti *et al.*, 2013; Qanbari and Simianer, 2014). A range of statistical tests have been developed to identify signatures left behind by positive selection in both SNP array and DNA sequence data (e.g. Tajima, 1989; Fay and Wu, 2000; Sabeti *et al.*, 2002; Nielsen, 2005; Voight *et al.*, 2006; Sabeti *et al.*, 2007; Rubin *et al.*, 2010; Qanbari and Simianer, 2014; Saravanan *et al.*, 2020). The primary division of these methods is whether they compare data within populations (intra-population) or between populations (inter-population). Summarised succinctly by Saravanan *et al.*, (2020), intra-population statistics primarily focus on the reduced local variation, modified allele frequency spectra, or altered regional LD. Whereas

inter-population statistics analyse differentiation between either haplotypes or single sites (Figure 1.3).

Within this thesis, the research focus is not on single individuals, nor is it restricted to analysing macroevolution over a large evolutionary timescale, such as across species or genera. In the past 200 years it has become increasingly commonplace to develop distinct breeds in livestock species, leading to the formation of populations closely related through recent ancestry, yet oftentimes maintaining heavily restricted gene flow (Ajmone-Marsan *et al.*, 2009; Decker *et al.*, 2014). Therefore, this provides numerous biological replicates of breeds with similar origins and demographic histories that have recently begun experiencing different selective pressures since divergence. The application of intra-population statistics to effectively identify selective sweeps requires knowledge of neutral expectations or the ancestral state of the population, which is often difficult to ascertain. Importantly, demographic history such as population expansions, contractions, migrations, isolation, admixture and introgression can impact the allele frequency spectrum at variable rates across the genome making interpretation of signatures of selection difficult (Utsunomiya *et al.*, 2015). Contrastingly, inter-population statistics can be more robust against demographic interferences by utilising the abundance of genomic data available from closely related livestock breeds to provide numerous comparative populations and biological replicates. I will therefore focus more keenly on methods to define inter-population statistics,

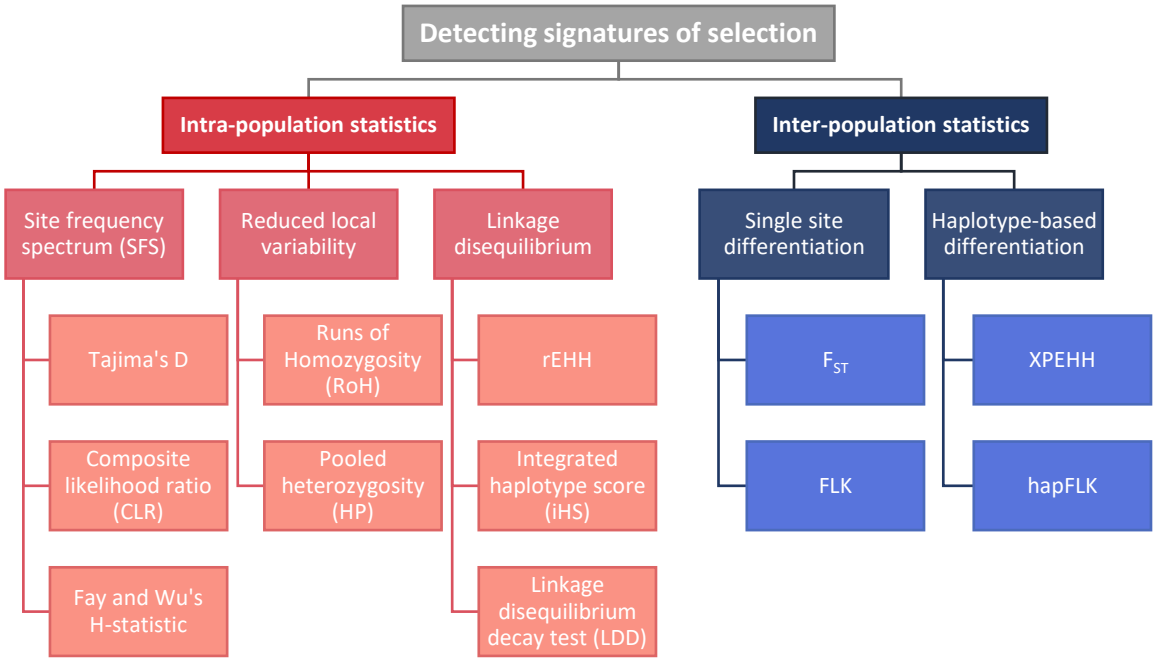


Figure 1.3. Common methodologies to detect signatures of selection in livestock populations. Adapted from Saravanan *et al.* 2020.

nonetheless it is important to consider the background of intra-population statistics as they often provide a theoretical basis for the development of inter-population methodologies.

1.6.2.1 Intra-population statistics

The site frequency spectrum (SFS) is the distribution of allele frequencies across the same type of loci (e.g. SNPs) genotyped within a population. Deviations from a neutrally evolving SFS can be indicative of selection (Figure 1.3). Statistics such as Tajima's D (D) compares the pairwise differences to the number of segregating sites between polymorphism data (Tajima, 1989). Fay and Wu's H (H) statistic uses the non-ancestral (derived) SFS to identify high frequency non-ancestral allele, however, for this it is necessary that ancestral states of polymorphisms are determined which is often substituted with an outgroup, used to infer the ancestral state of each polymorphism (Fay and Wu, 2000). A negative D value, indicating an excess of rare polymorphisms, can arise through mechanisms such as a rapid population expansion after a recent bottlenecks or selective sweeps; H is often used as a complementary method alongside D to determine the presence of high frequency derived alleles and thus a selective sweep. A third method, the maximum composite likelihood ratio (CLR), compares the SFS generated under neutral expectations with the SFS generated through a selective sweep (Nielsen, 2005). CLR can provide sensitive estimates of positive selection, however, the establishment of a null model is achieved through simulations that require demographic data, and information on regional recombination rates and mutation rates within the population for accurate identification of signatures of selection. SFS-derived statistics are less powerful than most modern alternatives due to the higher density of markers regularly used and the ability to assess haplotypes and LD with high confidence and at high genomic resolution (Saravanan *et al.*, 2020). Additionally, common quality control methods for SNP array data that are aimed at reducing genotyping error, such as minor allele frequency filtering thresholds, also disrupt the SFS with particular biases against rare alleles and thus disrupt SFS-derived statistics.

Genetic measures that are often utilised for whole-genome based analyses of demographic history, population structure, and genomic inbreeding can also be investigated at local levels to identify selection. Crucially, assumptions are made that demographic factors impact the genome uniformly, whereas selective pressures modify a particular gene and the linked region surrounding it (Qanbari and Simianer, 2014). Runs of homozygosity (RoH) are contiguous lengths of a diploid genome with homozygous genotypes which correspond to two haplotypes that share a recent common ancestor and are identical-by-decent (IBD) (Gibson *et al.*, 2006). Particularly long RoH are expected to occur around loci that have undergone strong and rapid directional

selection. Contrastingly, pooled heterozygosity measures allele counts in sliding windows, and local deviation of heterozygosity from genome-wide averages (Rubin *et al.*, 2010).

Selective sweeps can produce heightened levels of LD associated with particular haplotypes that arises during the rapid increased frequency of the selected loci that does not allow sufficient time for recombination to shuffle nearby genomic regions (Sabeti *et al.*, 2002). The extended haplotype homozygosity (EHH) methodology identifies these core genotypes through the decay of linkage over distance. Following this, normalisation through relative EHH (rEHH) allows comparisons of haplotypes segregating in the same genomic region to identify signals of positive selection (Sabeti *et al.*, 2002). Compared to the previously discussed methods, LD-based methods are more sensitive to detecting partial or incomplete sweeps while also dampening the effects of ascertainment bias that can heavily influence single marker SFS estimations (Wolf and Ellegren, 2017). The integrated haplotype score (iHS) incorporates the ancestral state into EHH analyses to identify positive selection of novel mutations in derived haplotypes (Voight *et al.*, 2006). Although iHS can effectively identify sweeps at intermediate frequencies, after fixation the effectiveness decreases due to the lack of variability at the selected loci (Sabeti *et al.*, 2010). A similar methodology is the linkage disequilibrium decay (LDD); this test examines only homozygous genotypes within an individual and the relative LD surround the locus, thus removing the requirement for phased data and reliable haplotype information (Wang *et al.*, 2006).

1.6.2.2 Inter-population statistics

A widely used single site differentiation metric in livestock is Wright's fixation index (F_{ST}), that measures the variance in allele frequencies at a given locus between two or more populations (Wright, 1949). F_{ST} ranges between 0, indicating that there are no differences between the allele frequencies of the two populations compared, to 1, indicating that differences in allele frequency between the two populations are maximum with different alleles fixed in each population. In neutrally evolving regions and within a system with a low mutation rate relative to migration rate, F_{ST} can provide an estimate of population differentiation through genetic drift. Unexpectedly high values of F_{ST} for a given locus relative to the genome-wide average F_{ST} or neutral expectations can be indicative of selection, potentially occurring through reproductive isolation and divergent selection in both populations or positive selection in one population (Randhawa *et al.*, 2014). Unlike haplotype- or SFS-based statistics, as F_{ST} is calculated per SNP it is possible to identify the specific variant under selection. Key limitations of F_{ST} include the assumption of an infinite population size which can result in overestimation of the metric when sampling is limited, as well as assuming all populations derive independently from the same ancestral population. Several more recent iterations adjust for the number of sampled populations, sample size, multi-allelic

loci, (Weir and Cockerham, 1984) and differential demographic histories of the diverged populations (Bhatia *et al.*, 2013). For example, FLK can consider the heterogeneity of population sizes over time and hierarchical clustering by calculating a kinship matrix between populations, thus producing a more complex expectation of genetic drift (Bonhomme *et al.*, 2010). Additionally, adopting windowed-based approaches when estimating F_{ST} can improve the reliability in identifying selective sweeps by incorporating adjacent loci in the estimation of F_{ST} and thus reducing reliance on high-F_{ST} outliers that are false positives with respect to targets of selection (Wolf and Ellegren, 2017).

Extensions of the haplotype-based identification of selective sweeps is the cross-population extended haplotype homozygosity (XPEHH) test. This method essentially combines iHS and EHH but differs in that EHH-curves are generated then integrated within each of two separate subpopulations. A directional ratio is then calculated between the two integrated EHH profiles for each SNP. This eliminates the requirement for ancestral allelic state identification and can identify selective sweeps where the selected allele is approaching or is in fixation in one population but polymorphic in the second population (Sabeti *et al.*, 2007). XPEHH has almost twice the statistical power to detect complete selective sweeps than iHS (Sabeti *et al.*, 2007; Qanbari and Simianer, 2014).

A key theme throughout this thesis will be the interpretation of selection under differential environmental conditions, however, there are often difficulties in gathering sufficient environmental data that is both high-resolution and with an appropriate temporal span. Inter-population statistics compare divergent groups to identify signatures of selection present in at least one of the populations. Comparisons of populations that predominantly differ only in their habitat, the metrics can provide a proxy for local adaptation – an invaluable advantage over intra-population statistics.

1.6.3 LANDSCAPE GENOMICS

Selective pressures are not uniform across a landscape, but instead form a complicated array of spatial and temporal conditions to which organisms may adapt to or migrate from to avoid extinction (Kawecki and Ebert, 2004). Exposure to differential selective pressures over a geographical range can result in differential selective responses known as local adaptation; this is where beneficial alleles conferring an improved fitness to the specific regional conditions have undergone positive selection in a population. Methodologies to detect locally adaptive variation can be described as either top-down or bottom-up. Top-down approaches utilise the relationship between phenotypic and genotypic data, including genome-wide association studies (GWAS) and

quantitative trait loci (QTL) mapping (Stinchcombe and Hoekstra, 2008). In the absence of phenotypic data, it is more common to use bottom-up approaches that relate genomic data directly with environmental conditions, through either population or landscape genomics (Rellstab *et al.*, 2015). As discussed previously, the population genomics approach largely relies upon the identification of selective sweeps (Figure 1.2). The use of biological replicates to identify consistently occurring signatures of selection can improve the reliability in detecting local adaptation and extracting enriched biological functions through gene ontology or previously mapped QTLs can indicate locally adaptive phenotypes where annotation is available (Kawecki and Ebert, 2004). Population genomics is however limited in detecting the sometimes-subtle alterations to allelic frequencies that occur through local adaptation, for example, during polygenic or epistatic interactions of additive genetic variation (Pritchard and Di Rienzo, 2010). Furthermore, intra-population outlier tests do not consider the quantifiable variation in environmental conditions and thus potentially the underlying selective pressures between populations (Rellstab *et al.*, 2015).

Landscape genomics directly incorporates environmental data into models to identify genome-environment associations. Originally, direct correlations between the frequency of alleles with environmental variables were calculated (Hedrick *et al.*, 1976), but the increasing number of markers (e.g. AFLPs and microsatellites) required the development of multiple simultaneous univariate logistic regressions to analyse associations (Joost *et al.*, 2007). Various developments have considered differential sample sizes, population structure (Günther and Coop, 2013), spatially explicit modelling (Guillot *et al.*, 2014), LD (Gautier, 2015) and confounding demographic factors (de Villemereuil and Gaggiotti, 2015). High-density genome-wide markers (e.g. SNPs) and modern high resolution environmental measures expose the computationally intensive (oftentimes exponential) scaling of previous methodologies. More recent software, such as Samβada allows for efficient multivariate regressions of multiple environmental variables and population structure with genotypic data whilst also accounting for spatial autocorrelation to enable the identification of local adaptation in a population (Stucki *et al.*, 2017).

1.7 AIMS OF THIS THESIS

The aims of this thesis are broadly divided into two objectives: firstly, to understand how demographic processes have affected the distribution of genetic variation in extant livestock breeds, and secondly, to understand how selective processes have contributed to modifying the extant genetic variation in livestock species. Each objective was investigated using genomic data

in a range of commercial, hobbyist, and feral livestock (namely cattle and sheep). This is carried out over different scopes, ranging from broad global panels with representative diversity across an entire species, to narrow focus on a single breed. Allowing investigation into the domestication process, large-scale gene flow, common signals of selection as well as breed-specific demographic challenges and local adaption. This was carried out with the expectation of informing management decisions of the breeds studied to ensure their survival and adequate use of their genomic resource. Simultaneously providing targetable farm animal genetic resources (FAnGR) for the plethora of extant breeds that are predicted to face ecological and environmental challenges in the near future at magnitudes exceeding the rate at which local adaptation can typically naturally occur.

While this thesis intends to provide information directly applicable to the populations analysed, livestock provide an excellent model species. Livestock have been the topic of detailed historical records for centuries, often heavily influencing human culture and have received significant focus in the genomics era, with high precision reference genomes, quantitative trait analysis, and genome annotations. Our understanding of the history of livestock populations provides additional context and validation, thus allowing calibration for many genomic techniques enhancing their application to other species.

1.7.1 Demographic processes

Inferences of population structure and divergence are some of the first steps in understanding the genetic variation and segregation within a dataset. In itself it provides valuable information on the evolutionary history, population splitting, and gene flow between populations this thesis aims to discover. For example. the expansive dataset collated in **chapter two** allows for characterisation of the structure and overall genetic composition of both *Bos taurus* and *Bos indicus* cattle, answering the questions of hierarchical divergences and potential admixture events. Within **chapter three** and **chapter four** population structure is also used to investigate potential adaptive introgression. Population structure is not only descriptive of demographic history, but it is essential to understand to ensure the assumptions of a panmictic subpopulation are met prior to subsequent analysis (where the software relies on such assumptions). **Chapter four** aims to briefly demonstrate the consequences of excluding population structure when considering local adaptation.

In addition to the unsupervised inferences of population structure, explicitly testing complex, parameter-rich demographic models allow for a more complete narrative and estimates of specific parameters when supplementary information is available to define model priors.

Chapter two aims to estimate ancient intra- and inter-species divergences, as well as the timing and frequency of domestication events within cattle. Whereas **chapter three** uses similar methodology to reconstruct recent colonisation and restocking events in *Bos taurus*. Overall, the aim is to both define the specific demographic nuances of each system and how this explains the contemporary distribution and variation, as well as providing evidence for the flexibility of this approach in incredibly complex systems.

Characterising long term effective population size trends can estimate the genetic drift the population is exposed to and the potential impairment of selection efficacy. For more ancient and static estimates of N_e, **chapter two** and **chapter three** aim to gather posterior distribution estimates of all populations utilised in ABC modelling. **Chapter three** adds the novel N_e Slope (NeS) analysis to attempt to disentangle fine resolution changes in the rate of decline of Creole cattle breeds - to test wider application of this method, it is also applied to Ryeland sheep in **chapter four**. By estimating both N_e change and rate of change, the aim is to investigate declines of N_e that are consistent between populations, and to monitor accelerating decline which may indicate increasing extinction risk. Furthermore, this is expanded on in **chapter five**, tracking both N_e declines and cross coalescent rates in a small, inbred population thus is capable of validating the timing and consequence of long-term isolation.

1.7.2 SIGNATURES OF SELECTION

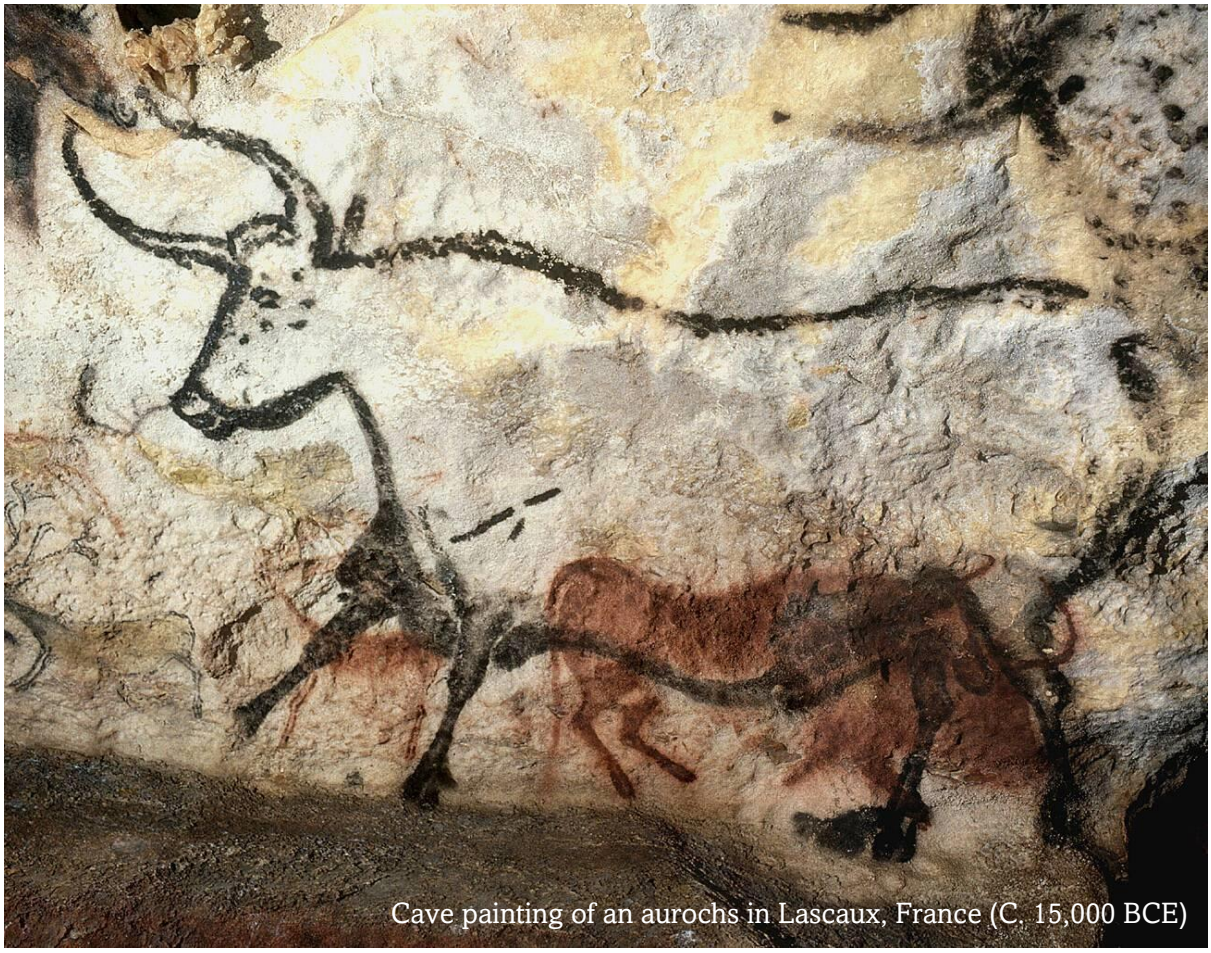
Adaptive change improves the fitness of the individual and can improve the overall persistence of a population as a whole. Under shifting demographic or environmental pressures, selective pressures and advantageous genotypes also shift. Understanding the underlying genetic mechanisms of how populations adapt to external changes and the genomic regions under selection can provide us with insight into evolutionary significant loci and biological processes associated with particular environmental conditions.

This thesis approaches detection of selection from various angles. **Chapter three** utilises a dataset with two recently diverged populations (Iberian and Creole cattle); largely isolated since divergence, Creole cattle have been translocated from a temperate to tropical environment, thus this system provides a biological experiment on the rapid evolution to drastically different environmental conditions, including humidity, temperature, disease, and diet. **Chapter four** studies environmental selection in the British population of Ryeland sheep, quantifying the association of a range of environmental variables against individual genotypes. This finer-scale approach aims to detect more subtle positive selection and correlate it directly with specific environmental conditions. **Chapter five** uses the Chillingham cattle to investigates the limitations

and consequences of selection on severely bottlenecked populations. This chapter aims to quantify the accumulation of deleterious mutations that accumulate in a population with a limited N_e, as well as the balancing selection that maintains variation. While each natural system experiences a variety of forms and strengths of selection, these studies represent challenges faced by a plethora of livestock breeds and wild populations.

Chapter Two

Domestication of cattle: Two or three events



Cave painting of an aurochs in Lascaux, France (C. 15,000 BCE)

2.1 ABSTRACT

Domestication of cattle from aurochs (*Bos primigenius*) first occurred more than 10,000 years ago, providing humans with a reliable source of meat, milk, draught power, and secondary products. Two primary regions have hosted independent domestication events, firstly the Fertile Crescent in the Middle East, giving rise to taurine (*Bos taurus [B.t.]*) cattle, and secondly the Indus Valley in the Indian subcontinent, resulted in indicine (*Bos indicus [B.i.]*) cattle. A third domestication event in the Western Desert of Egypt is hypothesised by some genetic and archaeozoological evidence, potentially responsible for taurine cattle predominantly found in Africa. A global panel of 3,196 individuals and 180 cattle (*B.t.* and *B.i.*) breeds was used to investigate global population structure and divergence. Furthermore, explicit models of domestication history, inter- and intra-species divergence and migratory patterns were iteratively developed with approximate Bayesian computation (ABC). While African taurine showed distinct genetic components to that observed in European conspecifics, comparative analyses between scenarios modelling two and three domestication events consistently favour a model with only two episodes. Differentiation of African taurine was likely catalysed with post-domestication introgression with wild aurochs. Similarly African indicine have experience high levels of gene flow from African taurine, likely representative of attempts to introduce local adaptation after the more recent introduction to the continent. Scenarios with unidirectional or bidirectional migratory events between European taurine and Asian indicine cattle are plausible but less well supported – further studies are needed to disentangle the complex human-mediated dispersion patterns of domestic cattle. This study therefore helps to clarify the effect of past demographic history on the genetic structure of modern cattle, providing a basis for further analyses exploring alternative migratory routes for early domestic populations.

2.2 INTRODUCTION

Aurochs (*Bos primigenius*) were the most populous cattle species from the middle Pleistocene and into the early Holocene (Zeuner, 1963). The maximal distribution of the three subspecies of aurochs spanned the breadth of Eurasia, into the Middle East, India, and north Africa (Zeuner, 1963); with populations persisting in eastern Europe until 1627 AD (Götherström *et al.*, 2005), much later than domestication. Similarly to sheep and goats, archaeological and genomic evidence suggests that the ancestors of taurine cattle (*Bos taurus*) were domesticated in the Fertile Crescent from *Bos primigenius primigenius* during the Neolithic, more than 10,000 years ago (YA; Bruford *et al.*, 2003; Ajmone-Marsan *et al.*, 2010; MacHugh *et al.*, 2017). A second

domestication occurred approximately 8,500 YA in the Indus Valley from *Bos primigenius nomadicus,* resulting in the establishment of Indicine cattle (Zebu cattle; *Bos indicus*) (Loftus *et al.*, 1994). Despite these domestications occurring only 1,500 years apart, it is thought that the respective founding subspecies of aurochs diverged approximately 250,000 – 330,000 YA (Loftus *et al.*, 1994).

Substantial support exists for the extant distribution of cattle arising from two main domesticated lineages; however, a third domestication event has been hypothesised to have occurred in northeast Africa about 8,000 – 9,000 YA, giving rise to the divergent African taurine cattle. Support for this hypothesis derives from archaeozoological evidence and genetic data from contemporary cattle and aurochs. Archaeozoological evidence can be diverse and indicate both the time point at which livestock were domesticated and how the individuals were utilised, including secondary production and draft. For example, milk production can be inferred from ruminant dairy fats identified through biochemical of Neolithic pottery (Bollongino *et al.*, 2012), isotopic analysis of calf teeth allows estimation of weaning age and thus allows inferences about milk exploitation of lactating cows (Vigne and Helmer, 2007), the age and sex ratios of slaughtered animals can indicate the intensity of management and culling for beef production, and overall conformation, craniofacial morphology and size, often associated with "domestication syndrome" (Wilkins *et al.*, 2014). The archaeozoological evidence is supplemented by comparative osteological analyses between ancient wild and domestic cattle from Europe, Asia and Africa (Grigson, 1991; Applegate *et al.*, 2001; Stock and Gifford-Gonzalez, 2013). Bradley *et al.* (1996), using maternal mitochondrial DNA (mtDNA), showed that African cattle feature a higher frequency of the T1 mitochondrial haplogroup than is common in other regions, estimated that the separation between African and European taurine ancestors occurred 22,000 – 26,000 YA (earlier than the Fertile Crescent domestication), and found patterns of population expansions consistent with domestication that were more recent than the corresponding signature of African/European divergence. These results are supported by analyses of extant taurine cattle from Europe, the Middle East and Africa, as well as extinct British aurochs (MacHugh *et al.*, 2001). Furthermore, evidence from nuclear DNA analyses (Hanotte, 2002; Pérez-Pardal *et al.*, 2010; The Bovine HapMap Consortium *et al.*, 2009) are consistent with a local domestication of African taurine cattle and the subsequent admixture of Near East and the Indus Valley cattle in Africa. However, Bonfiglio *et al.* (2012) support the Near Eastern origin of the T1 mitochondrial DNA haplogroups and defended that the North African sub-haplogroup T1d could have originated in already domesticated cattle shortly after their arrival from the Near East. Along this line, analyses on whole-genome SNP arrays supported two domestication events for taurine and indicine lineages followed by introgression from wild aurochs in Africa, East Asia and Europe (Decker *et*

al., 2014). Taken together, controversial archaeological and genetic data seem to support both two and three domestication events in cattle; therefore, an analysis using genealogical modelling of alternative scenarios that could give rise to the extant patterns of cattle genetic diversity needs to be addressed.

Interestingly, additional domestication events in cattle are still debated – an independent domestication has been hypothesised for the Turano-Mongolian (Asian taurine) group, including extant breeds such as Yakut and Buryat, which share considerable ancestry with both Hanwoo (Korean) and Wagyu (Japanese) breeds (Yurchenko *et al.*, 2018). Collating evidence from contemporary (Mannen *et al.*, 2004; Xia *et al.*, 2021) and ancient mitochondrial sequences as well as radiocarbon dating of an ancient bovid mandible, Zhang (*et al.*, 2013) estimates the presence of domesticated cattle over 10,000 YA. More recent SNP array analysis partially refutes this, suggesting the apparent divergence of Turano-Mongolian cattle is possible through early separation and subsequent low effective population size and isolation from taurine conspecifics post-domestication (Yurchenko *et al.*, 2018).

Taurine cattle dispersed from the Fertile Crescent northwest through Anatolia reaching the Balkans and then central Europe by 7,000 – 8,000 YA; dispersal likely followed two major routes – the Danube River and the Mediterranean coast – with subsequent radiation across

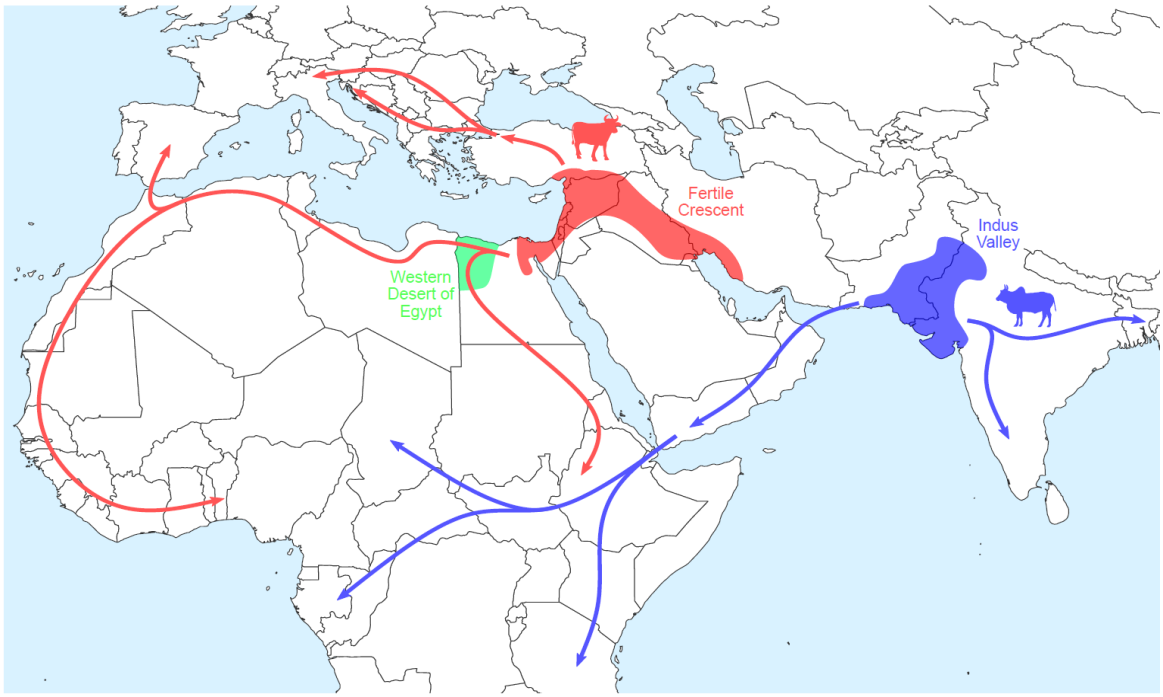


Figure 2.1. Hypothesised major migration routes of taurine (*Bos taurus*; red) and indicine (*Bos indicus*; blue) cattle from the respective domestication centres (shaded). Including the postulated third domestication site in Egypt (green).

Europe (Figure 2.1) (Beja-Pereira *et al.*, 2006; Pellecchia *et al.*, 2007; Ajmone-Marsan *et al.*, 2010). Dispersal into Africa may have started as early as 9,000 YA (Stock and Gifford-Gonzalez, 2013), following both the western coastline of the Red Sea to eventually reach central Africa and also a northern African dispersal that would have allowed migration into the Iberian Peninsula and admixture with local cattle (Beja-Pereira *et al.*, 2006; Decker *et al.*, 2009). Indicine cattle dispersed from the Indus Valley across India, China, and most of Southeast Asia (Ajmone-Marsan *et al.*, 2010). By 4,000 YA Indicine were present in the Middle East and had dispersed into central Africa with evidence of admixture between both aurochs and the established taurine cattle (Hanotte, 2002; Decker *et al.*, 2014; Verdugo *et al.*, 2019).

Previous studies using traditional molecular markers (e.g., mitochondrial DNA and microsatellites) generated insights into domestication, dispersal, and admixture (Bradley *et al.*, 1996; Machugh *et al.*, 1997; Beja-Pereira *et al.*, 2006). However, the development of newer, high-throughput genomic technologies, such as SNP arrays, has enabled the rapid collection of information for thousands of markers with high precision, facilitating comprehensive surveys of genome-wide variation in domestic cattle (The Bovine HapMap Consortium *et al.*, 2009).

Here, a dataset of ~54,000 SNPs was genotyped in 180 cattle populations from across the world, representing both indicine and taurine breeds (Table S2.1), to characterise the genetic diversity, population structure and demographic history of extant cattle. To disentangle the long-standing debate around the existence of a third domestication event in northeast Africa, approximate Bayesian computation was used to model a variety of possible scenarios that could give rise to the extant patterns of cattle genetic diversity, including analyses to determine whether modern cattle derive from two or three domestication events and migratory patterns among breeds.

2.3 Materials and Methods

2.3.1 SNP array data

Illumina's BovineSNP50 v.1, v.2 and Bovine High Density BeadChip 770k SNP array data were merged from several previously published studies (Table S2.1; The Bovine HapMap Consortium *et al.*, 2009; Decker *et al.*, 2009; Gautier *et al.*, 2010; McTavish *et al.*, 2013; Decker *et al.*, 2014; Mbole-Kariuki *et al.*, 2014; Orozco-terWengel *et al.*, 2015; Park *et al.*, 2015; Iso-Touru *et al.*, 2016; Upadhyay *et al.*, 2017). All SNPs were mapped to the UMD3.1 bovine assembly reference genome (RefSeq:GCF_000003055.5) and merged in PLINK v1.90 (Purcell *et al.*, 2007; Chang *et al.*, 2015). After removing duplicated copies of individuals, the final dataset comprised

~54,000 SNPs in most samples and corresponded to 3,196 individuals representing 180 breeds/populations of both taurine (n = 2,041) and indicine (n = 408) cattle, including hybrids of the two subspecies (Table S2.1). Additionally, one individual was genotyped from the archaeological remains of a *B. p. primigenius* animal from England (Park *et al.*, 2015). Across the dataset only SNPs mapped to autosomal chromosomes were retained. Markers with a minor allele frequency below 5% or a call rate less than 90% were removed to account for erroneous calls and missing data, respectively, resulting in a final dataset of 8,081 SNPs (8K) for downstream analyses.

2.3.2 GENETIC VARIATION AND POPULATION DIVERGENCE

The molecular inbreeding coefficient (F_{IS}), and observed (H_o) and expected (H_e) heterozygosities per breed were calculated using PLINK on all breeds (note that estimates from breeds with lower sample sizes may be increasingly unrepresentative of the breed as a whole). Statistical differences between H_e and H_o for each breed was assessed with Welch's t-tests. Population structure across breeds and *Bos* species was investigated with ADMIXTURE v1.3 (Alexander *et al.*, 2009).

The number of user-defined ancestral population clusters (K) assessed for population structure ranged between 1 and 150. The most probable clustering solution was identified by the 5-fold cross-validation (CV) procedure. This process effectively reruns the model 5 times, each time masking (temporarily flagged as missing) a different partition of non-missing genotypes, resulting in a reduction of overfitting and selection bias (Alexander et al. 2009). As the procedure assumes that markers are unlinked, the 8K dataset was pruned further for linkage disequilibrium (LD) in PLINK – using a sliding window of 50 kb, a step size of 10 SNPs, and randomly removing one SNP from each pairwise comparison with an r^2 of 5% or higher. Due to the computational power required for assessing higher values of K and the diminishing returns on population structure for each successive value of K, the analysis was restricted to all values of K between 1 and 50 and used intervals of 5 for K = 55 – 100 and examined K = 150. Resulting in sixty-one values of K ranging between 1 and 150 being tested. The clustering solutions were visualised using the POPHELPER package in R (Francis, 2017; R Core Team, 2018).

Dissimilarity between individuals was measured using Hamming distances and reduced to 20 dimensions with multidimensional scaling (MDS) in PLINK. The first two major axes explaining the highest proportion of the variance in the dataset were graphically represented in R (R Core Team, 2018). F_{ST} was also estimated between all pairs of populations featuring more than one sample and the resultant matrix was used to construct a neighbour-net tree in SPLITSTREE V.4.14.4 (Huson and Bryant, 2006).

2.3.3 APPROXIMATE BAYESIAN COMPUTATION (ABC)

Domestication and demography of taurine and indicine cattle was modelled with Approximate Bayesian Computation (ABC) estimations. This was carried out within the framework of ABCTOOLBOX (Wegmann *et al.*, 2010) which provides a suite of software to estimate model parameters using ABC algorithms. Initially, a large quantity of simulations are performed based on user-defined prior variables; here FASTSIMCOAL2 was used to generate reverse coalescent simulations for each model (Excoffier and Foll, 2011; Excoffier *et al.*, 2013). Summary statistics (e.g., measures of genetic diversity and population divergence) are then calculated in ARLEQUIN V3.5 (Excoffier and Lischer, 2010) both for the observed SNP dataset and each of the simulated datasets. The simulations with summary statistics closest to the observed summary statistics are retained and marginal posterior distributions are calculated for each parameter, thus producing model estimates.

For the ABC analysis, two geographically disparate breeds were selected for both taurine and indicine to produce a single "breed set": European taurine (TaurEU); African taurine (TaurAF); African indicine (ZebuAF); and Asian indicine (ZebuAS). Two additional breed sets were selected as biological replicates (Table 2.1). The breeds chosen avoided outliers to the major clusters and minimised admixture based on the MDS, neighbour-net, and population structure analyses.

Table 2.1. Breeds sets used on the modelling of domestication history of cattle with approximate Bayesian computation (ABC).

Breed set	TaurEU	TaurAF	ZebuAF	ZebuAS
1	Normande (NOR)	Somba (SOM)	Zebu from Madagascar (ZMA)	Kankraj (KAN)
2	Abondance (ABO)	N'dama (NDA)	Zebu Fulani (ZFU)	Bhagnari (BAG)
3	Montbeliard (MON)	Baoule (BAO)	Zebu Bororo (ZBO)	Dajal (DAJ)

The SNP dataset was further reduced to minimise computational load and ascertainment bias; linkage-based pruning was carried out again but at the more conservative r^2 maximum value of 2.5%, reducing the dataset to 2,202 SNPs (2K). The population structure generated in Admixture was qualitatively compared between the 2K and 8K datasets for each breed set to ensure population discrimination remained with the reduced number of markers.

The baseline model describes first the divergence between the two aurochs ancestors of taurine (TaurEU and TaurAF) and indicine (ZebuAF and ZebuAS) occurring approximately 250,000 YA (Bradley *et al.*, 1996; The Bovine HapMap Consortium *et al.*, 2009), followed by the subsequent intraspecies divergences occurring. Domestication events were modelled as bottlenecks within the respective populations. Indicine domestication was always modelled as prior to, or at the time of, divergence between ZebuAf and ZebuAS. This was imitated for TaurEU and TaurAF in modelled scenarios with a total of two domestication events, however, domestication events were positioned after intraspecies divergence of taurine populations when testing the hypothesised third domestication.

Initially, for breed set 1, four model scenarios (Figure S2.1, scenarios 1-4) were parameterised with uniformly distributed priors (Table S2.2), each scenario had 1 million simulations generated. To ensure summary statistics calculated within each simulation were not heavily linked, Spearman's ranked correlation assessed the monotonic association of summary statistics generated from a random sample of 5,000 simulations. Bonferroni correction was applied to account for multiple statistical comparisons, resulting in a reduction from 34 summary statistics to 17 after removal of any that were significantly (positively or negatively) correlated. Furthermore, to improve model fit it was ensured that the observed value fell within the 95% quantiles of the simulated distribution for each summary statistic.

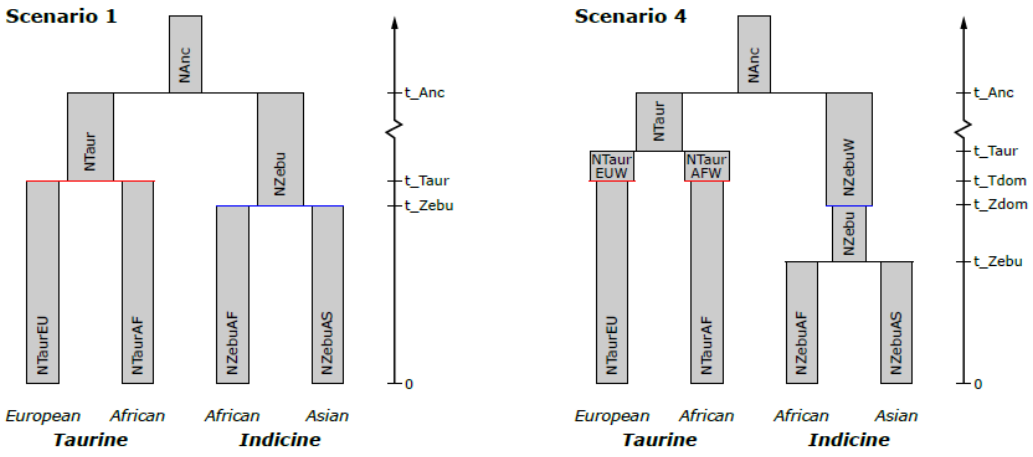


Figure 2.2. Example of two initial modelled scenarios for determining the domestication history of cattle using approximate Bayesian computation (ABC). Scenario 1 models only two domestication events (red and blue lines) that coincide with the divergence of taurine and indicine cattle. Scenario 4 models three domestication events: one in indicine cattle prior to divergence within the indicine group; two within taurine cattle, after divergence within the taurine group.

For each scenario, the 1,000 simulations with summary statistics most similar to the observed data were retained to generate posterior distributions. Model fit was assessed using a generalised linear model that computes the likelihood of the observed data and compares it to the likelihood of each of the retained simulations. A posterior probability, "*p*-value", is reported as the fraction of retained simulations with an equal or smaller likelihood than the observed data; thus, a small *p*-value may indicate the model poorly describes the observed data. Model discrimination was assessed by calculating the Bayes Factor (BF) by taking the quotient of marginal densities of two models; a BF greater than 3 suggests strong support of the first model over the second. After model comparisons between scenarios 1 to 4, strongly supported modelled scenarios were further developed in an iterative process until a total of 15 scenarios were described and tested (Figure S2.1). The eight models with highest support – good model fit and high relative marginal densities – were also simulated with breed sets 2 and 3 (Table 2.1).

TREEMIX v1.17 (Pickrell and Pritchard, 2012) was used to address any ambiguity between ABC models differing by only migration. The three replicate breed sets used in the ABC analyses were merged and a maximum likelihood tree was constructed using TREEMIX under default parameters. Iteratively, weighted migration edges were added to the network, until 99.8% of the variance of the ancestry between the populations was explained by the model.

2.4 RESULTS

2.4.1 ADMIXTURE AND POPULATION STRUCTURE

Population assessed across *B. taurus*, *B. indicus* and *B. p. primigenius* was assessed using Admixture for population clusters between 1 and 150. The most probable clustering solution, where CV error was lowest was at K = 70 (0.545). It is notable that over half (56%) of the total CV error reduction was observed by K = 5, suggesting the clusters separated beyond this point have relatively smaller discriminatory power. At K = 2, divisions between indicine and taurine cattle first emerge, with proportional ancestry most mixed in African animals, even within breeds not explicitly considered hybrids (Figure 2.4). A third cluster separates divides African and European taurine. A fourth cluster separates African indicine, a cluster that is largely shared with breeds indicated as African (*B. t.* X *B. i.*) hybrids. At higher levels of K, many taurine breeds could be differentiated from each other, whereas for indicine breeds, the main differentiation was observed for Asiatic breeds, while the African indicine breeds where more similar to each other, presenting a similar level of admixture.

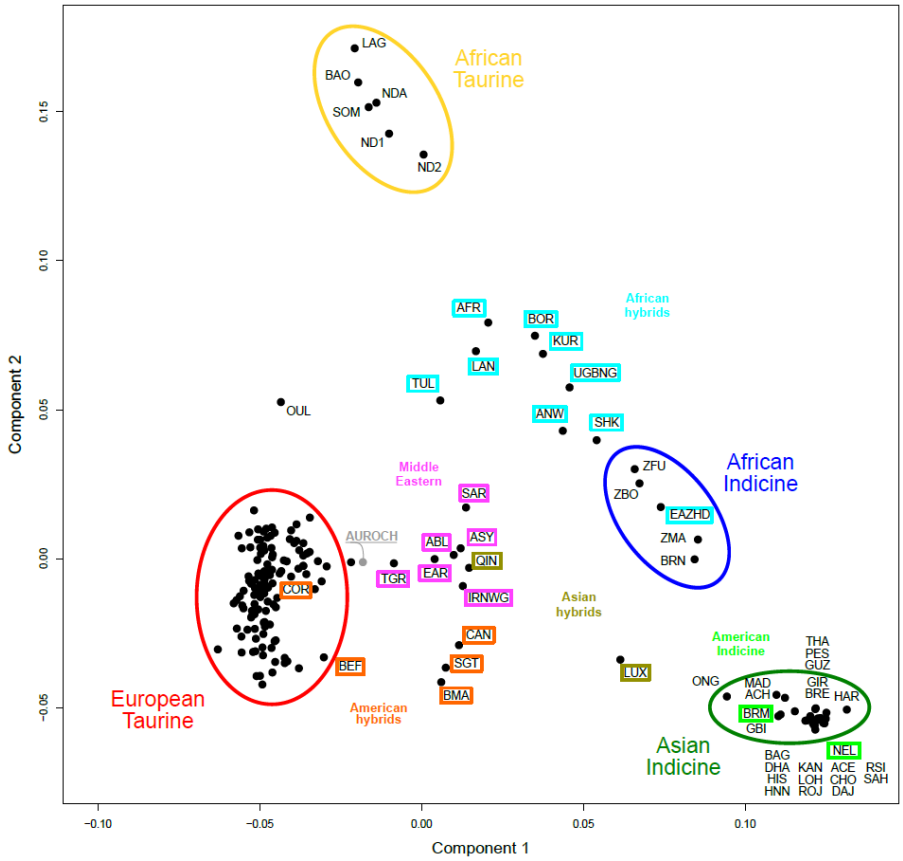


Figure 2.3. Multidimensional scaling (MDS) plot of 3,197 individuals belonging to 180 populations of *Bos primigenius primigenius*, *Bos indicus*, *Bos taurus*, and hybrids (see Table S1 for label information).

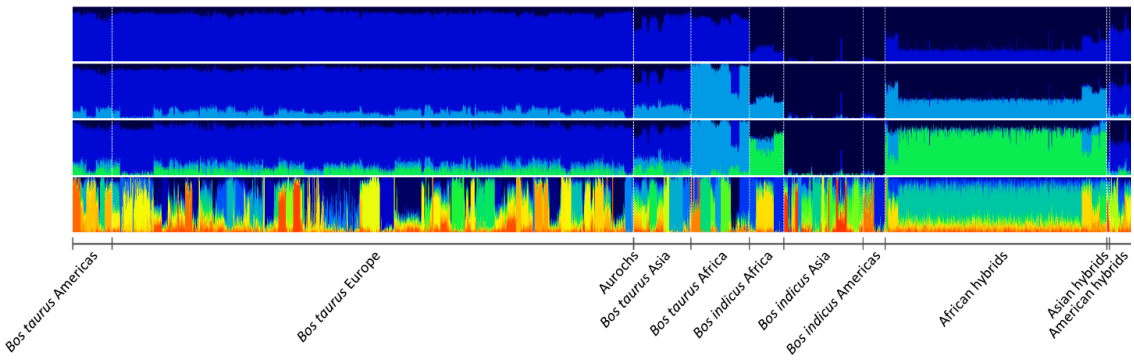


Figure 2.4. Admixture individual assignment plots for 179 cattle populations and one aurochs sample for K = 2, 3, 4 and 70. Each vertical bar represents an individual, the proportion of each colour in that bar corresponds to the ancestry (genetic variation) of an individual deriving from a given cluster.

MDS analyses resulted in the two major axes explaining ~31% and ~15% of the variance, respectively. The first component separated taurine from indicine cattle, with hybrid breeds occurring in the middle between the two main cattle groups, while the second component

separated African taurine from the rest of the cattle breeds (Figure 2.3). The African hybrids were placed between the African taurine and the African indicine populations, while the Asian and American hybrids were place between the European taurine and Asiatic indicine cattle. These results were supported by the breed clustering observed in the neighbour-net analysis, forming three distinct branches isolating Eurasian taurine, African taurine and indicine breeds (Figure 2.5). African indicine cattle occurred between the Asiatic indicine populations and the area of the network where the taurine-indicine hybrids were clustered. Overall, both analyses depicted two main clusters of hybrids between *B. taurus* and *B. indicus* (Figure 2.3; Figure 2.5): i) African hybrids (AFR, ANW, BOR, BRN, EAZHD, KUR, LAN, SHK, TUL, UGBNG), closer to African indicine breeds; and ii) American hybrids (BEF, BMA, CAN, COR, SGT), closer to European taurine cattle. More intermediate positions were assigned to Asian hybrids (LUX, QIN). Apart from hybrids, the Middle Eastern breeds (IRNWG, TGR, EAR, ABL, ASY, SAR) showed a slightly higher proximity

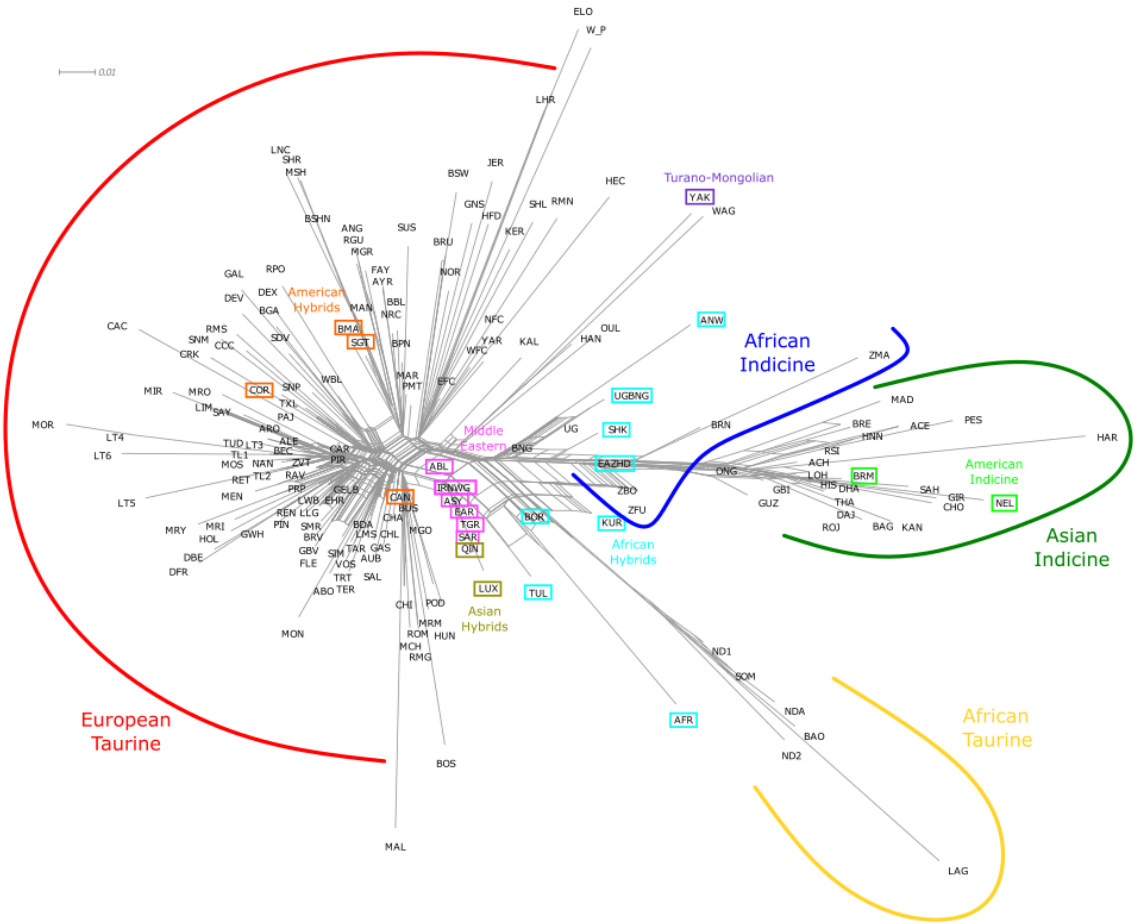


Figure 2.5. Neighbour-net using F_{ST} distances for 174 populations of *Bos taurus* and *Bos indicus* cattle, including known hybrids, with sample sizes greater than 1 (see Table S1 for label information). Scale for F_{ST} distance is displayed in the top left.

to the Eurasian taurine component. Creole cattle (CCC, CRK, RMS, SNM, SNP, TXL) were placed in the European taurine group, while the American indicine breeds (BRM, NEL) clustered along Asian indicine populations. Finally, the aurochs was placed between the European and Middle Eastern taurine groups.

2.4.2 Approximate Bayesian computation (ABC)

ABC was used to reconstruct the demographic history of taurine and indicine breeds. First, the reduced dataset (2K) was assessed, to qualitatively determine if the same admixture pattern as the larger dataset (8K). The CV error values for K from 1 to 5 were similar for both datasets (less than 1% difference for K = 3 and K = 4; Figure S2.2). For K = 4 each breed was assigned to one cluster with the ZebuAF showing admixture partially deriving from TaurAF but mostly from ZebuAS (Figure S2.3; Figure S2.4), while for K = 3 the clusters corresponded to TaurAf, TaurEU and ZebuAS. Both results reflected the divergence patterns observed in the population structure analyses.

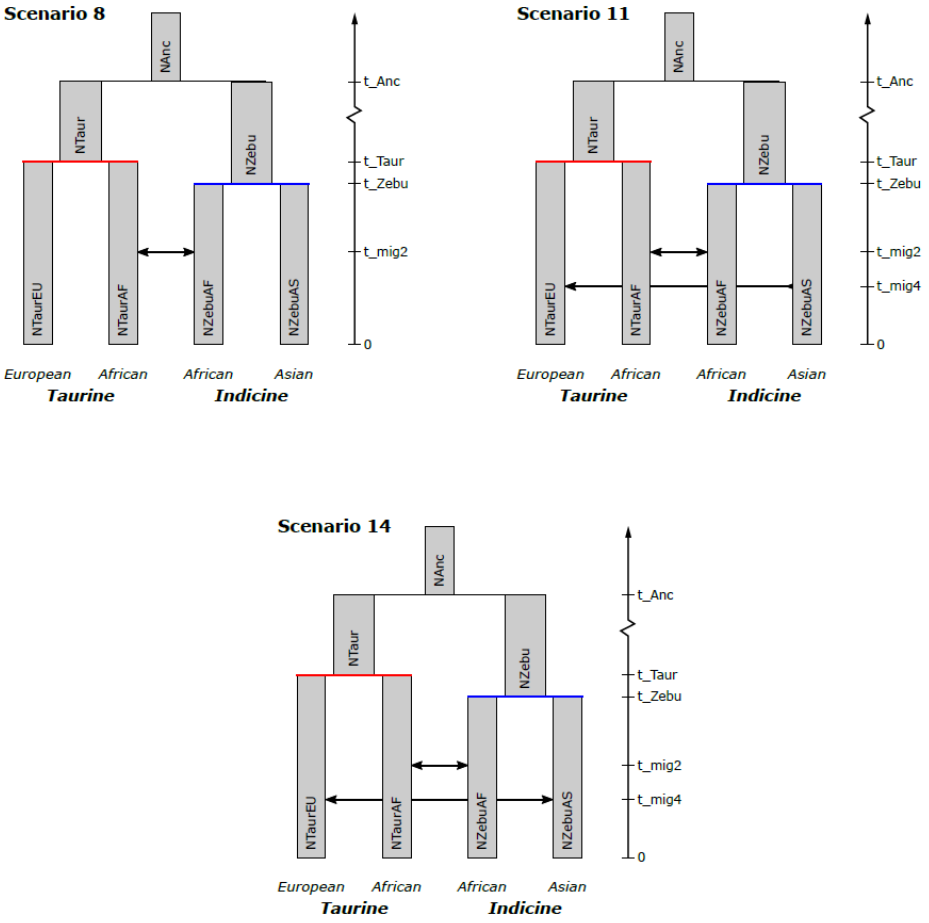


Figure 2.6. Schematic representation of demographic history modelled for taurine (*Bos taurus*) and indicine (*Bos indicus*) cattle using Approximate Bayesian Computation. The three models depicted are those with the highest support.

Of the 34 summary statistics originally chosen for the ABC analyses, half were removed because of correlations with other statistics, e.g., mean number of alleles per locus and pairwise differences between populations due to the high correlation with heterozygosity (Table S2.3; Figure S2.5). The remaining 17 summary statistics encompassed measurements of diversity (e.g., mean, and standard deviation of heterozygosity across loci for each population), and pairwise divergence (e.g., mean and standard deviation of F_{ST}; Table S2.3). It was ensured that all observed summary statistics fell within the 95% quantiles of simulated summary statistics before continuing as a final model.

The iterative methodology of producing additional models began by comparing BF between the first four scenarios, of which scenarios 1 and 2 had the highest support – both having only two modelled domestications (Table 2.2). Scenario 4 was less favoured; however, additional scenarios were still generated to test further test the hypotheses of three domestications. Despite additional models with three domestication events being tested, scenarios with only two domestications were better supported in almost all cases (Table 2.2). It should be noted that comparing MD between breed sets is not possible, however, the relative differences in MD between scenarios within each breed set remained largely consistent. Scenarios 8, 11 and 14 (Figure 2.6) produced consistently high magnitude MD (and thus BF > 3 when compared to most other scenarios) as well as high *p*-values, indicating the models fit the data well (emboldened in Table 2.2). Furthermore, there was difficulty distinguishing between these three models as BF was below 3. The models included only two domestication events, each occurring at the time of divergence within *B. taurus* and *B. indicus*, respectively, and bidirectional migration between the two African populations. Scenario 11 added migration from Asiatic indicine into European taurine, and scenario 14 made this migration bidirectional. Estimates of the posterior distributions of parameters such as the effective population size were relatively consistent across all three scenarios. These effective population size estimates suggest that following the divergence between the lineages that led the *B. taurus* and *B. indicus* (~250,000 YA) the ancestral taurine and indicine lineages grew demographically from ~10^3 until reaching effective population sizes of ~10^5 to 10^6, however, between ~3,600 and 7,900 generations ago these reduced until reaching ~10^4.

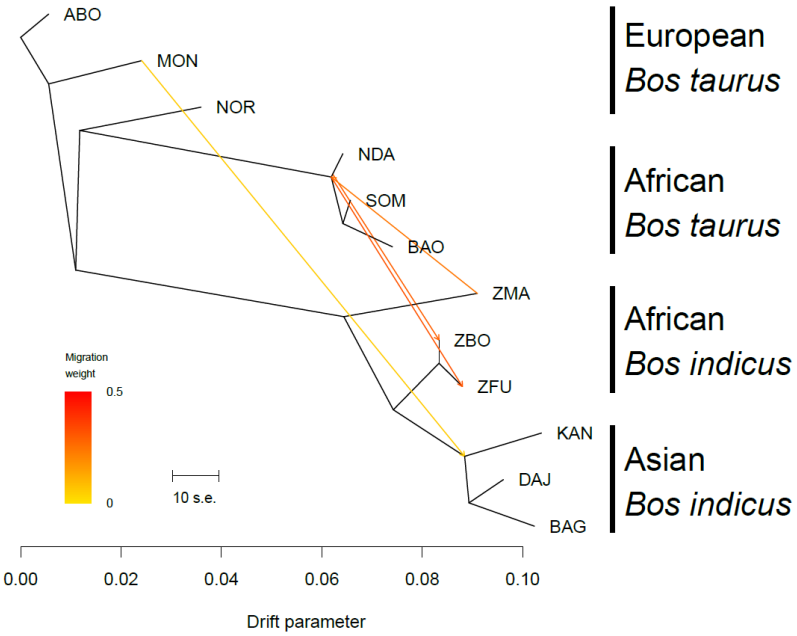


Figure 2.7. Phylogenetic network of the inferred relationships between 12 cattle breeds estimated using TREEMIX. Migration edges between breeds are shown as arrows pointing towards the recipient population and coloured according to the proportional ancestry received from the donor population. Scale bar is 10 times the mean standard error of the estimated entries in the covariance matrix.

The TREEMIX analysis recovered the expected phylogenetic relationships between the combined breed sets. As ABC model discrimination between scenarios 8, 11 and 14 was reliant on the specific migratory events between breeds selected in each breed set, migration edges added to depict primary movements of populations. Four migration edges were added before the model explained over 99.8% of the variance of the ancestry between the twelve populations (Figure 2.7). The first two edges were directed from TaurAF to ZebuAF, each weighted between 0.29 and 0.32 of the proportional ancestry received from the TaurAF into ZBO and ZFU. Additionally, a migration edge occurred in the reverse direction between ZebuAF (ZMA) and TaurAF at a similar strength (0.24). Finally, a weak (0.05) migration edge from TaurEU (MON) into the ancestor of ZebuAS was also observed.

Table 2.2. Model description and suitability for demographic history simulated for taurine (*Bos taurus*) and indicine (*Bos indicus*) cattle using Approximate Bayesian Computation. The three best scenarios with consistently high MD values across breed sets are emboldened. Models are shown in Figures 2, 5 and S2.1.

Scenario	Breed set 1[a]		Breed set 2[a]		Breed set 3[a]		Domestication Events	Scenario Description[c]
	MD[b]	*p-value*[b]	MD[b]	*p-value*[b]	MD[b]	*p-value*[b]		
1	13700.00	0.29	9810.00	0.49	17400.00	0.48	2	*B. t.* domestication at the time of *B. t.* divergence, *B. i.* domestication at the time of *B. i.* divergence.
2	1690.00	0.77	1720.00	0.46	803.00	0.50	2	*B. t.* domestication at time of *B. t.* divergence, *B. i.* domestication before *B. i.* divergence.
3	22.10	0.15	185.00	0.08	0.82	0.03	2	Domestications each before divergence of *B. t.* and *B. i.*
4	326.00	0.73	191.00	0.50	205.00	0.46	3	*B. t.* domestications after *B. t.* divergence, *B. i.* domestication before *B. i.* divergence.
5	0.51	0.08	1.46	0.13	1.32	0.07	3	Scenario 4 with ancient *B. t.* divergence (lower bound at 12,500 generations)
6	1060.00	0.97	728.00	0.53	686.00	0.77	3	*B. t.* domestications after *B. t.* divergence, *B. i.* domestication at the time of *B. i.* divergence.
7	5750.00	0.26	6020.00	0.27	6740.00	0.21	2	Scenario 1 with bidirectional migration before either divergence.
8	**578000.00**	**0.82**	**531000.00**	**0.66**	**238000.00**	**0.46**	**2**	**Scenario 1 with bidirectional migration between African *B. t.* and African *B. i.* after both divergences**.
9	329000.00	0.77	99100.00	0.58	10100.00	0.61	2	Scenario 8 with bidirectional migrations before either divergence.
10	245000.00	0.77	27400.00	0.55	32100.00	0.49	2	Scenario 8 with unidirectional migration from African *B. i.* to European *B. t.* after both divergences.

Scenario	Breed set 1[a]		Breed set 2[a]		Breed set 3[a]		Domestication Events	Scenario Description[c]
	MD[b]	*p-value*[b]	MD[b]	*p-value*[b]	MD[b]	*p-value*[b]		
11	**361000.00**	**0.86**	**592000.00**	**0.75**	**492000.00**	**0.70**	**2**	**Scenario 8 with unidirectional migration from Asian *B. i.* to European *B. t.* after both divergences.**
12	211000.00	0.78	328000.00	0.62	232000.00	0.50	2	Scenario 8 with unidirectional migration from African *B. i.* and Asian *B. i.* to European *B. t.* after both divergences.
13	2.37	0.00	599.00	0.14	22.40	0.06	2	Scenario 8 with constant ongoing bidirectional migration between *B. t.* and *B. i.* before either divergence
14	**194000.00**	**0.83**	**674000.00**	**0.84**	**621000.00**	**0.72**	**2**	**Scenario 8 with bidirectional migration between Asian *B. i.* and European *B. t.* after both divergences.**
15	124000.00	0.66	236000.00	0.68	364000.00	0.58	2	Scenario 14 with ancient bottlenecks within *B. t.* and *B. i.*

[a] Breed set 1: Normande (NOR), Somba (SOM), Zebu from Madagascar (ZMA), Kankraj (KAN); Breed set 2: Abondance (ABO), N'dama (NDA), Zebu Fulani (ZFU), Bhagnari (BAG); Breed set 3: Montbeliard (MON), Baoule (BAO), Zebu Bororo (ZBO), Dajal (DAJ).

[b] Marginal density (MD) and p-values are taken from the 1,000 simulations most similar to the observed data from 1 million simulations.

[c] Divergences here refer to events within Bos taurus (*B. t.*) and Bos indicus (*B. i.*) rather than the ancient divergence between them from *Bos primigenius primigenius*.

2.5 DISCUSSION

This worldwide survey on 180 populations using SNP array data depicted genetic ancestries, admixture, introgression, and migration patterns of cattle at a global scale and gained insights into the long-standing debate on the existence of a third domestication event from aurochs in northeast Africa. However, there are potential issues regarding ascertainment bias associated with the SNP arrays used in this study (Matukumalli *et al.*, 2009; Gautier *et al.*, 2010; Orozco-terWengel *et al.*, 2015), as reflected by the significantly higher genetic variation within taurine cattle when compared to indicine breeds. Higher H_o in European taurine populations and a potential overestimation of the inbreeding coefficient (F_{IS}) in indicine cattle would be an unexpected result had the dataset used an array not biased towards *B. taurus* variation (Table S2.1). Ascertainment bias was reduced through removing a proportion of the markers in LD, highly correlated markers with similar genealogical history were removed from the dataset, reducing multi-collinearity effects that would lead to overestimation of differentiation measurements. LD pruning of SNPs has been shown to be effective at reducing heterozygosity overestimations when compared to whole genome sequencing and, although this methodology underestimates F_{ST}, it is consistent regardless of relatedness to the ascertainment populations (Malomane *et al.*, 2018). Furthermore, replicating the ABC scenarios to increase the sampling scheme aimed to reduce any breed specific biases and more effective at capturing the demographic history representative of whole subgroups (Rougemont *et al.*, 2016).

Population structure analysis best supported the partition for K = 70; however, the low variation in CV error for K values above 15 suggests that additional clusters reflect only minor improvements in the resolution of proportional ancestry between breeds. This observation is likely to be an outcome of several factors influencing the statistical power to discriminate between the populations analysed, e.g., the recent divergence of many breeds compared to the genealogical depth observed across the whole dataset, hybridisation and gene flow, familial structure, and low sample sizes in some breeds (Kijas *et al.*, 2012; McTavish *et al.*, 2013; Orozco-terWengel *et al.*, 2015; Barbato *et al.*, 2017). Additionally, although breeds are often considered as phenotypically distinct populations, the shared genetic variation and incomplete lineage sorting that is resultant of recent divergences (only ~200 YA), makes it increasingly difficult to form high cluster resolution (Bradley *et al.*, 1996). An additional factor influencing the power to separate populations is the presence of gene flow between them, as reflected by the admixture signals in African indicine cattle. Deriving from Asiatic indicine populations introduced to Africa ~2,500-3,500 YA, African taurine cattle represent an important component of African indicine genetic diversity, probably reflecting gene flow between these populations due to their geographic

proximity, as well as deliberate hybridisation by farmers to incorporate local adaptations (e.g. trypanotolerance, heat tolerance) into the indicine animals following their more recent entry into Africa (MacHugh *et al.*, 2001). The differentiation between taurine and indicine was detected at two clusters, capturing the relatively ancient divergence between species and separate domestications (Loftus *et al.*, 1994). Additional clusters began separating geographically disparate groups within species, firstly, taurine split into African and Eurasia clusters and secondly indicine were split into African and Asiatic clusters (Figure 2.4). Each genetically differentiated group likely arising due to relative isolation by distance and distinct paths from the species-respective domestication centres. Interestingly, for low clusters (K = 2 to 4), African hybrids display similar proportional ancestry to African indicine. The introgression of indicine into African taurine cattle is predicted to have initially occurred ~4,000 YA, resulting in a consistently high indicine component within African hybrid cattle (Bahbahani *et al.*, 2017), possibly reinforced by modern admixture with African indicine.

The neighbour-net (Figure 2.5) and MDS (Figure 2.3) analyses identified similar patterns of divergence between indicine and taurine populations. The first two components of variation in the MDS analysis also identified the taurine vs. indicine (~31%) and African vs. Eurasian taurine (~15%) splits. While the division between the taurine groups explained much less genetic variance than the first component, it still explained a substantial proportion of the total (and more than twofold the third component), suggesting substantial differentiation between the two taurine groups, consistent with the hypothesis that the two groups harbour rather distinct genetic pools. In contrast to this, African indicine populations strongly resembled Asiatic indicine breeds. The relatively short distances between African indicine to both Asian indicine and African taurine populations in the neighbour-net suggest admixture between these groups. Furthermore, the central block of the neighbour-net depicts multiple complex connections among breeds, suggesting relatively recent divergence across many populations (Felius *et al.*, 2011). Despite being classified as *B. taurus*, Middle Eastern breeds occupy central positions both in the neighbour-net and MDS analyses, suggesting an admixed genetic background with an indicus component, as has been observed for several breeds from the Middle East and Fertile Crescent (Decker *et al.*, 2014; Karimi *et al.*, 2016). Consistent with the hypothesis of admixture of European taurine populations with local aurochs prior to their extinction approximately 400 YA (Achilli *et al.*, 2008; Park *et al.*, 2015; Upadhyay *et al.*, 2017), the *B. p. primigenius* sample in this study is placed close to Eurasian taurine groups in the MDS analysis. Although Decker *et al.* (2014) suggested a significant admixed aurochs ancestry for African taurine populations, with aurochs contributing up to 26% of their genomes, these current results failed to detect this influence, almost certainly because the British aurochs used is not an appropriate proxy for African aurochs

—often classified as *B. p. africanus* or *B. p. opisthonomous* (Clutton-Brock, 1989). Similarly to the admixture results (Figure 2.4), African hybrids from *B. taurus* and *B. indicus,* that were expected to be found in the central areas of the neighbour-net and MDS graphics, were placed near to the African indicine group, pointing towards a higher contribution of the African indicine gene pool in the formation of those hybrid breeds, as opposed to American hybrids that displayed higher influences from European taurine breeds. The genetic composition of Asian hybrids (QIN, LUX) was more balanced between Asian indicine and taurine origins. As expected, Creole cattle was placed closer to the European taurine cluster (Martínez *et al.*, 2012; Decker *et al.*, 2014; Sevane *et al.*, 2019), and American indicine cattle clustered with Asian indicine breeds (Orozco-terWengel *et al.*, 2015).

The substantial differentiation between the European and African taurine populations when compared to African and Asian indicine groups may be explained by several factors, such as lower gene flow between Africa and Europe than between Africa and Asia or a more recent divergence of indicine cattle. Another explanation may be a stronger effect of genetic drift in African taurine populations with respect to Eurasian taurine or indicine populations. Genetic drift would increase the rate of the change in allele frequencies in populations with a smaller effective population size (N_e), increasing the probability of accumulating differences between populations. An analysis of the trajectories of the effective population size over the last ~8,000 years shows that most cattle populations around the world exhibit relatively large effective population sizes at the beginning of the Holocene (~3,000 N_e); however, more recently all cattle populations have experienced a drastic decrease in N_e reaching modern N_e values of ~500 or less for many populations (Barbato *et al.*, 2015; Orozco-terWengel *et al.*, 2015), with the African taurine population showing, on average, larger N_e values than Eurasian taurine populations, albeit only marginally.

An additional domestication event in Africa from local aurochs has been also proposed to explain the large difference between taurine branches. The hypothesis of whether three domestication events better explain the genetic variation observed in modern domestic cattle when compared to the more widely accepted two-domestication scenario was explicitly tested. Demographic history incorporating a third domestication were successfully modelled with ABC, producing high posterior probabilities and simulated summary statistics proximal to observed values indicating the model reasonably reproduced the data (Wegmann *et al.*, 2010). Nonetheless, model discrimination strongly rejected all scenarios containing three domestications in favour similarly well-fitting models with only two (Table 2.2). The rejection of a hypothesised third domestication suggests an alternative explanation for the differences observed between African

and Eurasian taurine populations – admixture between migrating domestic cattle from the Middle East and indigenous African. Domestic cattle and their wild relatives occupied the same geographic regions for a long period of time, which raises the possibility that both taurine and indicine cattle naturally hybridised with aurochs, or that local farmers mixed them with local aurochs to restock their herds (MacHugh *et al.*, 2001; McTavish *et al.*, 2013). Uniparental loci such as mtDNA and Y-chromosome studies have generally underplayed the significance of admixture with wild aurochs, as haplotypes present in ancient DNA samples are often closely related but phylogenetically distinct from those in extant cattle samples (MacHugh *et al.*, 2001; Götherström *et al.*, 2005; Edwards *et al.*, 2010; Schibler *et al.*, 2014). Nonetheless, there is evidence of gene flow from wild aurochs prior to the extinction (~400 YA) into extant cattle in areas such as Italy, Iberia, southern Europe and the British Isles (Achilli *et al.*, 2008; Park *et al.*, 2015; Upadhyay *et al.*, 2017); interestingly, even the presence of taurine mitochondrial haplotypes have been identified in a Scandinavian medieval drinking horn made from an aurochs (Bro-Jørgensen *et al.*, 2018). Historical introgression into domestic populations from their wild counterparts has occurred across a breadth of animals, including horses (Cothran *et al.*, 2011), pigs (Frantz *et al.*, 2015), camels (Almathen *et al.*, 2016) and sheep (Cao *et al.*, 2021) often as a strategy to introduce local adaptation to the farmed population.

The presence of shared genomic regions between African taurine and African aurochs, such as the T1 mitochondrial haplogroup, has been interpreted as evidence for a more ancient divergence between European and African taurine populations (Bradley *et al.*, 1996; Bonfiglio *et al.*, 2012). While the current results refute this hypothesis in favour of local introgression from African aurochs post domestication, it highlights the difficulty in disentangling genomic introgression from shared ancestral variation. It is important to note that no explicit modelling of introgression from aurochs was modelled with ABC, thus the model posteriors may have been affected. The most notable consequences would include overestimates of N_e within the diverged African taurine population as heterozygosity would be inflated, or an earlier divergence between African and European taurine due to the more ancient coalescence of genomic regions originating from aurochs (Gray *et al.*, 2014). It is unclear if this had a significant impact on our estimates, as the divergence time within taurine for the three best-fitting scenarios (Figure 2.6) was estimated (using lower and upper quartiles of the 3 biological replicates for scenarios 8, 11, and 14) between 2965 and 8593 generations ago (Table S2.4), spanning reasonable estimates for the expected divergence (Bruford *et al.*, 2003; Ajmone-Marsan *et al.*, 2010; MacHugh *et al.*, 2017). A possible explanation for this is the heavily reduced SNP dataset used in the ABC modelling, potentially avoiding loci heavily influenced by introgression from aurochs. Future testing of the demographic models including higher density SNP data, an African aurochs population, and explicitly modelling

wild to domestic introgression would produce more accurate estimates of N_e, taurine divergence time, and allow identification of introgressed loci.

Previous studies using SNP data genotyped in taurine and indicine cattle have found an African taurine component in European taurine cattle reaching an average of 10% of their genotype, while between 5% and 10% of the genetic background in some European breeds is of indicine origin (Gautier *et al.*, 2010; McTavish *et al.*, 2013). This shared genetic variation may have originated in (or before) ancestral *Bos primigenius* populations, subsequent lineages may experience differential selection at a given site independent of the connectivity or relatedness of the resultant populations. This is hypothesised to have affected some variants present at high frequency in Turano-Mongolian (*Bos taurus*), African taurine and Asian indicine cattle, yet are at low frequency in European taurine (Buggiotti *et al.*, 2021). The identification of shared alleles with more distantly related *Bos* species – for example Yak (*Bos grunniens*) or Gaur (*Bos gaurus*) – could offer support to this mechanism within the current data (Naji *et al.*, 2021). Additionally, ABC was used to test alternative migration hypotheses between the sets of African and Eurasian taurine and indicine breeds used to model the domestication events. While two domestications fit the data better, it was not possible to further differentiate between a scenario with bidirectional gene flow between African breeds, and two alternative ones that also included unidirectional or bidirectional gene flow between the Eurasian breeds. This could be due the difficulty in accurately reconstructing demographic history and model differentiation, for example divergence and gene-flow can be indistinguishable to isolation followed by secondary contact, or brief bottlenecks remaining unidentified if preceded by large expansions due to the increased genetic diversity (Gray *et al.*, 2014; Rougemont *et al.*, 2016). These limitations restrict the intrinsic complexity in the models, ideally, events such as divergences and domestication would be reconstructed as gradual processes occurring over many generations as cattle were shifted from prey animals, to managed herds, to captive bred stock (Larson and Burger, 2013). Unfortunately, the increasing ABC model complexity also risks over-parameterisation as increasing the number of priors can exponentially increase the parameter combinations for the model, thereby reducing the accuracy of posterior estimates (Wegmann *et al.*, 2010). The maximum likelihood tree generated in TREEMIX recaptured the complex gene flow among the breed sets, with the first three migration edges mirroring the same migratory patterns between TaurAF and ZebuAF first defined in scenario 8 (Figure 2.6; Figure 2.7) and explained up to 99.78% of the variance in the tree. Adding a fourth migration edge marginally increased the explanation of the variance in the tree and corresponded to a weak migration edge from TaurEU into ZebuAS. Overall, this supports the ABC results and indicates that the more impactful gene flow occurred between *B. taurus* and *B. indicus* within

Africa, although minor movements of cattle between Europe and Asia may also contributed to shape the population structure that is seen today.

While SNP array data is an affordable way to collate genomic overview of the global cattle population, the high ascertainment bias of the current arrays introduces a lot of uncertainty defining the variation and demography of certain breeds, particularly indicine. Utilising whole genome sequence data independently sourced or through the growing 1000 Bull Genome Project database (Hayes and Daetwyler, 2019) would suffers less ascertainment bias than the current SNP arrays do, by capturing breed- and species-specific variation, thus allowing for more precise interrogation into demographic and introgressive processes. Furthermore, increasing flexibility of software facilitates the processing of larger datasets, for example Jiang (*et al.*, 2021) implemented similar demographic analysis on ducks using almost 4 million SNPs derived from whole genome sequencing with ABCTOOLBOX (Wegmann *et al.*, 2010) V2.0, suggesting little adjustment to the current pipeline would have to be made. Additionally, these techniques are transferrable to other systems or breeds – notably, mirror such analyses with focus on the Turano-Mongolian cattle to model the demographic history and the hypothesised additional domestication event (Zhang *et al.*, 2013). While it would be difficult to avoid over-parameterisation and obtain meaningful results by adding yet another extant population to the four populations tested above, a separate analysis incorporating Turano-Mongolian cattle, Asian indicine and African taurine cattle would provide valuable reference populations to test the domestication hypothesis. Furthermore, the addition of European taurine and contrasting models of migration may provide evidence for the mechanism (differential selection of ancestral variation, or, introgression) of differential variation observed between the former three groups and European taurine (Yurchenko *et al.*, 2018).

In conclusion, SNP array data collected in more than three thousand cattle samples belonging to 180 populations with a world-wide geographical distribution encompassing Africa, Europe, Asia and the Americas continents, provided a comprehensive picture of genetic diversity, population structure and demographic dynamics of cattle populations at a global level. The analyses confirm the large differentiation between African and Eurasian taurine and the high levels of admixture in African indicine cattle from both Asian indicine and African taurine cattle, revealing a higher contribution from African indicine genetic origin in the formation of African hybrids. Contrastingly, American hybrids, such as Creole cattle, exhibited a higher influence from taurine breeds. Modelling the domestication history of cattle using approximate Bayesian computation consistently favoured scenarios involving only two domestication events, discarding a third *B. p. primigenius* domestication in Egypt and suggesting the subsequent hybridisation from local aurochs to explain the additional genetic variation detected. Paleogenomic analyses of

Middle Eastern and African wild aurochs pre-dating domestication and early Middle Eastern and African domestic cattle will provide the data required address these questions. Further analysis utilising a wider panel of ancient DNA from extinct local aurochs and ancestral individuals of domestic populations will help to disentangle the complex human-mediated microevolution of domestic cattle.

2.6 ACKNOWLEDGEMENTS

Thank you to Licia Colli and Ezequiel Nicolazzi for their assistance with filtering and quality control of data from Orozco-terWengel *et al.* (2015). Similarly, thank you to Stephen Park and David MacHugh for their assistance in converting whole genome aurochs data from (Park *et al.*, 2015) for SNP array compatibility. A special thanks to Natalia Sevane for her feedback. The initial part of the ABC analysis was carried out for submission as part of my final year undergraduate project (BSc) at Cardiff University, however, these analyses were significantly extended and developed for this chapter.

Chapter Three

Demography and rapid local adaptation shape Creole cattle genome diversity in the tropics

---

Photo by Rodrigo Martinez

3.1 ABSTRACT

The introduction of Iberian cattle in the Americas after Columbus' arrival imposed high selection pressures on a limited number of animals over a brief period of time. Knowledge of the genomic regions selected during this process may help in enhancing climatic resilience and sustainable animal production. First, taurine (*Bos taurus*) and indicine (*Bos indicus*) contributions to the genomic structure of modern Creole cattle was determined. Second, their demographic history using approximate Bayesian computation (ABC), linkage disequilibrium (LD) and N_e Slope (NeS) analyses was inferred. Third, whole genome scans for selection signatures based on cross-population extended haplotype homozygosity (XPEHH) and population differentiation (F_{ST}) was performed to disentangle the genetic mechanisms involved in adaptation and phenotypic change by a rapid and major environmental transition. To tackle these questions, SNP array data (~54,000 SNPs) in Creole breeds was combined with data from their modern putative Iberian ancestors. Reconstruction of the population history of Creoles from the end of the 15th century indicated a major demographic expansion until the introduction of zebu and commercial breeds into the Americas ~180 years ago, coinciding with a drastic N_e contraction. NeS analysis provided insights into short-term complexity in population change and depicted a decrease/expansion episode at the end of the ABC-inferred expansion, as well as several additional fluctuations in N_e with the attainment of the current small N_e only towards the end of the 20th century. Selection signatures for tropical adaptation pinpointed the thermoregulatory slick hair coat region, identifying a new candidate gene (*GDNF*), as well as novel candidate regions involved in immune function, behavioural processes, iron metabolism and adaptation to new feeding conditions. The outcomes from this study will help in future-proofing farm animal genetic resources (FAnGR) by providing molecular tools that allow selection for improved cattle performance, resilience, and welfare under climate change.

3.2 INTRODUCTION

Until recently, selection has occurred at a relatively slow rate in cattle and has been largely passive, driven by adaptations to diseases, dietary variation and local climatic patterns (Russell, 2007). Since the domestication of cattle more than ~10,000 years ago (YA; Bruford *et al.*, 2003), farmers started to artificially breed animals with preferred phenotypes, although it was not until ~200 YA that European farmers began the formation of closed herds which developed into modern breeds (Taberlet *et al.*, 2011). Anthropogenic long-distance transportation of livestock

has forced even higher selective pressures on a limited number of domestic animals concentrated over brief time periods, one example of which is the introduction of Iberian livestock species in the Americas. After the first arrival of cattle on the tropical Caribbean Island Hispaniola in 1493, Creole livestock started to evolve into distinct ecotypes specifically adapted to a variety of environments and production systems. From this location, as well as reinforcements from Iberia and the Atlantic archipelagos during the 16th century, cattle populations expanded and spread throughout the Americas (Villalobos Cortés *et al.*, 2009), starting from an estimated founding stock below 1,000 individuals (Rodero *et al.*, 1992). Introductions of northern European cattle into North America were also reported between 1608 and 1640 (Felius *et al.*, 2014). After three centuries featuring the predominance of Creole cattle, population declines started with the introduction of other cattle around the middle of the 19th century, better suited to more intensive production and breeding systems (Willham, 1982). The introduction of European breeds (poorly adapted to the tropics but normally highly productive) and zebus (highly adapted to the tropics, but normally not as productive) resulted in the substitution of Creoles by a series of less adapted, admixed, or commercial populations, displacing them into marginal areas.

Reconstructing the demographic history of Creole populations is therefore key to disentangling American livestock colonization dynamics and can contribute to a better understanding of the genomic signatures of breed evolution. Additionally, ongoing climate change is likely to lead to reductions in animal production and welfare in the future, which makes an understanding of the genomic regions selected under the major and rapid environmental changes imposed on Creole cattle, a useful tool for enhancing resilience and sustainable production in the short term. Therefore, the aims of this study were first to determine the contributions of different taurine (*Bos taurus*) and indicine (*Bos indicus*) ancestors on the genomic make-up of Creole cattle. The second aim was to infer the demographic history of Creole cattle populations by combining different approaches to investigate trends in effective populations size (N_e): approximate Bayesian computation (ABC; Wegmann *et al.*, 2010); linkage disequilibrium (LD) structure (SNeP; Barbato *et al.*, 2015); and N_e Slope analysis (NeS). Finally, the third aim was to perform a whole genome scan for signatures of selection based on cross-population extended haplotype homozygosity tests (XPEHH; Sabeti *et al.*, 2007) and population differentiation (F_{ST}; Wright, 1949). To tackle these questions, SNP array data in modern Creole cattle was combined with modern day samples from breeds comprising their putative Iberian ancestors. Identifying genomic regions responding to these selection pressures provide valuable tools for improving cattle resilience, performance, and welfare under climate change.

3.3 MATERIALS AND METHODS

3.3.1 CATTLE POPULATIONS AND SNP ARRAY DATA

The data set comprised SNP array data from 412 individuals genotyped using the Illumina BovineSNP50 array versions 1 and 2, and the Bovine High Density BeadChip (Table S3.1; The Bovine HapMap Consortium *et al.*, 2009; Decker *et al.*, 2009; Gautier *et al.*, 2010; Decker *et al.*, 2014; Upadhyay *et al.*, 2017). Twenty-nine animals were newly genotyped using the Illumina BovineSNP50 version 2 and Geneseek Genomic Profiler Bovine 150k (Table S3.1). The dataset included six Creole populations adapted either to tropical humid (three Colombian breeds: Costeño con Cuernos, Romosinuano, San Martinero; a North American breed: Florida Cracker; and a Caribbean breed sampled in Brazil: Senepol) or dry conditions (Texas Longhorn). The main breeds comprising their putative Iberian ancestors included: (i) six different lineages of Lidia, a breed that has not been selected for productivity traits and may be the most representative modern descendent of Iberian cattle herds back in the 15th century, retaining high genetic variability among lineages; (ii) Mostrenca, Retinta, Berrenda en Colorado, Cárdena Andaluza and Pajuna breeds, distributed throughout central and southern Iberia; and (iii) Asturiana de los Valles and Cachena, reflecting the northern Iberian genomic pool. The remaining breeds represent a hypothesized African taurine influence on Creole cattle (Baoule, Lagune, N'Dama, Somba; Miretti *et al.*, 2004), representatives of commercial European stock introduced to the Americas around the middle of the 19th century (Angus, Red Poll, Holstein, Jersey, Shorthorn) and potential indicine introgression into Creole cattle from tropical areas (Brahman, Nelore, Gir). SNP array data were merged and filtered using PLINK v1.90 (Purcell *et al.*, 2007; Chang *et al.*, 2015) and the genomic positions for each SNP was mapped to the UMD3.1 bovine assembly (RefSeq:GCF_000003055.5). Only autosomal SNPs with a minor allele frequency (MAF) above 1% and a call rate of at least 90% across all breeds were retained for downstream analyses, leaving 33,342 SNPs. The data set was then phased with BEAGLE v3.3.2 (Browning and Browning, 2007).

3.3.2 ESTIMATION OF AUTOSOMAL ANCESTRY PROPORTIONS AND POPULATION DIVERGENCE IN CREOLE CATTLE

To determine the relative contribution of different potential taurine and indicine ancestors on the genomic structure of Creole cattle, population admixture analysis was carried out using the software ADMIXTURE v1.3 (Alexander *et al.*, 2009) with 2,000 bootstraps for eight population clusters (K), corresponding to the African (two clusters), Iberian, Angus, Shorthorn, Holstein and

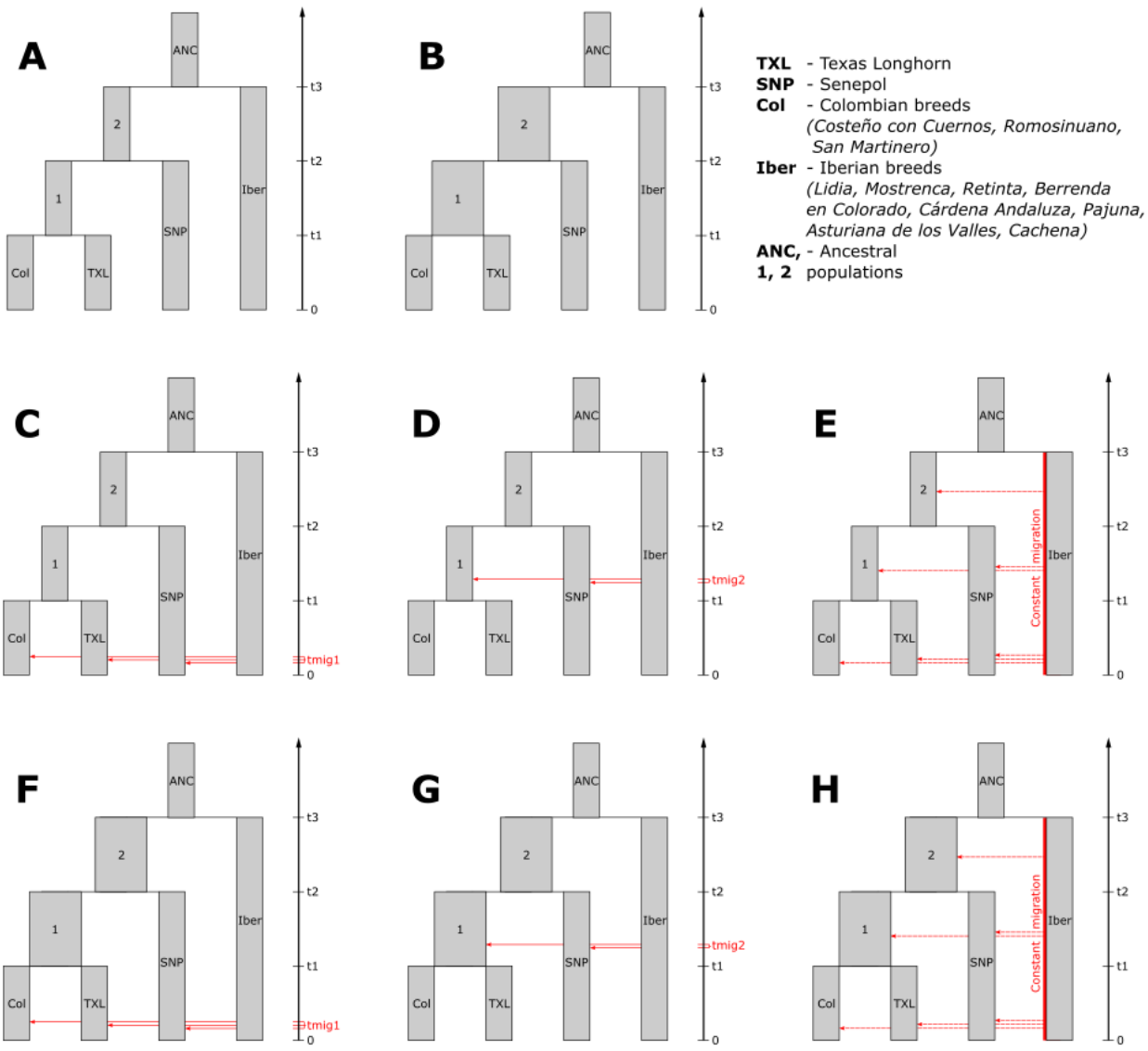


Figure 3.1. Modelled scenarios for reconstructing Creole cattle demographic history using approximate Bayesian computation (ABC). (a) Scenario 1: main model of cattle dispersion throughout the Americas. (b) Scenario 2: variation that includes expansions in Creole populations at t2 and t3. (c) Scenario 3: variation that includes recent migration. (d) Scenario 4: variation that includes migration before t1. (e) Scenario 5: variation that includes ongoing migration. (f) Scenario 6: variation that combines scenarios 2 and 3. (g) Scenario 7: variation that combines scenarios 2 and 4. (h) Scenario 8: variation that combines scenarios 2 and 5.

Jersey taurine ancestries, as well as the Asian zebu ancestry. For this analysis, autosomal SNP array data were further pruned for LD higher than 0.1 using a sliding window approach of 50 SNPs and a step size of 10 SNPs. The clustering solutions were visualised using the POPHELPER package in R (Francis, 2017; R Core Team, 2018).

Multidimensional scaling (MDS) was implemented using Hamming distances across 20 dimensions using PLINK. The first two (major) axes were visualized using R (R Core Team, 2018).

A Reynolds' distance matrix was estimated between population pairs using ARLEQUIN V3.5 (Excoffier and Lischer, 2010), and a neighbour-net tree was constructed in SPLITSTREE V.4.14.4 (Huson and Bryant, 2006).

3.3.3 DEMOGRAPHIC ANALYSIS

The population history of Creole cattle was reconstructed from the late 15th century to the present day using approximate Bayesian computation (ABC) as in Chapter 2. Briefly, a subset of the data was divided into four clusters: Col including all Colombian breeds (Costeño con Cuernos, Romosinuano, San Martinero), Senepol, Texas Longhorn and Iber (for all Iberian breeds). Eight alternative demographic histories were modelled based on historical records, results from Admixture, MDS and neighbour-net analyses, and prior N_e estimates obtained with SNeP. The scenarios included a model of Creole cattle dispersal throughout the Americas and variations of this model accounting for population expansions and alternative migration patterns representing restocking from Iberian populations (Figure 3.1). One million reverse coalescent simulations were generated for each of the eight scenarios with FASTSIMCOAL2 (Excoffier and Foll, 2011; Excoffier *et al.*, 2013) using a pipeline implemented in ABCTOOLBOX (Wegmann *et al.*, 2010), with a required computation time of eight days per scenario splitting simulations in ~50 parallel runs. Seventeen summary statistics were calculated in ARLEQUIN V3.5 (Excoffier and Lischer, 2010) for simulated and observed data (Table S3.2). A Spearman's rank correlation was calculated between each pair of summary statistics in R, and statistics with consistently high negative or positive correlation were removed (Figure S3.1, Table S3.2). ABCTOOLBOX was used to perform rejection sampling on the simulated data set, retaining the 5,000 (0.5%) simulations that closest fit to the observed data for each of the eight scenarios. Marginal density (MD) and posterior probability *p*-values (i.e., the proportion of simulations that have a smaller or equal likelihood to the observed data) were calculated from the retained simulations after a postsampling regression adjustment using a general linear model. Bayes factors (BF) were calculated between scenarios by taking the quotient of the MD from two scenarios to choose the best modelled scenario fitting the data; if BF > 3, the alternative scenario can be rejected (Wegmann *et al.*, 2010).

To examine the most recent changes in N_e, the software SNEP V1.11 (Barbato *et al.*, 2015) was used to estimate the demographic history for each population by the relationship between LD and N_e up until approximately 13 generations in the past. Default options were used apart from sample size correction for unphased genotypes and correction to account for mutation (Sved and Feldman, 1973). To identify subtle changes in the inferred N_e curve that might be diagnostic of changes in N_e not visually explicit when observed in the N_e plot, an "N_e Slope analysis" (NeS) was used to investigate the rate and directionality of N_e changes occurring in recent generations

(Figure S3.2). Mario Barbato derived the formulae for NeS and assisted with running the analysis. The slope of each segment linking pairs of neighbouring N_e estimates was first calculated and then normalised using the median of the two most proximal past N_e slope values as in:

$$N_e S_n = \left(S_n - \tilde{X}_n\right)\left(1 + \tilde{X}_n\right)^{-1}$$

where S_n is the slope of the n^{th} pair of neighbouring N_e estimates and

$$\tilde{X}_n = \text{med}\{S_n, S_{n+1}, S_{n+2}\}$$

3.3.4 SELECTION SIGNATURES

Recently generated selection signatures were scanned for to characterize differences observed between breeds that have remained in the Iberian Peninsula and those that colonized the Americas. Four Creole clusters were selected using the Admixture, MDS and neighbour-net results, one group (Col) including the three Colombian breeds (Costeño con Cuernos, Romosinuano, San Martinero) and three other breeds from the Americas, Florida Cracker, Senepol and Texas Longhorn. All pairwise comparisons were analysed between these four Creole clusters and three Iberian clusters used as biological replicates: (i) IB1, including Retinta, Berrenda en Colorado and Cachena; (ii) IB2, including Cárdena Andaluza, Asturiana de los Valles, Pajuna and Mostrenca; and (iii) a third group (LID), including the six Lidia lineages. The data set was separated per breed using VCFTOOLS v0.1.15 (Danecek *et al.*, 2011), and haplotype reconstruction was carried out using BEAGLE. All missing data were removed from the merged data set of the four groups using VCFTOOLS, leaving 15,375 SNPs.

Recent selective sweeps were identified in the Creole populations using XPEHH (Sabeti *et al.*, 2007) within the software SELSCAN v1.1.0b (Szpiech and Hernandez, 2014) with the IB1, IB2, and LID groups as references. The maximum distance between adjacent SNPs was 500 kb to allow for inconsistencies in bovine SNP arrays, whereas the remainder of the settings were left as default. The XPEHH scores were standardized across the whole genome. XPEHH scores exceeding the extreme 1% of the standardized distribution were identified as potential locations for positive selection in each given Creole cluster. All significant SNPs of a Creole breed validated with at least two Iberian clusters were merged regardless of the Iberian ancestral group to account for breed specific selection signatures. Contiguous significant SNPs were integrated to a common signature or region within each breed, allowing for up to one non-significant SNP in the middle, and including half of the physical distance to the neighbouring non-significant marker on both sides. As XPEHH searches for unusually long haplotypes, isolated significant SNPs were discarded, rendering this analysis conservative.

Selection signatures expected to have been generated prior to colonization of the Americas were explored using F_{ST} outliers compared to the null distribution generated in non-overlapping windows of 500 kb using VCFTOOLS. Windowed F_{ST} was used as a test statistic, retaining windows with values exceeding the 99% upper quantile as potential locations for selection. Given that F_{ST} analysis is not directional, that is does not differentiate between Creole or Iberian signatures of selection, only windows validated in the three Iberian replicates were consider for downstream analysis to isolate signals detected only in Creole cattle.

3.3.5 ANCESTRY ESTIMATION AT CANDIDATE REGIONS

LOCAL ANCESTRY IN ADMIXED POPULATIONS (LAMP) V2.5 (Paşaniuc *et al.*, 2009) kindly run by Natalia Sevane was used to estimate the ancestry proportions (Iberia, commercial, Africa and zebu) of Creole breeds at candidate regions. The LAMPANC method was applied for inferring the locus-specific ancestries providing the genotypes of the ancestral populations. Autosome-wide Creole ancestry proportions of 76% Iberian, 12% commercial and 3% African taurine groups, and 9% zebu cattle were estimated from the Admixture proportions α. An estimated number of 83 generations was set for the beginning of admixture in Creole cattle taking into account the introductions of North-European cattle in North America between 1608 and 1640 (Felius *et al.*, 2014), assuming an average generation length of 5 years, and otherwise using default parameters (Gautier *et al.*, 2016). The average excess/deficiency in the different ancestries was calculated by subtracting the average estimated ancestry at each significant SNP within candidate regions from the average estimated ancestry of all SNPs.

3.3.6 GENE ONTOLOGY

Gene ontology (GO) analyses were carried out on the annotated gene sets included in the genomic regions found to be under selection in the Colombian, Florida Cracker, Senepol and Texas Longhorn breeds using the Functional Annotation Cluster (FAC) tool from the DATABASE FOR ANNOTATION, VISUALIZATION AND INTEGRATED DISCOVERY (DAVID) V6.8 (Huang *et al.*, 2009). DAVID determines significantly enriched terms between a user-inputted and a reference list of genes using high stringency ease scores, effectively a more conservative Fisher's exact test. Enriched terms may indicate biological functions or processes positively selected in a breed. KEGG pathway analyses were also performed in DAVID to map clusters of genes involved in common pathways. In addition, the Bovine QTL Animal database (http://www.animalgenome.org) was used to identify any overlap with quantitative trait loci (QTL) described in the literature.

3.4 RESULTS

3.4.1 AUTOSOMAL ANCESTRY PROPORTIONS AND POPULATION DIVERGENCE IN CREOLE CATTLE BREEDS

When the number of clusters was set to eight in the Admixture analysis, it identified clusters formed by zebu (*B. indicus*), two *B. taurus* African clusters, a cluster formed by the Iberian breeds (*B. taurus* Iberia), as well as separate clusters for the main four commercial *B. taurus* (Angus, Shorthorn, Holstein and Jersey), while ancestry contributions to Creole populations ascribed the major genomic component to Iberian ancestry (0.76 ± 0.06 SD), with minor influences from zebu and European commercial breeds (Table 3.1; Figure 3.2), in concordance with previous studies (Martínez *et al.*, 2012; Decker *et al.*, 2014). Among Creole breeds, the Florida Cracker displayed the highest level of introgression from commercial genomes (0.36 ± 0.07 SD), mainly from Jersey, Angus and Shorthorn, whereas the Indicine component was higher in Senepol (0.15 ± 0.02 SD) and Romosinuano (0.10 ± 0.02 SD).

Table 3.1. Average taurine and indicine ancestries in Creole cattle breeds.

Breed	***B. taurus* Iberia** Mean ± SD	***B. taurus* commercial** Mean ± SD	***B. taurus* Africa** Mean ± SD	***B. indicus*** Mean ± SD
Costeño con Cuernos	0.80 ± 0.08	0.07 ± 0.06	0.04 ± 0.01	0.09 ± 0.03
Florida Cracker	0.60 ± 0.03	0.36 ± 0.07	0.01 ± 0.01	0.03 ± 0.02
Romosinuano	0.80 ± 0.06	0.07 ± 0.04	0.03 ± 0.01	0.10 ± 0.02
San Martinero	0.86 ± 0.06	0.04 ± 0.03	0.05 ± 0.01	0.06 ± 0.04
Senepol	0.69 ± 0.05	0.14 ± 0.04	0.02 ± 0.01	0.15 ± 0.02
Texas Longhorn	0.81 ± 0.07	0.06 ± 0.03	0.05 ± 0.01	0.08 ± 0.06
Mean	0.76 ± 0.06	0.12 ± 0.05	0.03 ± 0.01	0.09 ± 0.03

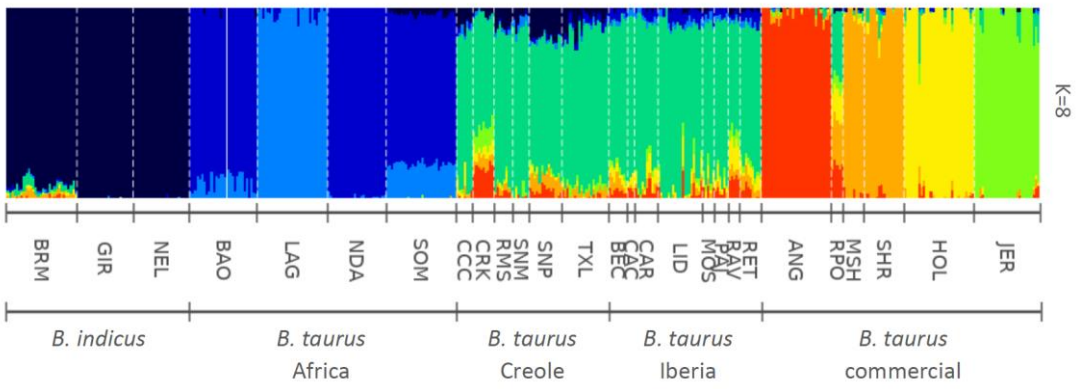


Figure 3.2. Ancestry proportions in Creole breeds at $K = 8$. Complete breed names are included in Table S1.

Multidimensional scaling allocated ~25% and ~21% of the variance to the first two axes, respectively, which separated taurine from zebu cattle breeds, and African taurine from the remaining populations (Figure 3.3). Among the relationships displayed by Creole, Iberian, and commercial breeds, Senepol showed the highest differentiation, driven by the influence of zebu breeds, and Florida Cracker was grouped most closely with the commercial breeds. These results were supported by the neighbour-net analysis, which clustered the breeds into five main groups (zebu, Africa, commercial, Iberia and Creole), with Florida Cracker intermediate between Iberian and the commercial breeds (Figure 3.4).

3.4.2 DEMOGRAPHIC HISTORY

ABC modelling was used to explore the recent demographic history of Creole cattle from the arrival of the first individuals to the Americas at the end of the 15th century to present. Thirteen summary statistics were retained after removing correlated measurements (Figure S3.1, Table S3.2). All observed summary statistics were within the 95% quantiles of the simulated summary statistics for each scenario. Comparison of the different scenarios showed a BF > 3 between scenarios 2, 6 and 7 and all the others (Table 3.2). Among the three best fitting scenarios, scenario 2 displayed the highest MD value, with a BF of 1.4 and 1.7 when compared with scenarios 6 and 7, respectively (Table 3.2). Scenario 2 supports the participation of a small number of animals (84) in the development of American breeds, followed by a major expansion up to a N_e of 57,278 by 180 YA, that later collapsed to the reduced population sizes detected nowadays, ranging between 497 for Senepol and 638 for Texas Longhorn (Table 3.3, Figure S3.3). Higher N_e values were retrieved for the Colombian (755) and Iberian (2,577) breeds derived from

Table 3.2. Approximate Bayesian computation (ABC) results for the different scenarios (shown in Figure 1) modelling Creole cattle demographic history.

Scenario	*P*-value	Marginal density	Bayes factor							
			Sc. 1	Sc. 2	Sc. 3	Sc. 4	Sc. 5	Sc. 6	Sc. 7	Sc. 8
Sc. 1	0.42	308.1	-	0.05	0.66	0.75	3.31	0.08	0.09	2.16
Sc. 2	0.67	5627.8	18.27	-	12.06	13.75	60.45	1.41	1.69	39.38
Sc. 3	0.56	466.5	1.51	0.08	-	1.14	5.01	0.12	0.14	3.26
Sc. 4	0.38	409.2	1.33	0.07	0.88	-	4.40	0.10	0.12	2.86
Sc. 5	0.52	93.1	0.30	0.02	0.20	0.23	-	0.02	0.03	0.65
Sc. 6	0.82	3993.2	12.96	0.71	8.56	9.76	42.89	-	1.20	27.94
Sc. 7	0.69	3324.4	10.79	0.59	7.13	8.12	35.71	0.83	-	23.26
Sc. 8	0.42	142.9	0.46	0.03	0.31	0.35	1.53	0.04	0.04	-

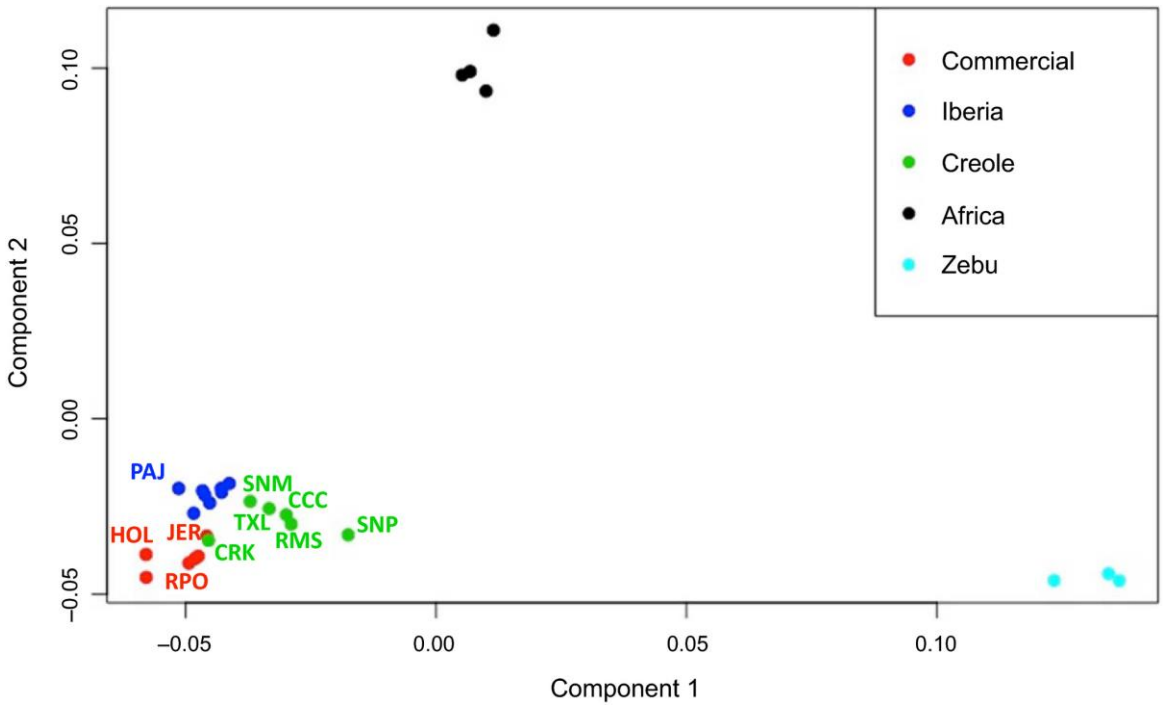


Figure 3.3. Multidimensional scaling (MDS) plot for 27 taurine and indicine cattle populations.

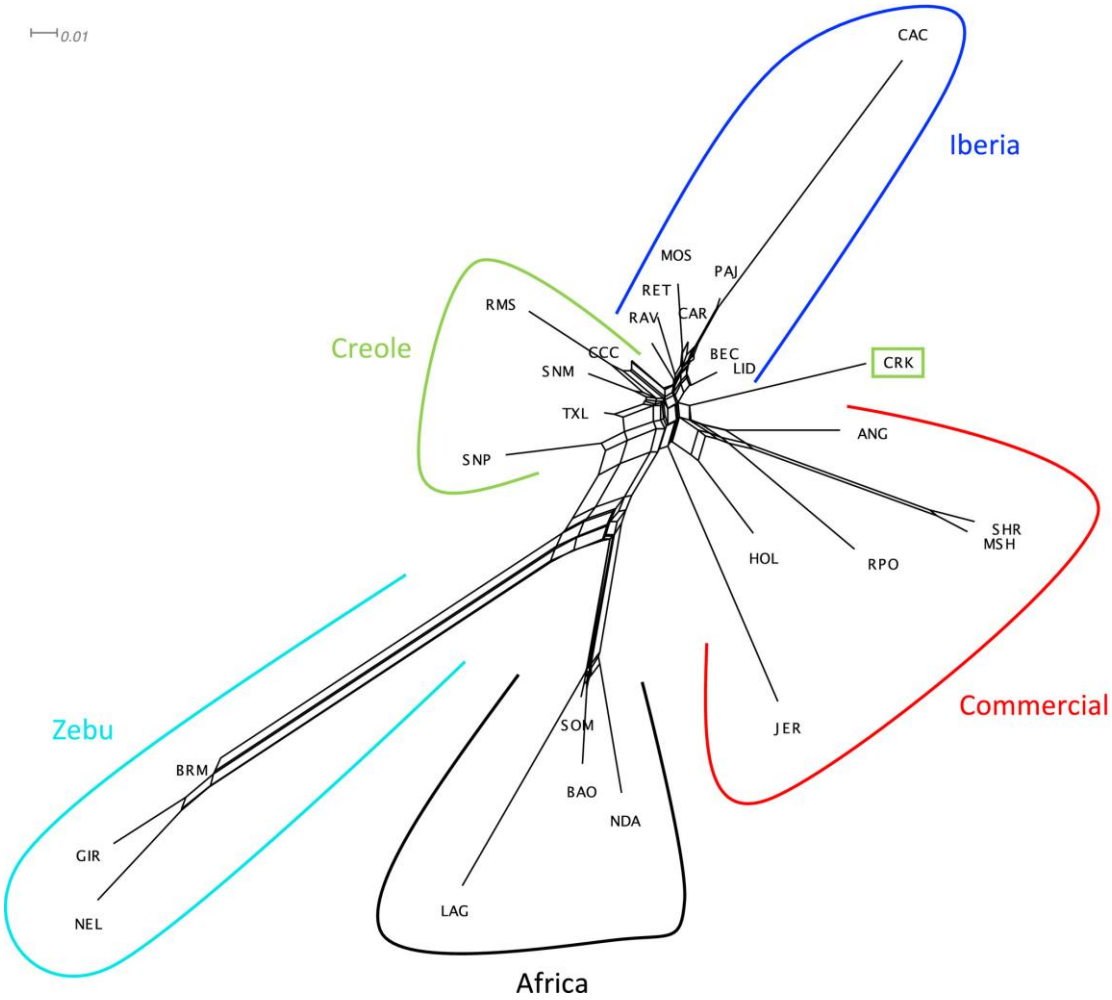


Figure 3.4. Neighbour-net using Reynolds' distances for 27 taurine and indicine cattle populations. Scale for Reynolds' distance is displayed in the top left.

the grouping of three and eight populations, respectively, which overestimated diversity values and therefore provide a rough estimation of effective population sizes of around 252 (Colombia) and 322 (Iberia) genomes per breed in each group.

The LD approach implemented in the SNeP program recorded a declining trend in N_e for all cattle breeds since 250 YA (Figure 3.5). The Iberian populations converged in three distinct clusters, one including Berrenda en Colorado, Lidia and Cárdena Andaluza, with a second including Cachena, Asturiana de los Valles, Retinta and Pajuna, and a third including Mostrenca (Figure 3.5a). These distinct demographic trajectories may correspond to relatively ancient branches such as Black Iberian for Lidia and Cárdena Andaluza, Cantabrian for Cachena and Asturiana de los Valles, or the individual trajectory of Mostrenca, a very ancient semi-feral breed uniquely adapted to the seasonally inundated marshes of Las Marismas in Andalucia (MARM, 2010). Creole breeds produced more homogeneous demographic trajectories, apart from the Texas Longhorn (Figure 3.5b). To further investigate the complex, recent demographic trajectories NeS was used. A constant rate of change is shown as a flat line proximal to 0 in the *y*-axis, whereas deviations above and below 0 represent relative increases and reductions in N_e, respectively (Figure S3.2). This analysis depicted a decrease in N_e towards the end of the expansion period, followed by a temporary recovery in effective size before a collapse to the small N_e detected in the present day (Figure 3.6). Thus, after several recent fluctuations, the current very small N_e was attained only towards the end of the 20th century. The majority of the Iberian breeds recorded similar overlapping NeS patterns (Figure 3.6a). A slowly increasing reduction in N_e being recorded until ~35 generations ago, followed by several fluctuations in N_e, until ~16 generation in the past where a marked reduction in N_e is shown. Among the Iberian breeds, Cachena showed the opposite pattern between ~22 and ~18 generations ago, depicting a sharp increase followed by a reduction in N_e. In contrast (also with all other Iberian breeds), Mostrenca expanded ~25 generation in the past, as well as Asturiana de los Valles in recent generations (~15). The majority of Creole breeds recorded overlapping NeS patterns (Figure 3.6b) and mirrored those recorded by the Iberian breeds, with Senepol displaying a different pattern until very recently (~18 generations ago), whereupon it converges with the other breeds showing an increase followed by a steep population decline.

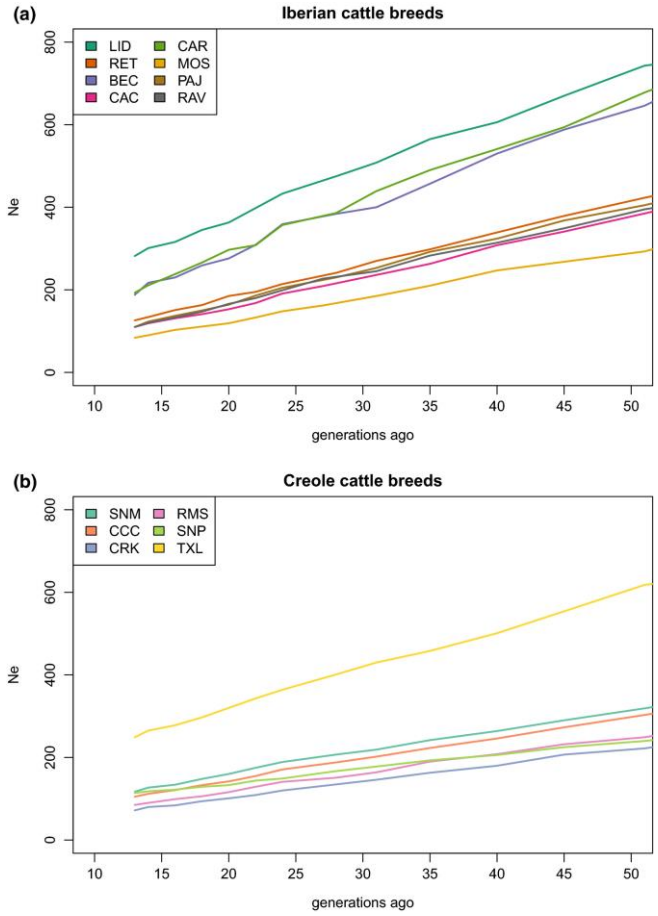


Figure 3.5. Estimation of Ne change between 13 and 50 generations (assumed to be 5 years per generation) ago using SNeP.

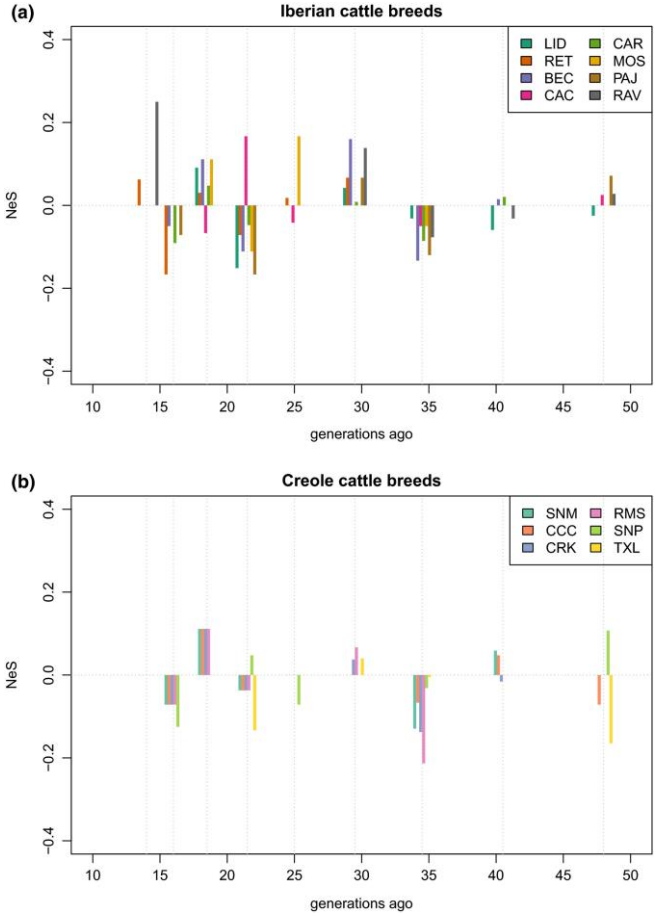


Figure 3.6. N_e Slope analysis (NeS) between 13 and 50 generations ago.

3.4.3 SIGNATURES OF SELECTION

Two methodologies were applied that analyse different patterns of genetic variation, mainly related to evolutionary timescale, to investigate selection pressures enforced by the new tropical environment in six Creole populations, five of which are adapted to humid and hot conditions and one to dry and hot conditions. F_{ST}, better suited to detect signals in the more distant past (Sabeti *et al.*, 2010) that might reflect the zebu ancestral component found in Creole populations, and the LD-based XPEHH method, which provides better resolution for recent selection (Cadzow *et al.*, 2014) and is more suitable for disentangling the differences between Creole and Iberian populations expanding over the last 500 years.

Figure 3.7 and Figure S3.4 and Figure S3.5 depict the genome-wide distribution of outliers on each autosome detected by XPEHH and F_{ST} scans for signatures of selection. The total number of significant SNPs and windows identified per cluster is listed in Tables S2.3 and S2.4. Using the criteria of contiguous blocks of at least two SNPs from the XPEHH analysis (confirmed with more than one Iberian group), or windows containing two or more SNPs from the F_{ST} analysis (confirmed with the three Iberian groups), 10–14 genomic regions under selection per Creole cluster were retreived — two shared between Colombian and Texas Longhorn breeds, one between Colombian and Senepol clusters, and one between Florida Cracker and Texas Longhorn (Table 3.6). Annotation of genomic regions under selection from both analyses retrieved 38, 66, 72 and 61 different genes in the Colombian, Florida Cracker, Senepol and Texas Longhorn clusters, respectively (Table 3.6). GO analysis using DAVID produced a total of 12 enriched functional clusters (Table S3.5) and 13 enriched KEGG signalling pathways (Table 3.4).

Table 3.3. Prior distributions and posterior characteristics for scenario 2, the preferential ABC model with and expanded Creole population between t3 and t1.

Parameter	Prior distributions[a]			Posterior characteristics				
	Scale	Minimum	Maximum	Mode	Q50 lower	Q50 upper	Q90 lower	Q90 upper
Mutation rate	Log_{10}	0.0001	0.05	0.00214	0.00185	0.00292	0.00143	0.00413
N_e_1	Log_{10}	100	500000	57278	10936	116464	2015	343384
N_e_2	Log_{10}	100	500000	40765	8262	99131	1467	325147
N_e_ANC	Log_{10}	100	5000	84	61	111	39	167
N_e_Iber	Log_{10}	100	50000	2577	1725	3975	949	7236
N_e_TXL	Log_{10}	10	5000	638	376	1157	176	2515
N_e_Col	Log_{10}	10	50000	755	378	1622	137	4676
N_e_SNP	Log_{10}	10	5000	497	356	694	224	1094
t1[b]	Linear	5	150	36	28	68	11	100
t2[b]	Linear	20	150	89	64	110	36	136
t3[b]	Linear	50	150	127	92	130	64	145

a Priors were sampled uniformly.

b Time in generations, assuming a generation length of 5 years.

Log10 scaled priors have been converted back from Log10.

Q50, 50th quantile range; Q90, 90th quantile range; Ne_1, effective population size at t1; Ne_t2, effective population size at t2; Ne_ANC, ancestral effective population size; Ne_Iber, Iberian cluster effective population size; Ne_TXL, Texas Longhorn effective population size; Ne_Col, Colombian cluster effective population size; Ne_SNP, Senepol effective population size.

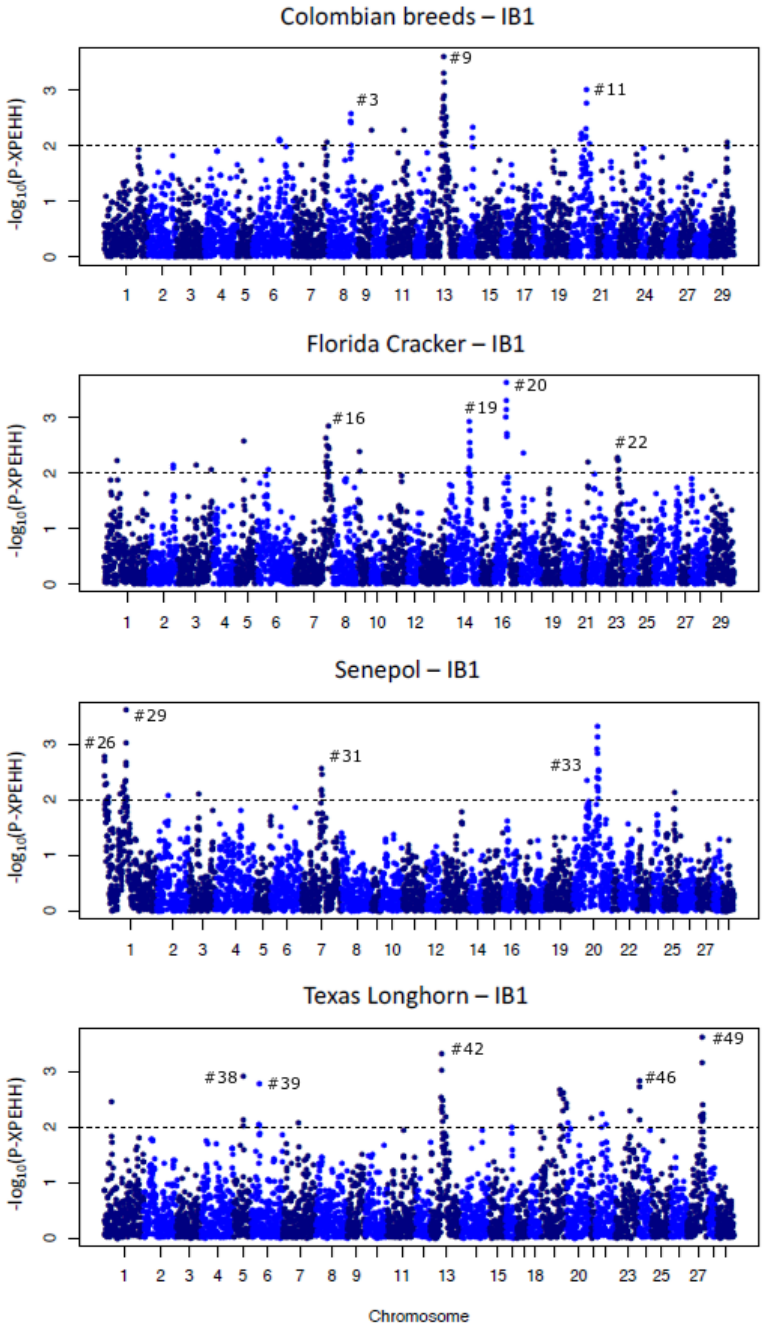


Figure 3.7. Manhattan plots of genome-wide distribution of selection signatures detected with XPEHH for Creole clusters when compared to the Iberian ancestral group IB1. Threshold is set at $-\log_{10}$(P-XPEHH) = 2. The dominant visible peaks are labelled by region, described in Table 3.6. Regions #11 and #33 span the putative location of the slick genotype.

Estimation of different ancestries using LAMP allocated slightly different contributions to Iberian, commercial, African and zebu genomic components (Table 3.6), when compared with the Admixture results (Table 3.1). Several regions under selection in Creole populations showed strong deviations in ancestry contributions (two standard deviations [SD] above or below the

genome-wide average, see Table 3.6), mostly detecting increases in the zebu component. Florida Cracker and Senepol displayed higher proportions of regions under selection with strong ancestry deviations (54% and 60%, respectively), all involving zebu haplotypes except for one in Florida Cracker with an increase in Iberian ancestry. The two regions showing strong deviations in Texas Longhorn were driven by African ancestry, one of them coupled with a higher zebu component. In Colombian breeds, only one region displayed a clear increase above the genome average, again with zebu ancestry. Regions showing an increase in zebu ancestry have been associated with traits important for tropical adaptation, such as the sleek hair coat, conformation and stature, reproduction (including a region associated with reproduction traits in Tropical Composite bulls) and heat tolerance (Table 3.6).

Table 3.5. Enriched KEGG signalling pathways for genomic regions under positive selection in Florida Cracker, Senepol and Texas Longhorn breeds.

KEGG pathway	**Genes**	***P*-value**	**Fold enrichment**
Florida Cracker (CRK)			
bta05031: Amphetamine addiction	*GRIN1, SLC18A2, CAMK2D*	0.007	22.64
bta05030: Cocaine addiction	*GRIN1, SLC18A2*	0.087	20.63
Senepol (SNP)			
bta04060: Cytokine-cytokine receptor interaction	*IFNAR2, FLT3, LIFR, IFNGR2, IFNAR1*	0.002	8.85
bta04630: Jak-STAT signalling pathway	*IFNAR2, LIFR, IFNGR2, IFNAR1*	0.004	11.16
bta04620: Toll-like receptor signalling pathway	*IFNAR2, CD80, IFNAR1*	0.023	12.04
bta04650: Natural killer cell mediated cytotoxicity	*IFNAR2, IFNGR2, IFNAR1*	0.028	10.80
bta04380: Osteoclast differentiation	*IFNAR2, IFNGR2, IFNAR1*	0.035	9.43
bta05162: Measles	*IFNAR2, IFNGR2, IFNAR1*	0.038	9.03
bta05164: Influenza A	*IFNAR2, IFNGR2, IFNAR1*	0.056	7.31
bta05168: Herpes simplex infection	*IFNAR2, IFNGR2, IFNAR1*	0.066	6.65
Texas Longhorn (TXL)			
bta04970: Salivary secretion	*CD38, BST1, LYZ*	0.007	21.08
bta04972: Pancreatic secretion	*CD38, BST1, SCTR*	0.010	18.23
bta00760: Nicotinate and nicotinamide metabolism	*CD38, BST1*	0.050	36.46

Table 3.6. Genomic regions under positive selection detected with F_{ST} and XPEHH analyses in Creole breeds.

Region	SNPs	Method	BTA position	Region length (kbp)	ΔAI[a]	ΔAC[a]	ΔAA[a]	ΔAZ[a]	Candidate genes	QTL
					Whole genome ancestry (mean ± SD)[b]					
Colombian breeds (Col)					0.90 ± 0.08	0.03 ± 0.04	0.01 ± 0.02	0.06 ± 0.06		
#1	2	XPEHH	1:96530234-97142235	612	0.08	-0.03	-0.01	-0.04	*PLD1, TNIK, ENSBTAG00000031795*	-
#2	2	F_{ST}	3:83000001-83500000	500	0.03	-0.01	-0.01	-0.01	*ATG4C, U6, ENSBTAG00000048179*	Milk, reproduction
#3	3	F_{ST}	5:19500001-20000000	500	0.07	-0.03	0.01	-0.05	*ATP2B1, 5S_rRNA*	Tick resistance, weight, performance, milk
#4	3	XPEHH	6:95081924-95220410	138.5	-0.01	0.04	-0.01	-0.01	*ANXA3*	Milk
#5	3	XPEHH	6:115330158-115581341	251.2	-0.04	0.02	-0.01	0.03	*C1QTNF7, CC2D2A, bta-mir-2448, FBXL5*	-
#6	3	XPEHH	8:104528330-104765557	237.2	-0.01	-0.03	0.04	0.01	*RGS3, ENSBTAT00000011467*	-
#7	2	XPEHH	13:38621621-38870180	248.6	-0.07	-0.01	0.01	0.07	*KAT14, ENSBTAG00000004620, ENSBTAG00000004620, DZANK1, POLR3F, RBBP9*	Weight
#8	5	XPEHH	13:39065177-39654999	589.8	-0.08	-0.01	0.01	0.08	*SLC24A3*	Milk, reproduction, feed intake
#9	11	XPEHH	13:39880764-40951781	1071	-0.08	-0.03	0.01	0.10	*NAA20, CRNKL1, CFAP61, INSM1, RALGAPA2, SNORA70, KIZ*	Feed intake, conformation, weight, reproduction, milk
#10	3	XPEHH	13:41726060-42207435	481.4	-0.06	-0.03	0.01	0.08	*FOXA2, U6*	Conformation, weight, reproduction
#11	9	XPEHH	20:35850633-37219008	1368.4	**-0.39**	-0.03	0.04	**0.38**	*LIFR, EGFLAM, SNORA17, U6, GDNF, WDR70, NUP155, bta-mir-2360, ENSBTAG00000000586, NIPBL*	Milk, mastitis, feed intake, meat, reproduction, weight, coat texture
Florida Cracker (CRK)					0.57 ± 0.16	0.27 ± 0.13	0.03 ± 0.07	0.13 ± 0.13		
#12	2	XPEHH	5:44487133-44773477	286.3	-0.07	-0.05	0.03	0.09	*ENSBTAG00000039170, ENSBTAG00000026323, ENSBTAG00000026088, ENSBTAG00000020564, ENSBTAG00000046511, ENSBTAG00000046628, ENSBTAG00000026322, U6, LYZ, CPSF6*	Reproduction (tropical breed)

Region	SNPs	Method	BTA position	Region length (kbp)	ΔAI[a]	ΔAC[a]	ΔAA[a]	ΔAZ[a]	Candidate genes	QTL
					Whole genome ancestry (mean ± SD)[b]					
#13	3	XPEHH	6:12946249-13195583	249.3	-0.18	-0.10	0.08	0.20	*CAMK2D*	Weight, milk
#14	2	XPEHH	7:97937246-98208707	271.5	-0.18	-0.16	-0.03	**0.37**	*PCSK1*	Weight, conformation, meat
#15	4	XPEHH	7:98584703-98897359	312.7	-0.18	-0.16	-0.03	**0.37**	*ERAP1, ERAP2, LNPEP*	Reproduction, weight, performance
#16	6	XPEHH	7:107116333-107853496	737.2	-0.24	-0.16	-0.03	**0.43**	*ENSBTAG00000000360*	Meat
#17	2	XPEHH	9:8595005-9165093	570.1	**-0.35**	0.06	-0.03	**0.31**	*LMBRD1*	Milk
#18	2	F_{ST}	11:105500001-106000000	500	-0.13	0.12	0.03	-0.02	*ZMYND19, ARRDC1, DPH7, PNPLA7, MRPL41, NSMF, NOXA1, ENTPD8, EXD3, NRARP, TOR4A, FAM166A, NDOR1, CYSRT1, RNF224, NELFB, SSNA1, TUBB4B, LRRC26, SLC34A3, RNF208, ENSBTAG00000046416, ENSBTAG00000046223, TMEM203, ENSBTAG00000047715, TMEM210, GRIN1, RXRA*	Milk, reproduction, conformation, fatty acids
#19	15	XPEHH	14:25682788-26937892	1255.1	-0.13	-0.10	0.03	0.20	*FAM110B, ENSBTAG00000047136, UBXN2B, CYP7A1, U1, SDCBP, NSMAF, TOX*	Tick resistance, reproduction, insulin growth factor, milk, weight
#20	6	XPEHH	16:76321854-77156723	834.9	**-0.51**	0.17	0.03	**0.31**	*SNORA48, PLXNA2*	Weight, reproduction, daily gain, performance, conformation, milk
#21	2	F_{ST}	21:29500001-30000000	500	**0.32**	-0.21	-0.03	-0.07	*PCSK6, SNRPA1, ENSBTAG00000003957, ENSBTAG00000047130*	Reproduction, performance
#22	3	XPEHH	23:45450666-45712365	261.7	0.15	-0.05	-0.03	-0.07	*TFAP2A*	Tuberculosis susceptibility, weight, milk
#23	3	XPEHH	23:46206000-46470169	264.2	0.15	-0.05	-0.03	-0.07	-	-
#24	4	F_{ST}	26:37500001-38000000	500	-0.24	-0.16	-0.03	**0.43**	*ENO, SHTN1, VAX1, KCNK18, SLC18A2, PDZD8*	Heat tolerance, temperament, reproduction, milk
Senepol (SNP)					0.64 ± 0.15	0.12 ± 0.10	0.02 ± 0.04	0.21 ± 0.13		
#25	6	XPEHH, F_{ST}	1:1000001-1908934	908.9	**-0.35**	-0.08	-0.02	**0.47**	*ITSN1, CRYZL1, DONSON, SON, GART, DNAJC28, TMEM50B, SNORA20, IFNGR2, IFNAR1, ENSBTAG00000019404, IFNAR2, HIST1H4G, OLIG1*	Polled, milk, reproduction

Region	SNPs	Method	BTA position	Region length (kbp)	ΔAI[a]	ΔAC[a]	ΔAA[a]	ΔAZ[a]	Candidate genes	QTL
					Whole genome ancestry (mean ± SD)[b]					
#26	7	XPEHH, F_{ST}	1:2188833-3000000	811.2	**-0.35**	-0.08	-0.02	**0.47**	*EVA1C, URB1 (splice region variant), MRAP, MIS18A, HUNK*	Weight, performance, reproduction, milk, conformation, stature
#27	2	XPEHH	1:6557886-7047431	489.5	**-0.35**	-0.08	-0.02	**0.47**	*LTN1, ENSBTAG00000038433, N6AMT1, U6*	-
#28	3	XPEHH	1:63614355-64126677	512.3	-0.21	-0.12	-0.02	**0.36**	-	Milk, meat
#29	9	XPEHH	1:64565646-65264693	699	-0.21	-0.12	-0.02	**0.36**	*UPK1B, B4GALT4, ARHGAP31, TMEM39A, POGLUT1, TIMMDC1, CD80, ADPRH, PLA1A, POPDC2, COX17, MAATS1, SNORA31, NR1I2*	Meat, milk, reproduction
#30	2	F_{ST}	5:90000001-90500000	500	-0.14	0.09	0.02	0.04	*ENSBTAG00000044467, 5S_rRNA*	Reproduction, milk, conformation, stature
#31	5	XPEHH	7:64259215-65018918	759.7	0.07	-0.12	-0.02	0.08	*GPX3, TNIP1, ANXA6, CCDC69, ENSBTAG00000045615, GM2A, SLC36A3, SLC36A2, SLC36A1, ENSBTAG00000003498, ENSBTAG00000047625, SPARC, ATOX1, G3BP1*	Milk, reproduction, conformation, performance,
#32	2	F_{ST}	12:32000001-32500000	500	0.11	-0.08	-0.02	0.00	*FLT3, URAD, ENSBTAG00000001819, PDX1, ENSBTAG00000009166*	Weight
#33	14	XPEHH	20:35850633-38012333	2161.7	**-0.37**	-0.08	0.02	**0.45**	*LIFR, EGFLAM, SNORA17, U6, GDNF, WDR70, NUP155, bta-mir-2360, ENSBTAG00000000586, NIPBL, ENSBTAG00000047208, SLC1A3, RANBP3L*	Slick hair coat, milk, mastitis, feed intake, meat, reproduction, weight
#34	2	F_{ST}	21:16500001-17000000	500	0.04	-0.05	-0.02	0.04	*U6, ENSBTAG00000037383, KLHL25*	Reproduction, milk, tuberculosis susceptibility
Texas Longhorn (TXL)					0.87 ± 0.07	0.04 ± 0.04	0.02 ± 0.03	0.07 ± 0.05		
#35	2	XPEHH	1:147742602-147862779	120.2	0.01	0.01	0.03	-0.05	*PCNT, DIP2A*	Weight
#36	3	F_{ST}	2:33500001-34000000	500	0.01	-0.02	-0.02	0.03	*ENSBTAG00000047523*	Fat, conformation, reproduction, performance, stature
#37	3	XPEHH	2:71689118-72269356	580.2	0.01	0.04	-0.02	-0.02	*SCTR, CFAP221, TMEM177, PTPN4, ENSBTAG00000048209, EPB41L5*	-
#38	4	XPEHH	5:44373006-45012918	639.9	**-0.17**	-0.02	**0.06**	**0.13**	*U5, ENSBTAG00000022971, ENSBTAG00000000198, ENSBTAG00000039170, ENSBTAG00000026323, ENSBTAG00000026088, ENSBTAG00000020564, ENSBTAG00000046511, ENSBTAG00000046628,*	Reproduction (tropical breed), meat, milk

Region	SNPs	Method	BTA position	Region length (kbp)	ΔAI[a]	ΔAC[a]	ΔAA[a]	ΔAZ[a]	Candidate genes	QTL
					Whole genome ancestry (mean ± SD)[b]					
									ENSBTAG00000026322, U6, LYZ, CPSF6, ENSBTAG00000002741, SNORA44	
#40	3	XPEHH	6:19530474-20352600	822.1	0.00	-0.02	0.01	0.00	*GIMD1, AIMP1, TBCK, ENSBTAG00000043620*	Milk, meat, weight
#41	5	XPEHH	6:115330158-115857381	527.2	0.01	-0.04	-0.02	0.06	*C1QTNF7, CC2D2A, bta-mir-2448, FBXL5, U6, BST1, CD38*	-
#42	10	XPEHH	13:39981697-40951781	970.1	-0.12	-0.04	**0.16**	0.00	*CFAP61, INSM1, RALGAPA2, SNORA70, KIZ*	Feed intake, conformation, weight, reproduction, milk
#43	4	XPEHH	13:46248573-46574633	326.1	0.08	-0.04	0.01	-0.05	*ADARB2, ENSBTAG00000039356, ENSBTAG00000037833*	Milk
#44	3	XPEHH	14:7907751-8007829	100.1	-0.05	-0.02	0.01	0.06	-	Milk
#45	3	XPEHH	15:59919746-60219850	300.1	-0.05	-0.02	0.03	0.03	-	Reproduction, conformation, performance
#46	3	XPEHH	23:51003189-51242238	239	0.01	-0.02	0.01	0.00	*ENSBTAG00000037624, ENSBTAG00000012058*	Meat, milk, reproduction
#47	3	XPEHH	24:57909464-58455030	545.6	0.08	-0.04	0.01	-0.05	*NEDD4L, bta-mir-122, ALPK2, MALT1*	Abomasum displacement, milk, meat
#48	4	XPEHH	27:35316858-36020332	703.5	0.11	-0.04	0.01	-0.07	*ZMAT4, 5S, SNORA70, SFRP1*	Milk
#49	6	XPEHH	27:36220416-36884199	663.8	0.11	-0.04	0.01	-0.07	*GPAT4, ENSBTAG00000047361, ENSBTAG00000004244, bta-mir-486, ENSBTAG00000004242, ENSBTAG00000063621, KAT6A, U6, AP3M2, PLAT, IKBKB*	Milk, meat, performance, conformation, weight

[a]ΔAI, ΔAC, ΔAA, ΔAZ: estimated excess/deficiency of the Iberian, commercial, African and zebu ancestries, respectively. In bold, substantial increase/decrease in ancestry by more/less than two standard deviations (SD) from the whole genome mean, respectively.

[b]Whole genome ancestries obtained with the software LAMP.

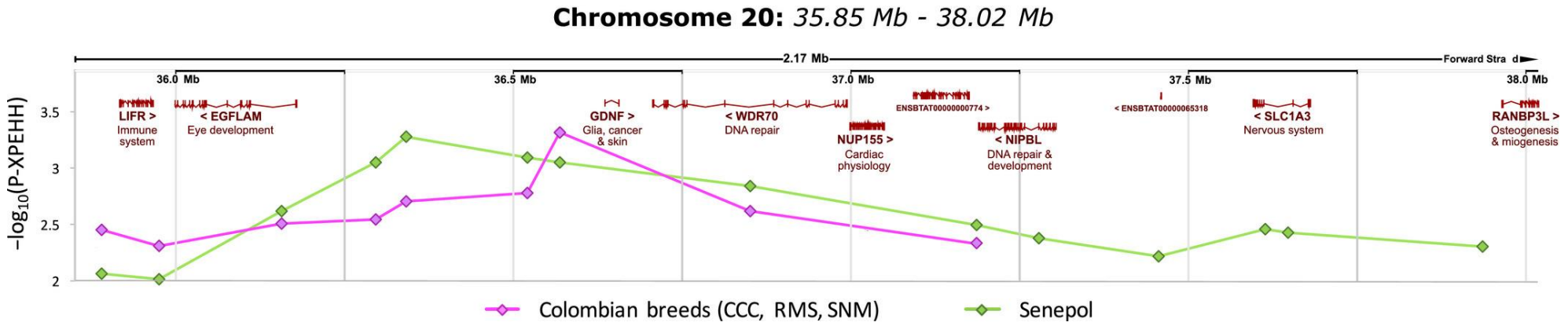


Figure 3.8. Selection signatures in the BTA20 genomic region shared by the Colombian cluster (Costeño con Cuernos, Romosinuano, San Martinero) and the Senepol breed. Plot of $-\log_{10}$(P-XPEHH) values (*y*-axis) around loci (*x*-axis in Mb). Points mark significant SNPs.

3.5 DISCUSSION

Tropical adaptation, that is the ability to tolerate heat stress, high humidity, tropical diseases, and parasite infections while maintaining standards of performance and reproduction, constitutes the most valuable asset of Creole cattle, assuring protein production within its region and providing insights into genomic and physiologic mechanisms selected during the transition to a tropical environment. Most Creole breeds included in this study (Costeño con Cuernos, Romosinuano, San Martinero, Florida Cracker, Senepol) have been developed under physiologically challenging tropical conditions and tolerate high temperatures and humidity, poor soils, drought, high rainfall, and are tick resistant, all while maintaining good performance (Alba, 1987). In addition, breeds such as Texas Longhorn have adapted to very hot and dry tropical conditions including the ability to reproduce very effectively with minimal human intervention where forage is sparse.

Creole populations included in this study were largely unaffected by the introduction of African taurine cattle into the Americas, which reached its highest proportion in the San Martinero and Texas Longhorn (0.05 ± 0.01 SD; Table 3.1; Figure 3.2). This residual African genomic component may be explained by ancient introgression in the Iberian Peninsula and the Canary Islands (McTavish *et al.*, 2013). The high contributions of zebu (15%) and European commercial breeds (14%) with minor elements of African taurine ancestry (2%) found in Senepol are in accordance with the results obtained by Flori *et al.* (2012) and Huson *et al.* (2014), and they argue against the reporting of direct incorporation of N'Dama into Senepol breeding (Miretti *et al.*, 2004). Although these authors attributed all European taurine contribution to Red Poll ancestry, these results strongly imply a major Iberian origin (68%) with a much lower ancestral contribution from commercial breeds (14%), including Red Poll. Despite the claimed admixture of Romosinuano with polled British breeds to incorporate polledness into its phenotype (Huson *et al.*, 2014), contribution from European commercial breeds (including Red Poll samples) was inferred to be low and equal to that of Costeño con Cuernos (7%), from which the Romosinuano was developed. Although as far as it is known, Florida Cracker has not been crossed with European commercial breeds (Ekarius, 2008), this ancestry represents 36% of its genomic pool. Finally, despite indicine introgression having been described in the Texas Longhorn (Decker *et al.*, 2014), the values detected here are within the mean range for all Creole cattle populations (8%). These results illustrate the influence of taurine and indicine ancestry that may underlie some of the demographic patterns and selection signatures found in Creole populations.

Alternatively to introgression from indicine and African taurine, it is possible that the exposure to an environment similar to that which African taurine experience applied selective pressure to ancestral components within the genome of Creole populations while residing in the Canary Islands or later in the Americas. Inclusion of more distantly related species such as other *Bos* or *Bovinae* to the dataset and identification of the alleles shared across taxa may indicate a retained ancestral state (Naji *et al.*, 2021). TREEMIX may provide insight into the source of the shared ancestral components, constructing a maximum likelihood tree and then additional migration edges between branches to further explain variance in the data (Pickrell and Pritchard, 2012). This could be combined with the f_4-statistic that uses allele frequencies of four populations (perhaps Creole, Iberian, African taurine and Indicine breeds) to distinguish between incomplete lineage sorting and introgression to determine if differential selection on ancestral genotypes or direct gene flow was causative in the shared ancestral components (Reich *et al.*, 2009).

ABC analysis described events in close agreement with the known history of a small founding population, relatively unrestricted expansion, and later contraction and marginalisation of cattle of Iberian origin in the Americas (Willham, 1982; Alba, 1987; Rodero *et al.*, 1992; Villalobos Cortés *et al.*, 2009). This is reflective of the general trend displayed by populations that successfully colonize new habitats, undergoing a bottleneck followed by rapid growth usually due to lack of competition (Gray *et al.*, 2014). However, in the case of Creole cattle, population growth was likely aided by extensive habitat modification (Alba, 1987). The building of realistic models and priors in the ABC analysis was guided by historical population and migration records, ADMIXTURE, MDS, neighbour-net results, and recent N_e estimations based on LD, and included a wide representation of the Iberian populations sharing a common ancestor with Creole breeds in the recent past. However, obtaining exact parameter estimates can be complex (Gray *et al.*, 2014), which may explain the discrepancy found between the colonization time t3 (635 YA) and known dates such as the arrival of cattle to the Americas after 1492 (524 YA; although within the 50th quartile range of 460–650 years). However, the drastic N_e reduction from t1 (180 YA) to present closely correlates with the introduction of zebu and commercial cattle breeds to the Americas, starting around the middle of the 19th century and causing the gradual replacement of Creole populations that has led to their small current effective population sizes (Willham, 1982; Alba, 1987; Taberlet *et al.*, 2011; MacLeod *et al.*, 2013). Despite the influence of European commercial breeds and zebu cattle detected here and supported by historical records (Decker *et al.*, 2014; Felius *et al.*, 2014), computational constraints hampered their incorporation in the models. It is possible that the potential oversimplification of the models analysed here may underestimate the complex demography of Creole breeds and obscure recent Iberian, European and zebu influences.

Contrasting the sudden decline in observed in the ABC analysis 180 YA, SNeP analysis using a greater frequency of timepoints allowed visualisation of a more gradual population decline in recent history. The difference in inference gained using ABC and SNeP is likely to reflect their resolution of temporal complexity, where ABC only allows comparison among competing demographic scenarios whereas SNeP applies a single, model-free algorithm and its application enables the inference of more complex, short-term, events instead. Thus, ABC reveals general trends and their relative likelihood, while LD-based analysis provides an insight on the short-term complexity within these trends. The novel NeS method records the change in slope of the inferred N_e trend obtained from LD-based demography analysis implemented in SNeP, potentially offering a more detailed picture of population changes 13–50 generations ago. Interestingly, the greatest decline detected by the NeS approach and the only timepoint in which all six Creole breeds displayed unanimous directionality occurred at 34 generations ago (170 YA), indicated that although the ABC approach may provide lower resolution of more subtle changes there is substantial power in both methods to detect major demographic events.

The region in BTA20 shared by Colombian (region #11) and Senepol (region #33) populations showed signals of selection with the XPEHH analysis and demonstrated a strong increase in zebu ancestry of 38% (more than 6 *SD*) in Colombian breeds and 45% (almost 4 *SD*) in Senepol (Table 3.6), implying that zebu haplotypes, otherwise representing a small proportion genome-wide, are strongly selected for in this region and that anthropogenic selection and/or local adaptation rather than genetic drift is driving their presence. Among the genes included in this area, *LIFR* is implicated in immune processes, *NUP155* displays functions in cardiac physiology, *RANBP3L* is implicated in osteogenesis and myogenesis (Chen *et al.*, 2015), and *WDR70* and *NIPBL* are involved in DNA repair processes, highly conserved in nature to remove or tolerate DNA damage caused, among other exogenous factors, by ultraviolet daylight, especially intense in tropical latitudes (Menck and Munford, 2014). This region overlaps with several cattle loci associated with milk traits, mastitis, feed intake, meat attributes, reproduction, and weight. Importantly, it also partially overlaps with the region for the slick hair coat, a phenotype that plays an important role in thermotolerance in some tropical Creole breeds, including Senepol and Romosinuano (Flori *et al.*, 2012; Huson *et al.*, 2014). Slick hair coat is characterized by sleek, short hair coupled with increased perspiration. The sleek and shiny properties of this coat may reflect solar radiation more efficiently, and the hair coat thickness and hair weight per unit surface increase heat loss via convection and conduction. As a result, slick animals show lower temperature and respiration rates and an increased production under tropical conditions when compared with normal-haired individuals (Flori *et al.*, 2012). This broad region has already been introgressed into a Holstein lineage to improve thermoregulatory ability,

reducing summer milk yield depression (Pulina *et al.*, 2017). Several studies have associated a region in BTA20 to this phenotype and suggested different candidate genes (*PRLR,* Mariasegaram *et al.*, 2007; *RAI14,* Flori *et al.*, 2012; *SKP2, SPEF2,* Huson *et al.*, 2014). However, the causative mutation is still unknown. Here, the detected region under selection in BTA20 is located slightly downstream (36–38 Mb) compared to the others studies (37–40 Mb), with the most significant SNPs peaking around the *GDNF* gene both in Senepol breed and Colombian group, which included the Romosinuano breed (Figure 3.9, Table 3.6, Table S3.3). A possible explanation for the lack of complete overlap with other studies may be the inclusion in the analyses for the first time of the Iberian populations sharing a common ancestor with Creole cattle in the recent past. The candidate gene for the slick phenotype identified here, the glial cell-derived neurotrophic factor (*GDNF*), has important roles in skin homeostasis, is involved in the migration and differentiation of melanocytes and shows a strong expression in sebaceous and sweat glands (Adly *et al.*, 2006). It is also implicated in hair follicle morphogenesis and cycling control, increasing the number of the proliferating HF keratinocytes (Adly *et al.*, 2006). However, as in previous studies, the associated SNPs are located in noncoding regions and further studies are needed to narrow down the causative mutation.

Another region in BTA06 showing selection signal with the XPEHH methodology in two clusters, Colombian group (region #5) and Texas Longhorn (region #41), has not been associated with any QTL in cattle so far and includes genes such as *C1QTNF7*, related to *Trypanosoma cruzi* cardiomyopathy (Deng *et al.*, 2013), *FBXL5*, which controls iron metabolism processes key for the regulation of reactive oxygen species that augment with the exposure of animals to high environmental temperatures (Paital *et al.*, 2016), *BST1* that has immune functions facilitating pre-B-cell growth, and *CD38* that has pleiotropic functions in T-cell activation (Würsch *et al.*, 2016), social behaviour through its effect on the release of oxytocin (Krol *et al.*, 2015) and cancer. *BST1* and *CD38* are also implicated in salivary and pancreatic secretion and nicotinate and nicotinamide metabolism pathways (Table 3.4). The genes in this region represent adaptations to new and challenging environments, including immune function, nervous and behavioural processes that may be key for animals to adapt to new environmental conditions, metabolism, high environmental temperatures, and diet.

Although the genes included in the region under selection in BTA05 shared by Florida Cracker (region #12) and Texas Longhorn (region #38) and detected with XPEHH are mostly uncharacterized novel genes in Ensembl, as well as the antimicrobial agent lysozyme (*LYZ*) and other genes with no clear role in reproduction, this region has been associated with reproduction traits in Tropical Composite bulls. Concordantly, a substantial increase in zebu (by 13%) and

African (by 6%) ancestries was found in the Texas Longhorn, although this was not found in the Florida Cracker. Another region under selection in two clusters, Colombian (region #9) and Texas Longhorn (region #42), was also detected with XPEHH methodology and included genes in BTA13 with roles in reproduction (*CFAP61*), neuroendocrine differentiation (*INSM1*), cancer (*RALGAPA2*) or cell cycle (*KIZ*). This region has been previously associated with QTLs related to production traits in cattle (Table 3.6) and displayed a strong increase in African ancestry (10%, more than 5 *SD*) in Texas Longhorn, but again imperceptible in the Colombian cluster.

Apart from these genomic regions under selection in more than one cluster, signatures of selection associated with a variety of traits were detected (Table 3.4; Table 3.6; Table S3.5). These include regions of the genome enriched for genes involved in immune system activation in response to infectious diseases (tick resistance in the Colombian group and Florida Cracker, tuberculosis susceptibility in Florida Cracker and Senepol, mastitis in the Colombian group and Senepol), or enriched immune pathways in Senepol (cytokine–cytokine receptor interaction, Jak-STAT signalling, Toll-like receptor signalling, natural killer cell-mediated cytotoxicity, osteoclast differentiation, and responses to viral diseases -measles, influenza A, herpes simplex-). In addition, regions enriched for genes associated with heat tolerance, including regulation of blood pressure and, importantly, thermoregulation in lactating cows exposed to heat stress in the Florida Cracker were found (region #24). This region in BTA26 showed a strong increase in zebu ancestry (43%, more than three *SD*) and was also implicated in temperament, with the *SLC18A2* gene involved in the dopamine and serotonin pathways associated with temperament in cows (Garza-Brenner *et al.*, 2017). Phenotypic variation driven by production aims, such as beef or dairy traits, may have had an impact in the genomic areas under selection, highlighted here by the regions detected within QTLs associated with milk and meat production, fatty acid profile, performance, conformation, and reproduction.

Finally, the signal for the polled locus (Flori *et al.*, 2012; Medugorac *et al.*, 2012) in Senepol (BTA01 region #25) was also validated, with both XPEHH and F_{ST} methodologies. This region showed a strong zebu component increase of 47% (almost four *SD* deviations above the genome mean). None of the previously described polled mutations are located in known coding regions. Within the candidate region identified in this study, the most significant SNPs peaked around three genes, *GART*, *DNAJC28*, and *TMEM50B*, none of them with a clear role in polled ontogenesis. The key immune functions displayed by several genes in this region (*IFNAR2*, *IFNGR2*, *IFNAR1*; Table 3.6), which could be important in responses against tropical diseases and parasite infections, may distort the signal from the polled locus.

Although F_{ST}- and LD-based methodologies are widely used, there are other possible factors apart from selection that may mimic the signals obtained, including demographic events such as the bottlenecks and expansions detected with the ABC and SNeP analyses (Vitti *et al.*, 2013). Moreover, the use of SNP array markers may under- or overestimate genetic diversity through ascertainment bias, distorting allele frequencies, interfering with the accuracy of distinguishing admixture from differential selection of ancestral components, and analysis of derived statistics such as LD (Vitti *et al.*, 2013). Also, selection response for complex traits caused by weak selection at many sites across the genome may leave few or no classical signatures (Kemper *et al.*, 2014), reducing the signal obtained. However, other studies on cattle adaptation to new environments (Porto-Neto *et al.*, 2014; Makina *et al.*, 2015), including tropical adaptation, reported the slick hair coat and QTLs associated with tick resistance, heat tolerance and reproduction in tropical populations. A key assumption made in this study is that the Creole and Iberian differ only in environment, however, farm management, farmer selection and preference are likely to contribute in defining each breed and selective pressures the two populations are exposed to (Garforth, 2015). While it is logical that artificial selection of some traits – such as survival – are often aligned with natural selection and adaptation to the local environment, oftentimes without careful management there is antagonism between two beneficial traits such as productivity and fertility (Bieber *et al.*, 2020). The inclusion of regional wild or feral conspecifics, or even other locally adapted mammals, would allow for separation of environmental- and farmer-driven selection between the two continents (Stronen *et al.*, 2019; Buggiotti *et al.*, 2021).

In conclusion, modern Creole cattle was compared with modern day samples from breeds comprising their putative Iberian ancestors for the first time to reconstruct their demographic history and search for selection signatures enforced by American environments on a small number of founder animals during a brief period of time. Showing that despite strong evidence for rapid genomic adaptation to their new tropical environments (e.g., for slick hair coat genes improving thermotolerance), Creole cattle have recently undergone a major decline and will require genetic conservation measures to be put in place to avoid their potential disappearance and ensure they thrive again. The outcomes from this study will contribute to the design of innovative breeding schemes that will include, apart from traditional performance traits, resilience biomarkers, allowing sustainable production in harsh environments and improving sanitary conditions in farms under the ongoing climate changes.

3.6 ACKNOWLEDGEMENTS

Thank you to Javier Cañón and Susana Dunner from Universidad Complutense de Madrid for their contribution of samples and SNP array data. Thank you to Rodrigo Martinez for collection of Colombian cattle samples. Thank you to Mario Barbato for useful discussions regarding SNeP analyses and contributing NeS analysis. A special thanks to Natalia Sevane for her supervision and advice.

Chapter Four

Signatures of selection and landscape genomics of Ryeland sheep

---

Photo by John Donovan

4.1 ABSTRACT

The Ryeland is one of the oldest British sheep breeds, originally created in the 14th century. The breed has been developed under a range of shifting selective pressures and management systems over the past 700 years, yet defining characteristics have persisted, including ease of management, fine fleece, and tolerance to privation. Here demography, population structure, genetic diversity, and selection was investigated for 60 Ryeland sheep genotyped using Illumina's OvineSNP50 array. A total of 36,326 single nucleotide polymorphisms (SNPs) remained after filtering. Population structure analysis revealed a division in Ryeland sheep, separating the Northern-most nine individuals into a unique cluster. Further population genetic analysis (F_{ST}, Runs of Homozygosity [RoH], identity by descent [IBD]) suggested isolation and genetic drift as causative factors, rather than introgression or admixture. Over the past 200 years, Ryeland were estimated to have one of the lowest modern effective population sizes (N_e = 84), however, they were amongst the most stable of the twelve breeds tested (N_e decline/year = 1.65). Scans for selective sweeps using cross-population extended haplotype homozygosity (XPEHH) and subsequent gene ontology analysis isolated up to 13 functional annotation clusters and nine enriched KEGG pathways, including the identification of positive sweeps for multiple metabolic pathways and prion disease pathogenesis mediated through the complement system. Landscape genomic approach was implemented with environmental variables gathered through the Met Office weather data, digital elevation modelling and land cover mapping identified three genes associated with these variables. All three genes (*TMEM123, DNAJC25, BZW2*) are implicated in cancer regulation, the complement system. *DNAJC25* and *BZW2* are also closely associated with hepatocellular carcinoma growth. These results provide potential targets for maintenance of breed health and an understanding of unique variation within the Ryeland.

4.2 INTRODUCTION

Sheep were domesticated approximately 10,000-11,000 years ago and since have been one of the most economically important agricultural assets for humans (Zeder, 2008). Short generation times, high fecundity, lack of aggressive behaviour and manageable body size are a few traits that cause the rapid spread of sheep husbandry throughout Neolithic culture (Helmer *et al.*, 2007). Distinct mitochondrial lineages support at least three separate domestication events occurring in sheep, indicating the species' suitability as livestock (Tapio *et al.*, 2006; Meadows *et al.*, 2007). Initially utilised primarily for meat, about 4,000-5,000 years ago, increasing selection

was evident for secondary products such as fleece and milk (Helmer *et al.*, 2007). In the majority of cases it seems that primitive sheep were progressively replaced by those specialised for generation of secondary products (Chessa *et al.*, 2009). Development of the breed concept in the last couple centuries has accelerated the fixation of population-specific phenotypes often favouring specialisation towards a single product (e.g. meat, milk or high-quality fleece). Modern domestic sheep (*Ovis aries*) have since been radiated from their native range and established across the world. A diverse array of unique and locally adapted breeds has emerged from exposure to differing environmental, production and anthropogenic selective pressures.

Ryeland sheep are an ancient British breed, first established in Herefordshire and the Welsh Borders, taking their name from the rye pastures on which they were grazed. The monks of Leominster abbey developed the modern breed in the 14th century, which quickly became renowned for the production of finest and highest-quality fleeces among British sheep (Youatt, 1837). The fleece was coined as "Lemster ore" due to the profitability and local commerce it provided (Ryeland Flock Book Society, 2019). Demand for wool dropped in the 18th century with the rise of the cotton industry. As a result, Ryeland fleece quality most likely declined as improving carcass weight and quality became increasingly incentivised.

Commercial viability of Ryeland sheep had dropped by the beginning of the 20th century, leading to the establishment of the Ryeland Flock Book Society (RFBS) in 1903 with a primary aim to maintain the purity of the breed as the population size dropped (Ryeland Flock Book Society, 2019). Despite increasing popularity abroad (e.g., New Zealand, Australia, South Africa and much of Europe), between 1920 and 1952, the number of flocks registered with the society halved from 80 to 40. Continued demand for meat production during the 1950s to 1970s and decline in Ryeland numbers led to introgression from two dual-purpose breeds, Cotswold and Southdown (Kelham *et al.*, 2013). Additionally, during this period a Ryeland tup was reintroduced into Southwest England from a productive Australian flock that had been established in the 1920s. Only 980 breeding ewes were registered with RFBS in the mid 1970's as Ryeland transitioned away from commercial relevance, becoming almost exclusively a hobbyist's breed. Consequently, the newly formed Rare Breeds Survival Trust (RBST) added Ryeland to a watchlist of threatened traditional British breeds. Recent years have seen Ryeland recover with over 3,000 breeding adult females in at least 550 flocks and subsequent removal from threatened status (Rare Breeds Survival Trust, 2019).

Although of slighter frame compared to most purpose-bred British meat breeds, Ryeland have been well-known since the 18th century as resilient in poor pastures and capable of thriving with sparse feed. Sir Joseph Banks who was well acquainted with the constitution of Ryeland

sheep stated that they "deserved a niche in the temple of famine" (Youatt, 1837). Additionally, it has been suggested qualitatively that British Ryeland have a relatively high innate resistance against transmissible spongiform encephalopathies (TSE), such as scrapie (Ryeland Flock Book Society, 2019). An indication of natural resistance was observed in 1957 when 24 breeds of British sheep were screened for susceptibility using intracerebral injections of SSBP/1 scrapie. Of the 34 Ryeland sheep inoculated, only 15% developed any signs of disease (Gordon, 1966). TSE or prion diseases have been economically and agriculturally relevant, particularly in the UK for the past 40 years, with peak incidence in 1993 (Bruce *et al.*, 1997). In an eight year monitoring study of Dutch sheep, Ryeland exhibit the highest frequency of the prion protein gene polymorphism (ARR) that is associated with resistance (Melchior *et al.*, 2011). Seventy eight percent of Dutch Ryeland are homozygous for the ARR genotype with over 95% carrying at least a single copy (Melchior *et al.*, 2011). The introduction of Ryeland to the Netherlands was before the modern strong selection and regular genotypic screening of prion resistance, suggesting long standing natural resistance. Similar results were observed in British Ryeland individuals, which displayed one of the highest ARR allelic (89.2%) and ARR homozygous genotype (~80%) frequencies among the 27 rare breeds sampled (Townsend *et al.*, 2005).

Selection can be efficiently detected using modern high-throughput genomic data even in the absence of corroborating phenotypic information (Akey, 2009; Schrider *et al.*, 2016; Grossen *et al.*, 2020). Loci under historic or current selection can be identified through interrogation of allelic fixation, haplotype structure and diversity loss (Utsunomiya *et al.*, 2015). Analyses of selection typically initially focus on comparing populations exposed to similar selective pressures (Sabeti *et al.*, 2007). Each livestock breed experiences different artificial pressures even within the same geographical or climatic environment, often via production traits and breeding structure. This paves the way for cross population analysis whereby comparisons between breeds within the same region produce signatures of selection descriptive of the unique variation to each breed whilst minimising the impact of environmental factors (Iso-Touru *et al.*, 2016). Landscape genomics methods identify associations between selection and the environment and are increasingly facilitated by the availability of processing power, high resolution environmental data, and genomics. It becomes possible to account for environmental variation within a breed, identifying subtle local adaptation within a structured population (Stucki *et al.*, 2017).

This is the first genome-wide (~54,000 SNPs) dataset produced for Ryeland sheep, used to investigate the British population with 60 individuals sampled from 19 flocks across England, Wales, and Scotland. The aims of this study were to first characterise population structure and genetic diversity within Ryeland, with particular attention to identifying and explaining any

population subdivisions within the breed. Introgression may be expected from other British sheep breeds that may explain phenotypes that differ between different flocks of Ryeland, such as body size or coat colour. Secondly, identifying local adaptation within Ryeland associated environmental variables, geographic features, and land usage. Finally, scanning for selective sweeps within Ryeland sheep relative to both British and more distantly related breeds. Aiming to characterise the regions associated with distinctive and agriculturally important traits for which Ryeland are renown, such as prion resistance and hardiness. Traditional livestock breeds are under threat of extinction due to increased demand for higher productivity. Identifying consistent biological functions and processes under selection in Ryeland will improve understanding of the value of older traditional breeds with respect to local adaptation, climate change resilience and disease resistance.

4.3 METHODOLOGY

4.3.1 SNP ARRAY DATA

Sixty domestic sheep samples from the Ryeland breed were analysed, renamed with anonymised internal IDs, collected from 19 flocks across England, Wales, and Scotland. A blood aliquot was collected from each animal by local veterinary surgeons, corresponding to less than 10% of the volume taken during routine veterinary analyses. DNA was extracted from the blood aliquots with the Qiagen Blood and Tissue Kit. DNA extractions were genotyped with Illumina's OvineSNP50 array by the Laboratory of Genetics and Services of the Italian Association of Animal Producers and within the context of the ClimGen Project (https://climgen.bios.cf.ac.uk). Markers with a call rate below 99% or a minor allele frequency below 5% were filtered out of the dataset using PLINK v1.90 (Purcell *et al.*, 2007). Heterozygosity (H_O), inbreeding coefficient (F_{IS}) and identity by descent (IBD) across Ryeland were calculated in PLINK.

4.3.2 POPULATION STRUCTURE AND ADMIXTURE

Within breed population structure of Ryeland sheep was analysed in ADMIXTURE v1.3 (Alexander *et al.*, 2009). The software assumes the markers are unlinked; therefore, an additional dataset was produced, retaining only markers within approximate linkage equilibrium using the –indep-pairwise function in PLINK. SNPs with r^2 greater than 0.2 were removed from a sliding window of 10 kb with a step-size of 5 SNPs. For the ADMIXTURE analysis, the number of clusters (K) tested ranged between 1 and 10 with each individual's ancestry assigned proportionately to each

cluster. The fivefold cross-validation (CV) procedure implemented in ADMIXTURE was used to estimate which value of K best discriminated the individuals into clusters as indicated by the lowest CV. Unexpectedly, Ryeland preferentially clustered into two groups, a Northern cluster of 9 individuals exclusively comprised of Scottish samples and a Southern cluster of 51 individuals across England and Wales (Figure 4.1; Figure 4.2). These results were used to inform suitable methodological choices downstream. Henceforth, the two clusters will be referred to as 'Northern Ryeland' (n = 9) and 'Southern Ryeland' (n = 51), respectively.

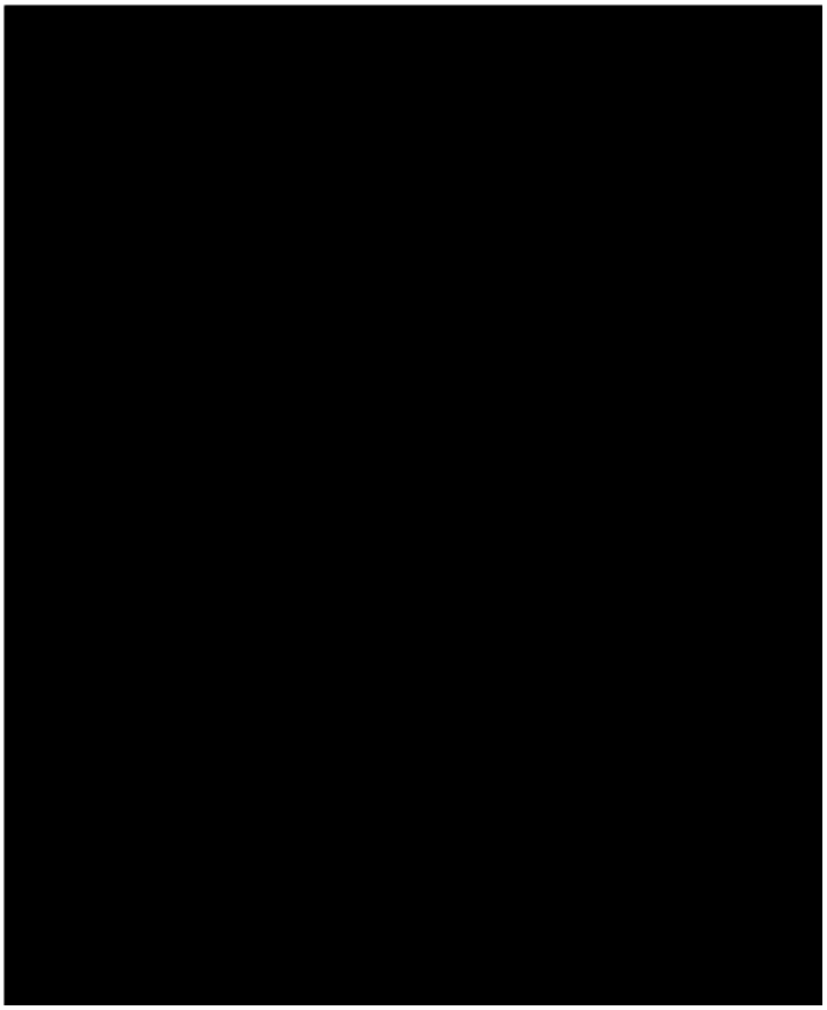


Figure 4.1. Ryeland sheep sample distribution in the UK. Individuals cluster into two groups (Figure 4.2), separable across a north-south divide. Numbers given are the anonymised internal IDs.

To assess potential admixture with other breeds that could explain the partition of the breed into two clusters, 99 global breeds/populations genotyped with the same SNP array and representing sheep in the UK, Europe, Asia, Africa, the Americas and the domestication centre in the Middle East (Kijas *et al.* 2012; Beynon *et al.* 2015; International Sheep Genomic Consortium initiative [www.sheephapmap.org]) were merged with the Ryeland data. Admixture between these breeds was tested for values of K in the range between 1-10 and then discontinuously, every 5 values, between 15-135. Mean F_{ST} was calculated in VCFTOOLS (Danecek *et al.*, 2011) between

the global breeds and the two Ryeland clusters. Additionally, a network was plotted with the resultant matrix in SPLITSTREE V.4.14.4 (Huson and Bryant, 2006). H_O and F_{IS} were calculated in PLINK for breeds containing at least 8 individuals.

4.3.3 RUNS OF HOMOZYGOSITY

Using the 36,326 markers, Runs of Homozygosity (RoH) were calculated separately for Northern and Southern Ryeland using the R (R Core Team, 2018) package *DETECTRUNS* (Marras *et al.*, 2015), which uses a similar approach to the one implemented in PLINK. For this analysis, the genome is scanned in windows of 50 SNPs and sliding each window a single SNP at the time. To be considered as a RoH, an extended region of adjacent homozygous SNPs, the run had to be at least 15 SNPs in length, adjacent SNPs could be no further apart than 1 Mb, the minimum length of a RoH had to be greater than 1 Mb, the average density had to exceed 1 SNP per 100 kb. The RoH must also contain no more than 1 heterozygous SNP, allowing some flexibility for genotyping error or a novel mutation. The remaining settings were left as default, recapturing the recommended parameterisation by Marras *et al.* (2015) to limit the identification of spurious RoH. Features of the RoH were compared to windowed F_{ST}, calculated in VCFTOOLS between Northern and Southern Ryeland using a window size of 1,000 Mb and a step size of 100 kb. Additionally, windowed F_{ST} was calculated for both Ryeland clusters against three breeds derived from British flocks (SUF, SBF and WMT) and Soay (Table 4.1). Regions with high F_{ST} within a RoH may indicate adaptive introgression, with the RoH possibly being the outcome of a hitchhiking effect. Average F_{ST} within a RoH is more typical of positive selection acting on a variant present in both populations but selected only in one, and finally, high F_{ST} outside of a RoH may indicate introgression or a novel mutation.

For both Northern and Southern Ryelands, RoH were summarised as the mean number of RoH per individual, the mean length of RoH, the number and proportion of RoH that exceeded 8 Mb, as well as an inbreeding coefficients derived from RoH: $F_{RoH} = \sum L_{RoH} \; / \, L_{genome}$, where $\sum L_{RoH}$ is the sum of the length of all RoH for an individual and L_{genome} is the length of the genome. Due to the sample size difference between Northern (n = 9) and Southern (n = 51) Ryeland, the differences in RoH statistics could be caused due to the small sample size of the Northern cluster. Therefore, 1,000 bootstrap replicates of 9 individuals randomly sampled from the Southern population were generated and calculated the same RoH statistics in order to obtain a null distribution of the values of those statistics conditional on a sample size of 9.

4.3.4 DEMOGRAPHIC HISTORY

Recent demography for each breed (Table 4.1) was investigated by calculating the relationship between LD and the effective population size (N_e) within the last 50 generations for each breed using SNEP v1.1 (Barbato *et al.*, 2015). Default settings were used, except for implementing the mutation rate modifier from Sved and Feldman (1973) which adjusts for increased probability of multiple recombination rates at higher linkage distances, sample size correction for unphased genotypes, and mutation correction (Ohta and Kimura, 1971). Additionally, N_e Slope analysis (NeS) was used to investigate fine scale slope changes for the estimated N_e curve at each time point.

4.3.5 LANDSCAPE GENOMICS

Long term environmental data (168 environmental variables; Table S4.1) averaged across 1981-2010, formatted into 5 km^2 areas, and assigned to the British National Grid (EPSG:27700) were downloaded from the Met Office (2017). The R package *RASTER* (Hijmans *et al.*, 2019) was used to extract the values for each variable in locations where sheep were sampled. Environmental variables with zero variance were removed.

An additional two geospatial datasets were utilised. Land cover mapping (LCM) data, which uses satellite data to divide land use into 21 separate classes defined by UK Biodiversity Action Plan Broad Habitat definitions, at a 1 km resolution (Rowland *et al.*, 2017) was downloaded from Edina Digimap (https://digimap.edina.ac.uk/). Digital elevation model (DEM) data at 25 m resolution (Copernicus Land Monitoring Service and European Environment Agency, 2019) was downloaded from Copernicus (https://land.copernicus.eu/). QGIS v3.6 (QGIS Development Team, 2019) was used to interpolate additional map layers from the base datasets, resulting in a further 18 environmental variables. Within QGIS, DEM data were reprojected to an Earth-centred projection (WGS84). Hill slope was calculated as the degrees of an incline. As aspect is a circular variable (i.e. North-facing is equivalent to 0° and 360°), aspect with respect to South was generated as a map layer using the equation $A_S = \sqrt{(Aspect - 180)^2}$. Using A_S, South- is 0°, East- and West- are 90°, and North-facing slopes are 180°, producing a linear proxy for insolation (solar exposure). Due to the expanse of each farm, a single geographical point is not representative of the entire land occupied; as a result, extraction of one cell of the relatively high-resolution LCM and DEM data would not encapsulate the land use and geography of the whole environment the sheep are exposed to. Therefore, zonal statistics were calculated for each variable in a circular buffer region with a 2.5 km radius centred on the farm coordinates. Mean and standard deviation for aspect, aspect with respect to South, altitude and slope were extracted from the same buffer regions.

Mean and standard deviation for the proportional cover of 5 potentially relevant classes within the buffer region were extracted from the LCM: improved grassland; neutral grassland; calcareous grassland; acid grassland; and heather grassland.

Three datasets were produced and tested for associations to genotypic data. Principal Component Analysis (PCA) was used to reduce the dimensionality of the datasets. Principal components were kept until cumulatively 80% of the variation of the data was explained, retaining a high proportion uncorrelated variability contained in the data whilst limiting the probability of false positives. This process was applied to: i) Met Office (MO) data; ii) a combination of DEM and LCM derived data; iii) and all three combined, MO, DEM and LCM.

The probability of environmental variables being associated with specific genotypes was calculated using logistic regressions in SAMβADA (Stucki *et al.*, 2017) and assessed with log-likelihood ratios (G). SAMβADA is able to consider population structure during the selection procedure. Due to Ryeland separating into two distinct clusters (K = 2), the coefficient of membership to a single cluster (K-1 clusters) was added to the multivariate modelling. Additionally, a univariate model excluding Northern Ryeland was also tested. Associations between genotypes and environmental variables were considered significant when the G score had a significance probability less than 0.05 after implementing a false discovery rate (FDR) threshold of 5% (Benjamini and Hochberg, 1966). Biological function of genes within 20 kb of the remaining loci were explored. This distance was selected after analysing linkage disequilibrium (LD) decay at increasing distances in Southern Ryeland by calculating r^2 between pairs of markers in PLINK. Furthermore, a distance of 20 kb represents approximately half the median gap size between SNPs in the OvineSNP50 array ensuring the majority of the genome is covered, and at distances greater than 20 kb, LD declines to relatively low levels ($r^2 > 0.2$) in British sheep (Beynon *et al.*, 2015).

4.3.6 SIGNATURES OF SELECTION

In addition to the landscape genomic approach, a population pairwise method to detect selective sweeps was applied. For this, cross-population extended haplotype homozygosity was calculated (XPEHH; Sabeti *et al.* 2007) using the R packager *REHH v3.2.1* (Gautier *et al.*, 2017). XPEHH detects regions in which alleles have reached or are approaching fixation by comparing the extent of linkage disequilibrium around those alleles in two populations (Sabeti *et al.*, 2007), with one of the populations referred to as the test population and the other as the reference population. This test of pairwise comparisons was applied between the Southern Ryeland population and nine other reference sheep breeds (Table 4.1; Kijas *et al.* 2012; Beynon *et al.* 2015;

https://www.sheephapmap.org/) originating from either Britain (UK), Europe (EU) or Asia (AS). Phased haplotype data was first generated for each breed with FASTPHASE V1.4 (Scheet and Stephens, 2006) under default settings and using 10 starts for the expectation-maximisation algorithm. For each reference breed, XPEHH values that exceeded 2 (equivalent to a *p*-value < 0.01) were retained to search for neighbouring genes that may be putative targets of selection. This threshold value represents the degree to which an allele is an outlier and presents a larger LD range relative to the rest of the genome, therefore, indicating positive selection favouring the Southern Ryeland allele compared to the reference breed. Conversely, as XPEHH produces positive values reflecting the results in the test population, and negative values for the reference population, extreme negative values would indicate positive selection favouring the reference breed. Regions 20 kb either side of retained SNPs were scanned for genes contained within using BIOMART (Durinck *et al.*, 2005). Additionally, if two or more adjacent SNPs exceeded the threshold XPEHH value, the neighbouring SNPs were instead considered as a single window. Genes extracted in the XPEHH analysis between Southern Ryeland and each of the UK breeds were considered for Gene Ontology (GO) analysis only if they were present against at least two of the three UK reference breeds. Additionally, a combined median XPEHH score was calculated for Southern Ryeland against UK, EU and AS breeds, to identify common windows under selection across multiple breed pairwise comparisons. Genes within those windows were identified in the same way. The median was selected as it is not as heavily skewed by a single, high magnitude outlier as mean values are.

4.3.7 GENE ONTOLOGY

The Functional Annotation Cluster (FAC) tool from the Database for Annotation, Visualization and Integrated Discovery (DAVID) V6.8 (Huang *et al.*, 2009) was used for GO analysis on the filtered genes to determine significantly enriched user-inputted terms that may be indicative of biological function or processes under selection. Mapping clusters of genes involved in common pathways were also highlighted through KEGG pathway analysis in DAVID.

4.4 RESULTS

4.4.1 DATA FILTERING

A total of 54,241 single nucleotide polymorphisms (SNPs) were genotyped. SNPs relatively uniformly distributed across the genome, with a mean gap size of 50.9 kb and a median

gap size of 42.5 kb. After removing loci with a call rate below 99% and loci with a minor allele frequency less than 5%, the dataset was reduced to 36,326 markers.

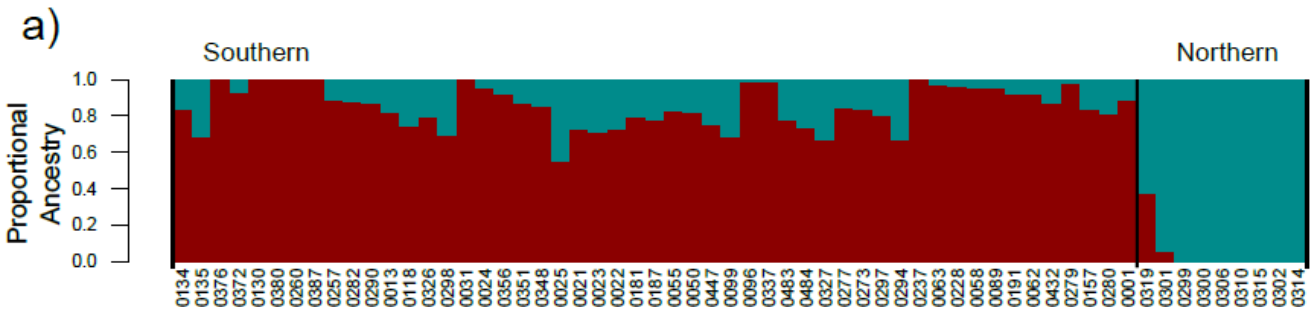
Table 4.1. Sheep Breeds and genetic diversity. Ryeland sheep are subdivided into a Northern and Southern cluster within the breed.

Breed	Abbr.	Number	F_{IS} (± SD)	H_O (± SD)	Group	Ref[a]
Ryeland	RYE	60	0.096 ±0.045	0.388 ±0.019		1
- ***Northern***		*9*	0.132 ±0.043	0.372 ±0.018		1
- ***Southern***		*51*	0.089 ±0.042	0.391 ±0.018		1
Australian Suffolk	SUF	109	0.051 ±0.036	0.408 ±0.015		2
Scottish Blackface	SBF	56	0.088 ±0.025	0.392 ±0.011	Britain (UK)	2
Improved Welsh Mountain	WMT	15	0.028 ±0.018	0.418 ±0.008		3
Swiss White Alpine	SWA	24	0.081 ±0.026	0.395 ±0.011		2
Rambouillet	RAM	102	0.131 ±0.051	0.373 ±0.022	Europe (EU)	2
Castellana	CAS	23	0.055 ±0.027	0.406 ±0.012		2
Awassi	AWS	24	0.113 ±0.045	0.381 ±0.019		3
Qezel	QEZ	35	0.112 ±0.030	0.382 ±0.013	Asia (AS)	2
Afshari	AFS	37	0.111 ±0.035	0.382 ±0.015		2
Soay	SOA	110	0.328 ±0.025	0.289 ±0.011		2
Boreray	BOR	27	0.332 ±0.073	0.287 ±0.032	Outgroup (OT)	2

a - References: 1 - This study; 2 - Kijas *et al.* 2012; 3 - www.sheephapmap.org

4.4.2 POPULATION STRUCTURE AND ADMIXTURE

After pruning against LD for ADMIXTURE analysis, 8,011 markers remained. Population structure analysis with ADMIXTURE produced the lowest CV error at K = 2, preferentially separating Ryeland into two clusters (Figure 4.2). Difference in CV error between K = 1 (0.616) and K = 2 (0.614) was marginal and represented the smallest shift in CV error for all values of K tested (K = 1 to 10, Figure 4.2b). At K=2 the individuals separated based on their ancestry to form two spatially distinct clusters. One cluster was comprised of the animals located in Scotland (n = 9), the most Northern location in the data, and the second cluster was comprised of individuals present across England and Wales (the Southern cluster; Figure 4.2). Additional Admixture analysis of the Southern Ryeland and excluding the Northern Ryeland population, favoured a single cluster in



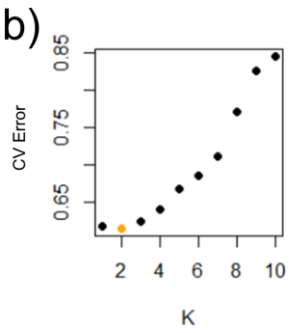


Figure 4.2. ADMIXTURE results for Ryeland sheep, each four-digit number represents a single individual. a) Proportional ancestry at K = 2. Individuals are sorted according to latitude of the sampling location, with more Southerly locations on the left, increasingly more Northerly locations to the right. Individuals are divided according to the majority cluster, Southern (red, n = 51), or Northern (blue, n = 9). b) Cross-validation (CV) error for K = 1 to 10. Lowest value at K = 2.

tests for K ranging between 1 and 4, confirming Southern Ryeland as a single population. Across the two groups, Northern Ryeland were relatively homogeneous with little evidence of crossings with the Southern Ryeland population (average northern ancestry coefficient 0.952 ±0.125 SD), including seven out of nine individuals exclusively belonging to the Northern cluster. On the other hand, the genetic components to many Southern Ryeland individuals presented a higher affinity to the Northern cluster (average Southern ancestry coefficient 0.853 ±0.113 SD). The divergence between these two groups, measured with F_{ST} was 0.062.

A large admixture analysis among 99 global breeds and including Ryeland indicated the preferential clustering solution at K ≈ 130. Ryeland started to for its own cluster at K = 10, and from K ≈ 90 onwards Ryeland displayed the same within-breed partition as when tested in isolation (Figure S4.1). Furthermore, above K = 90 through to the highest K value tested, 135, neither Northern nor Southern Ryeland shared the majority portion of their proportional ancestry with any other breed in the dataset. Northern Ryeland, exhibited significantly higher F_{ST} (0.188 ±0.33 SD) than Southern Ryeland (0.151 ±0.26 SD) against all the global breeds tested (t-test = 37.027; df = 98; $p < 0.001$; Figure S4.2; Figure S4.3). Soay had the highest F_{ST} when compared to both Northern (0.301) and Southern (0.244) Ryeland.

4.4.3 GENETIC DIVERSITY AND IBD

The inbreeding coefficient of global breeds ranged between 0.017 – 0.420 and a mean of 0.144 (Table S4.2), with F_{IS} observed in Ryeland at 0.096 (±0.045; Table 4.1). However, when separated Southern Ryeland had lower inbreeding (0.089 ±0.042) than Northern Ryeland (0.132

±0.043). Higher than average H_O was detected in Ryeland (0.388 ±0.019). Similarly, H_O deviated further from the overall global trend (mean: 0.368; range: 0.249 – 0.422) for Southern Ryeland (0.391 ±0.018) than Northern Ryeland (0.372 ±0.018), suggesting greater genetic diversity in the Southern cluster of the breed. Mean proportional IBD calculated pairwise between Ryeland sheep formed a highly connected network even when plotting only connections that exceeded 10% (Figure 4.3), leaving six individuals unconnected. IBD was statistically no different when considering all (0.028 ±0.070), or just Southern individuals (0.030 ±0.064; Table S4.5). Contrastingly, IBD between Northern individuals (0.242 ±0.204) was significantly higher than either group ($t\text{-test}_{All}$: -6.277; df = 35; $p < 0.001$. $t\text{-test}_{Southern}$: -6.204; df = 35; $p < 0.001$). This indicates that on average the Northern cluster are 2nd-degree relatives with 11 connections having IBD = ~0.5, thus indicating 1st degree relations.

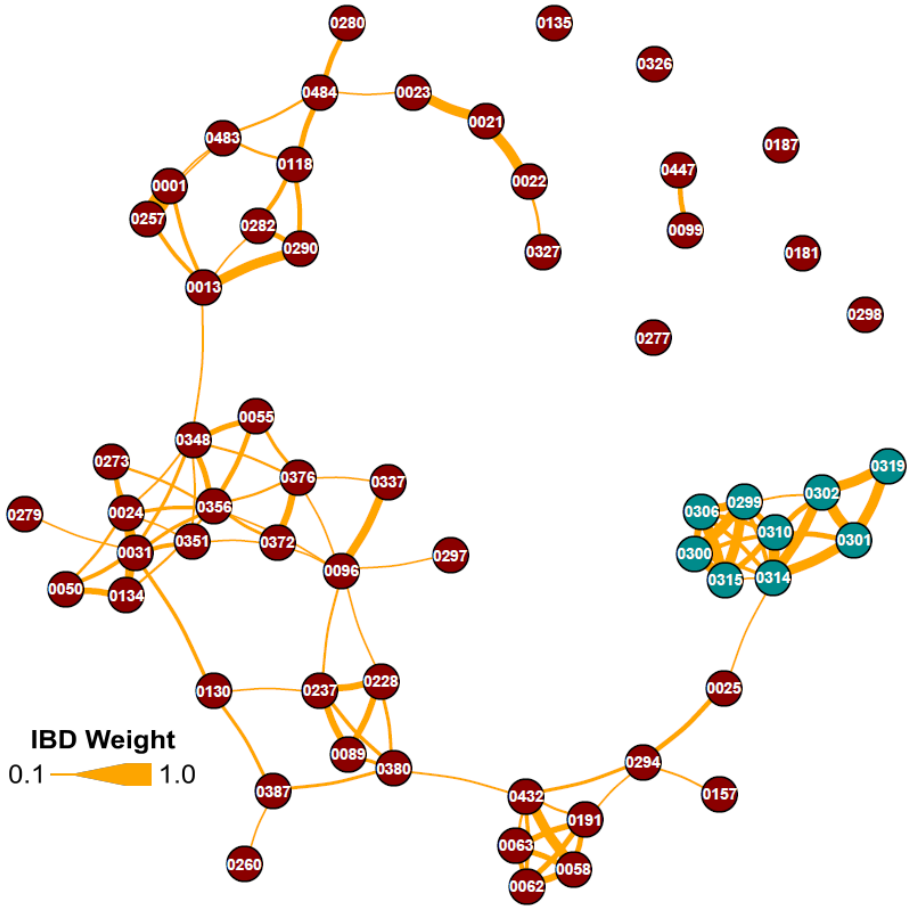


Figure 4.3. IBD network of Ryeland sheep. Nodes represent individuals belonging to Northern (blue) and Southern (red) populations. Edges represent IBD. Only values for IBD > 0.1 are visible, thicker edges indicate higher IBD between individuals. Line length is arbitrary. Raw IBD measures are given in Table S4.5.

4.4.4 RUNS OF HOMOZYGOSITY

RoH and F_{ST} were calculated to attempt to identify the presence of genomic regions that differentiate the two Ryeland groups with respect to each other and with respect to the other sheep breeds tested, and which may represent signatures of introgression from other breeds or adaptation within a single population. As previously shown, windowed F_{ST} was lower for Southern than for Northern Ryeland against each of the test breeds (SOA, SBF, WMT and SUF). However, although the patterns of F_{ST} across the genome varied with some areas presenting higher F_{ST} than others, the trend in F_{ST} across the genome was similar in both Northern and Southern clusters when compared to the test breeds (Figure S4.4). Consistently, the distribution of RoH across the genome paralleled each other in the two Ryeland groups, although RoH were longer and more frequent in Northern than Southern Ryeland. Genomic regions in which only Northern Ryeland presented RoH maintained similar F_{ST} trends across both regional clusters of the breed, implying the causative affect for the difference observed is more likely to be genetic drift rather than introgression or selection.

Table 4.2. Runs of Homozygosity (RoH) derived statistics for Northern and Southern Ryeland. Mean values and one standard deviation reported.

Statistic	Northern (n = 9)	Southern (n = 51)	Southern bootstrap (n = 9)
Number of RoH[a]	61.9 ±8.1	52.7 ±8.1	52.6 ±7.9
Length of RoH (Mb)	7.68 ±6.19	7.18 ±5.55	7.18 ±5.48
Number of RoH >8 Mb[a]	18.1 ±6.5	14.4 ±5.2	14.4 ±5.1
Proportion of RoH >8 Mb	0.292 ±0.087	0.269 ±0.080	0.270 ±0.077
F_{RoH} [a]	0.185 ±0.045	0.147 ±0.036	0.147 ±0.034

a – Northern Ryeland observed value fell outside of 95% quantiles of the Southern bootstrap distribution.

One of the main differences between the two Ryeland groups is the small sample size of the Northern cluster (n = 9) when compared to the Southern cluster (n = 51). In order to determine if the RoH values observed for the Northern Ryeland group were affected by sample size, bootstrapping was used to resample nine individuals from the Southern population 1,000 times and estimated RoH-derived statistics on each of the subsampled data. The observed values for the Northern Ryeland fell outside of the 95% bootstrap quantiles for 3 out of the 5 statistics measured suggesting that the difference between the two groups remains after sample size adjustment (Table 4.2; Figure S4.5). The Northern Ryeland cluster had more RoH per individual (61.9 ±8.1) than Southern Ryeland whether estimated with all 51 individuals (52.7 ±8.1) or the

bootstrapped estimate (52.6 ±7.9). On average, the runs were longer and more variable for the Northern (7.68 ±6.19 Mb) than the Southern population (all: 7.18 ±5.55 Mb; bootstrap: 7.18 ±5.48 Mb). In turn, Northern Ryeland displayed higher F_{RoH}, an inbreeding coefficient derived from RoH (0.185 ±0.045) than the Southern population (all: 0.147 ±0.036; bootstrap: 0.147 ±0.034). Runs exceeding 8 Mb were more frequent in Northern individuals (18.1 ±6.5) than Southern (all: 14.4 ±5.2; bootstrap: 14.4 ±5.1), however, the proportional number of runs exceeding 8 Mb of detected RoH was similar for both Northern (29.2% ±8.7) and Southern (all: 26.9% ±8.0; bootstrap: 27.0% ±7.7).

4.4.5 DEMOGRAPHIC HISTORY

All breeds analysed with the LD approach in SNEP showed a recent population decline between 50 and 13 generations ago (Figure 4.4). Assuming an average generation interval of 4 years (Beynon *et al.*, 2015) this reduction in N_e spans a time period between 52 – 200 years before present. Southern Ryeland was estimated to have a recent N_e of 84, the third lowest N_e of the breeds tested. However, despite the low N_e Southern Ryeland and SOA were the two most demographically stable breeds, retaining 55.1% and 67.4% of their historic N_e (i.e. 50 generations ago), respectively. A comparison of the yearly rate of N_e decline inferred by SNEP was calculated using a linear model for each breed, resulting in low adjusted R^2 value of 0.41 for the Southern Ryeland, similar in magnitude to the N_e decline of OT breeds SOA ($R^2 = 0.38$) and BOR ($R^2 = 0.37$). The remaining breeds exhibited steeper overall declines, ranging from SWA ($R^2 = 0.99$) to QEZ ($R^2 = 3.5$), a higher value indicating greater decline in N_e each year. Finer scale interrogation of the demographic fluctuations was calculated with the NeS analysis that showed the Southern Ryeland declined at a consistent rate between 1840s and 1920s, with an increased decline during the most recent time point in the 1960s (Figure 4.4b). However, in the 1950s, all breeds experience an increased decline, except Ryeland and SOA whose decline did not accelerate.

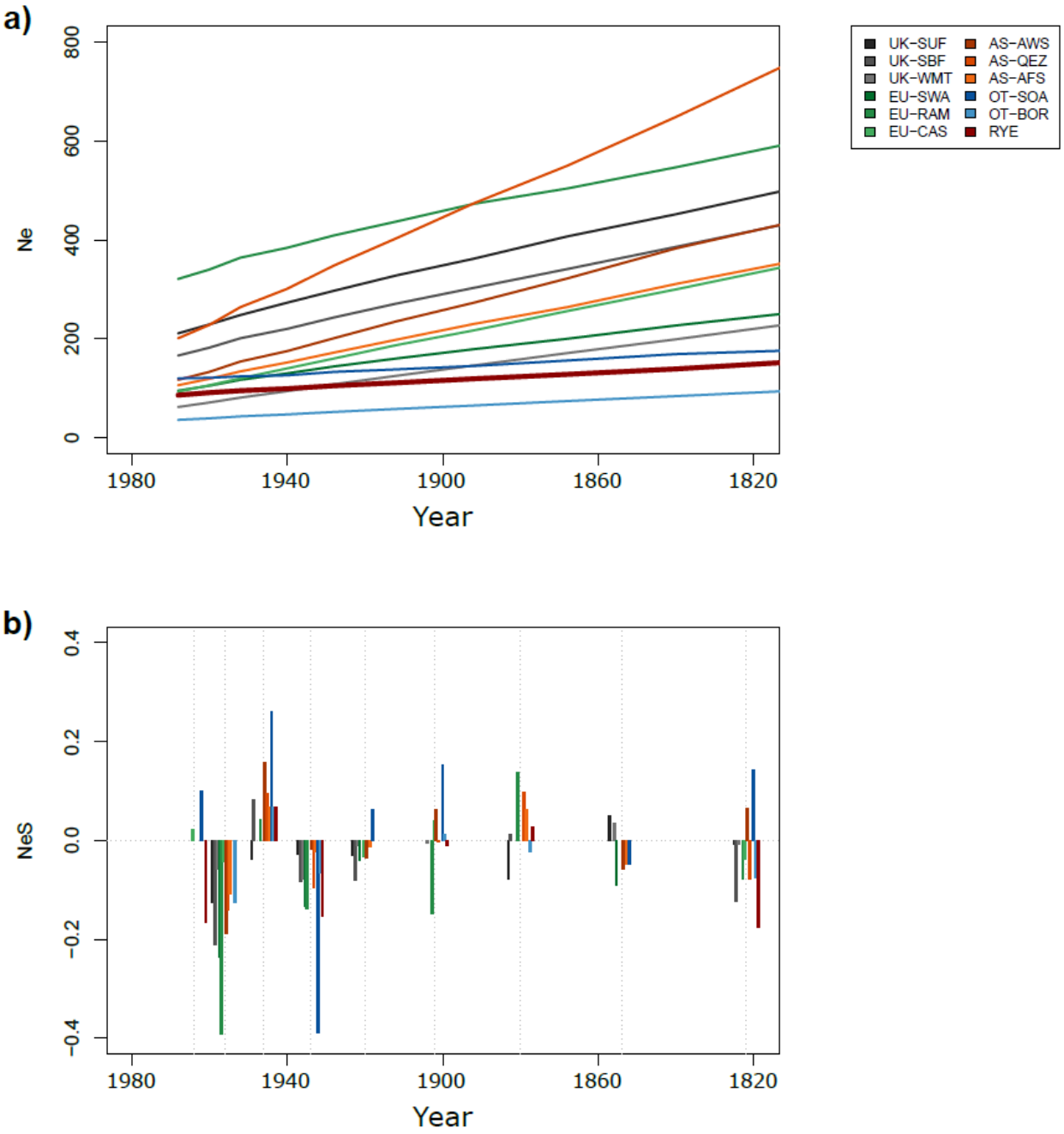


Figure 4.4. Population size trends of eleven sheep breeds between 13 and 50 generations ago calculated using SNeP. Calendar year is estimated using a mean generation time of 4 years. RYE is comprised solely of Southern Ryeland individuals. a) Estimation of N_e. RYE is shown in bold for visibility. b) N_e slope analysis (NeS).

4.4.6 Landscape genomics

PCA was used to reduce the dimensionality of the three datasets containing environmental variables until the cumulative sum of the variation explained by each dataset

exceeded 80%. The MO, combined DEM and LCM, and all combined environmental data, were reduced to 3, 5 and 4 principal components, respectively (Figure S4.6). The landscape genomic analysis correlating environmental variables to the samples' genotypes and accounting for population structure using SAMβADA'S multivariate approach did not result in significant associations (Table 4.3). This was despite the combinations of different datasets and relaxing the previously highly conservative false discovery rates when filtering the initial outputs. Contrastingly, univariate regression model excluding the Northern Ryeland, identified four significant associations between genotypes and environmental variables. These four associations were for the genotypes AA in OAR2_11836099.1 and AA in OAR15_5833133.1, both of which were associated with longitude, as well as for the genotypes AA and AG of the locus OAR4_25295648.1 that were significantly associated with the third component of the reduced environmental variables (Figure S4.7). Twenty environmental variables had r2 correlations greater than 0.70 when compared to longitude, three were related to monthly humidity measurements and the remaining 17 all associated with rainfall or number of rain day per month (Table S4.3). Measures related to windspeed, the presence of snow and slope aspect predominantly contributed the 20 highest loading variables for the third principal component (Table S4.3).

It was not viable to run GO analysis for so few associations, therefore, each locus was investigated separately and scanned for all genes within 20 kb. This distance was selected because LD remained relatively high in Southern Ryeland, with a median r^2 estimates of 0.63 derived from markers with a 20 kb (±0.5 kb) distance (Figure S4.8). Despite moderate linkage remaining even across larger distances (r^2 = 0.43 at 100 ±0.5 kb), beyond 20 kb, r^2 values begin to asymptote.

Table 4.3. Environmental-genotype associations retained after sample correction from landscape genomics analysis in SAMβADA. Threshold for significant associations was a G score less than 0.05 after multiple sample corrections. False discovery rate corrections used to estimate the number of true associations (Benjamini and Hochberg, 1966).

Environmental Dataset	**Reduction Method**	**Variables Remaining**	**Associations after FDR**
MO	PCA	3	0
DEM and LCM	PCA	5	0
MO, DEM and LCM	PCA	4	0
MO, DEM and LCM (Univariate)[a]	PCA	4	4

a - Univariate analysis with only Southern individuals.

Similar observations are made in Beynon *et al.* (2015) among British sheep, albeit at much lower overall levels of r^2.

TMEM123, a gene encoding a cell surface receptor called Porimin, was associated with the loci on chromosome 15 (OAR15_5833133.1[AA]). Porimin is a highly glycosylated transmembrane receptor involved in mediating oncotic cell death, dysregulation of which can lead to cancer (Ma *et al.*, 2002). The heat shock protein encoding gene, *DNAJC25*, associated with the chromosome 2 loci (OAR2_11836099.1[AA]) is also implicated in cell death. *DNAJC25* is highly expressed in liver tissues and increases cell apoptosis, additionally, it is suggested to be a tumour suppressor as downregulation of the gene is common in the liver cancer, hepatocellular carcinoma (HCC; Liu *et al.* 2012). The third locus, on chromosome 4 (OAR4_25295648.1[AA/AG]), is associated with *BZW2* – a member of the basic-region leucine zipper superfamily. *BZW2* is also linked with progression of cancers, with gene knockdown experiments inhibiting cell growth in osteosarcomas (Cheng *et al.*, 2017), muscle-invasive bladder cancers (Gao *et al.*, 2019) and HCC (Jin *et al.*, 2019). Upregulation of *BZW2* causes a significant increase of HCC progression in cell lines and tissues (Jin *et al.*, 2019).

4.4.7 Cross-population selection and gene ontology

Composite XPEHH was calculated using median scores across nine reference breeds from UK, EU, and AS (Table 4.1) and used to investigate selective sweeps unique to Southern Ryeland. 781 out of 36,326 SNPs (2.2%) had a median XPEHH score exceeding 2 (equivalent to a *p*-value < 0.01), indicative of a selective sweep within Southern Ryeland relative to the reference breeds.

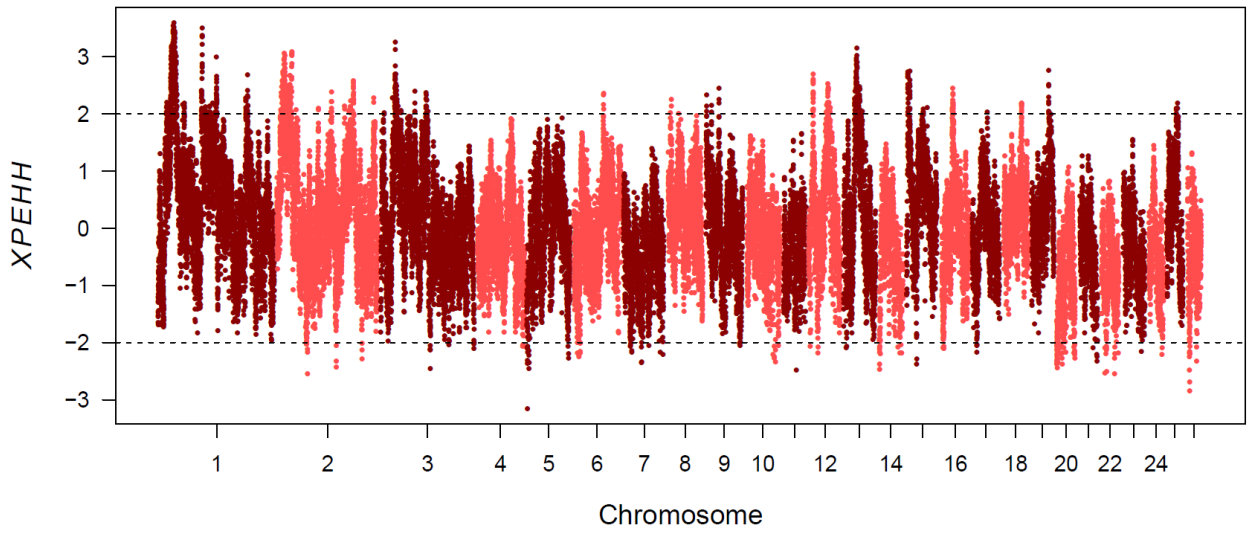


Figure 4.5. Median XPEHH scores for Southern Ryeland sheep against nine sheep breeds. Positive values indicate selective sweeps favouring Ryeland. Loci exceeding XPEHH scores of 2 were used for gene ontology analysis.

Conversely, only 162 SNPs (0.4%) had XPEHH scores less than -2 (Figure 4.5). From the 781 loci, a total of 191 regions were derived from either adjacent outlying SNPs or standalone SNPs, containing a total of two hundred genes. GO analysis identified 12 FACs filtered to a high stringency and 6 enriched KEGG signalling pathways with a *p*-value exceeding 0.05 (Table 4.4).

Individually, the comparisons between the Southern Ryeland and three UK breeds, SUF, SBF and WMT resulted in 240, 238 and 287 regions showing signatures of selection. These regions presented 309, 308 and 290 genes, respectively (Figure S4.9). Out of the 530 unique genes found within these selected regions, 262 were present in at least two of the comparisons and therefore retained for gene ontology. GO analysis resulted in 13 FACs and 9 enriched KEGG signalling pathways were identified, 4 of which are shared with the composite median XPEHH KEGG pathways, despite variation in the causative genes (Table 4.4). Among these, biosynthesis of antibiotics had the most significant p-value in both analyses with up to 12 different genes

Table 4.4. Enriched KEGG signalling pathways for genomic regions under positive selection in Southern Ryeland sheep. Gene list generated from XPEHH analysis against British breeds and median XPEHH values across 9 global breeds. Pathways and genes in bold are shared by both analyses.

KEGG pathway	**Genes**	***P*-value**	**Fold enrichment**
British breeds (UK)			
oas01130:Biosynthesis of antibiotics	**SDHB, CTH, ACSS1, ALDOB, PGM1**, PFKP, **TKT**, **AK4**, FBP2, HSD17B7, GART, **MDH1**	< 0.0001	5.40
oas00030:Pentose phosphate pathway	**ALDOB, PGM1**, PFKP, **TKT**, FBP2	< 0.001	17.08
oas01200:Carbon metabolism	**SDHB, ACSS1, ALDOB**, PFKP, **TKT**, FBP2, **MDH1**	0.001	5.98
oas01100:Metabolic pathways	**ALDOB**, **CYP2J**, **ALG6**, **MTMR2**, **ACSS1**, PIGF, P4HA1, **PLCH2**, **PLA2G12B**, HSD17B7, **SPTLC1**, PGAP1, PFKP, **TKT**, NADK, **AK4**, FBP2, GART, **PANK4**, **SDHB**, **CTH**, ITPA, **PGM1**, **MDH1**, **PYGB**	0.001	1.94
oas00010:Glycolysis / Gluconeogenesis	ACSS1, ALDOB, PGM1, PFKP, FBP2	0.004	7.51
oas04380:Osteoclast differentiation	FCGR2B, JUN, JAK1, PPP3CA, SYK	0.033	4.06
oas01230:Biosynthesis of amino acids	CTH, ALDOB, PFKP, TKT	0.034	5.50
oas05020:Prion diseases	C8A, C8B, C7	0.035	9.91
oas00051:Fructose and mannose metabolism	ALDOB, PFKP, FBP2	0.040	9.29
Median – British, European & Asian breeds			
oas01130:Biosynthesis of antibiotics	**SDHB, CTH, ACSS1, ALDOB, PGM1, TKT, AK4, MDH1**	0.001	4.85
oas01200:Carbon metabolism	**SDHB, ACSS1, ALDOB, TKT, MDH1**	0.010	5.74
oas00030:Pentose phosphate pathway	**ALDOB, PGM1, TKT**	0.019	13.79
oas04060:Cytokine-cytokine receptor interaction	IL12RB2, IL23R, CCL21, LEPR, CCL19, GHR	0.025	3.52
oas04630:Jak-STAT signalling pathway	IL12RB2, IL23R, LEPR, JAK1, GHR	0.034	3.99
oas01100:Metabolic pathways	**SPTLC1, ALDOB, CYP2J, TKT, ALG6, AK4, MTMR2, PANK4, SDHB, CTH, ACSS1, PLCH2, PGM1, PLA2G12B, PYGB, MDH1**	0.041	1.67

responsible for identifying the pathway. The pentose phosphate pathway had the highest fold enrichment, with carbon metabolism and metabolic pathways the other shared enriched KEGG pathways. The most enriched FAC was comprised of genes (*C7, C8A, C8B*; Table S4.4) associated with the complement system, an essential component of the innate immune system that ultimately synthesises the membrane attack complex, causing lysis of pathogenic cells (Fujita, 2002).

4.5 DISCUSSION

Ryeland sheep are one of the more ancient extant breeds in mainland Britain and have also experienced less artificial selection than many of their more recently established commercial conspecifics. The breed therefore provides an interesting model to investigate the effects of prolonged local adaptation and population-specific traits. Population structure was analysed, finding an unexpected division within the Ryeland population, splitting the breed latitudinally. Investigation with F_{ST}, RoH, and IBD suggested population structure was caused by genetic drift with additional admixture analysis confirming the divide was not caused by introgression from other breeds. Demographic analysis corroborated historical reports in Ryeland; overall showing a low, declining population, but also capturing the breed's recently improved stability relative to comparative breeds that continue to decline. Four genotypes significantly associated to environmental variables were identified across the British landscape. Neighbouring genes were implicated in cancer regulation, apoptosis, and, specifically HCC – a form of liver cancer common in ruminants (Nourani and Karimi, 2007; Xia *et al.*, 2015; Machicado *et al.*, 2016). Finally, genes in regions of putative selective sweeps were extracted through conservatively filtered XPEHH analysis and identify related GO terms, including multiple metabolic pathways, biosynthesis of antibiotics and prion diseases.

4.5.1 POPULATION STRUCTURE AND ADMIXTURE

Partitions within minority livestock breeds such as Ryeland is common (Deniskova *et al.*, 2018; Bray *et al.*, 2009), however, the preferential division into Northern (Scottish) and Southern (England and Wales) clusters was not anticipated. The RFBS identified five primary familial lines with highest prolificacy, yet similarly to Kelham et al. (2013), this study indicates no evidence of these forming distinct genetic lineages (Ryeland Flock Book Society, 2019). Furthermore, this differs from phenotypic observations, suggesting that there are three types of Ryeland in the UK: Welsh borders and Herefordshire host shorter, thicker, dark eared sheep; Scotland, North and South West England are notably larger sheep; and, finally, Ryeland with a brown fleece which does not pertain to a specific geographical region (Ryeland Flock Book Society, 2019). Ryelands are traditionally of white fleece colour, however, since the late 1960s a dark "coloured" phenotype

appeared in the breed. Although coloured coats in Ryeland are caused by a single variant of the agouti signalling protein (*ASIP*), it remains unknown how the variant appeared in the breed, however, it is possible to be the result of introgression (Kelham *et al.*, 2013). It is therefore noteworthy that coloured individuals (e.g. RYE_0001, 0021, 0024, 0300 and 0301) do not show any atypical clustering with respect to admixture and IBD analysis, suggesting that these phenotypically different animals are otherwise typical Ryeland animals. For this study attempts were made to capture geographical and genetic variation across the UK in Ryeland sheep, however, sampling was restricted to farmers volunteering blood samples. Nonetheless, the results of this study suggest that the phenotypic clustering and previously suggested geographical divisions are not representative of the overall genetic variation observed in the breed.

Despite phenotypic similarities between Scottish and English (N and SW) Ryeland, the Scottish sheep produce finer wool than their counterparts. The hypothesis that Northern Ryeland may be genetically distinct due to introgression from other breeds was tested, however, the current analyses unanimously refute this. Northern animals display higher differentiation than Southern Ryeland with respect to all other breeds tested and share with them, in no significant part, any ancestry as shown with the admixture analysis (Figure S4.1). This observation is further supported by genome-wide similarities in the distribution of F_{ST} across SNPs between the Northern and Southern clusters. Additionally, 3 of the 5 RoH-derived statistics for the Northern cluster exceed 95% bootstrap quantiles of Southern Ryeland, suggesting that a difference remains between the two groups after sample size adjustment (Table 4.2). It is possible that the Northern population is a subdivision of the Ryeland breed in which introgression has contributed minimally to. This cluster was comprised of only a single flock; this sampling bias of genotyping nine individuals from one farm was a result of a secondary investigation into coat colour and therefore contained individuals with particularly high relatedness (Figure 4.3). Within this flock, selection in which has largely been directed towards improving fleece quality. However, the higher inbreeding, lower genetic diversity (Table 4.1; Figure 4.3), and apparent isolation within this small (n = 9) cluster of related animals is likely to confound selective sweeps with genetic drift. Therefore, the rest of the analyses were focussed on the Southern Ryeland group.

4.5.2 DEMOGRAPHIC HISTORY

Demographic analysis estimating N_e from LD in SNEP showed a decline in all sheep breeds tested over the past ~200 years (Figure 4.4). Interestingly, although with relatively low N_e, Ryeland exhibit the second most demographically stable trend, similar to observations in both Boreray and Soay sheep breeds (Figure 4.4; Kijas *et al.* 2012; Stoffel *et al.* 2020). Notably, these three breeds are the least commercial breeds in the current dataset and the relative stability in N_e

may be attributed to the reduced exposure and intensity of artificial selective pressures and single-sire mating systems often adopted in modern commercial sheep farming. Despite low levels of genetic variation and prolonged small N_e, under sufficient levels of random or disassortative mating, the populations may experience lower inbreeding and prolonged stability in effective population size. An N_e of 84 was estimated in late 1960's as represented by most recent time point of the demographic analysis, shortly before the breed's integration with the RBST (Ryeland Flock Book Society, 2019; Rare Breeds Survival Trust, 2019). By using NeS to analyse the more subtle changes in demography, around the 1930's there is a steep decline as the breed reduced in popularity in Britain in favour of increased meat production from livestock. However, NeS in the 1940's displays a positive value for Ryeland, indicating a shift to a lower rate of decline and an increase in N_e relative to the previous trajectory. During this period, Ryeland were still in decline, although the introgression from Cotswolds and Southdown sheep to improve carcass quality may be responsible for the apparent increase in NeS. Introgression increases heterozygosity and reduces LD, therefore increasing N_e estimations during this time point.

4.5.3 SIGNATURES OF SELECTION

Regions and functional GO clusters associated specifically to Southern Ryeland were identified by filtering for consistent outliers of XPEHH, first with three British breeds and then nine global breeds. Three metabolic pathways were significantly enriched in the KEGG analysis for both XPEHH comparative methods, including metabolic and pentose phosphate pathways as well as carbon metabolism. Focussed on sugar production and processing the selection may be associated to the resilience observed in Ryeland for maintaining healthy form in spite of poor pastures that allowed the sheep to thrive around rye fields (Youatt, 1837). Furthermore, biosynthesis of antibiotics is the fourth and final KEGG pathway resulting from both XPEHH methodologies which could indicate an increased challenge from pathogens. In recent decades, it has been systematic practice to routinely apply antibiotics to entire flocks of sheep (Van Boeckel *et al.*, 2017; He *et al.*, 2020). As a hobbyist's breed and at lower stocking densities, Ryeland are not routinely managed with antibiotics and thus are likely under greater selection to defend against infection.

In addition to anecdotal evidence from RFBS and evidence collected from Dutch Ryeland (Ryeland Flock Book Society, 2019; Melchior *et al.*, 2011), GO analysis resulting from the XPEHH comparison to other British breeds, suggests that the British Ryeland population shows positive selection against TSE (Table 4.4). Prion diseases were propagated by the unknowing utilisation of meat and bone meal from scrapie-infected sheep as supplementary feed for livestock (Bruce *et al.*, 1997). In cattle, the outbreak of bovine spongiform encephalopathy (BSE) was aided by these

cannibalistic feeding regimes among ruminants (Bruce *et al.*, 1997). In turn, consumption of BSE infected meat was responsible for the 1996 emergence of variant Creutzfeldt-Jakob disease, a human TSE, in the UK (Will *et al.*, 1996). The complement system, which has been implicated in the early pathogenesis of TSE, and the prion diseases KEGG pathway were both highly enriched in the UK XPEHH analysis through the genes *C7*, *C8A* and *C8B*, suggesting a strong selective force among these genes within Southern Ryeland (Table 4.4; Table S4.4). During TSE infection, expression of components of the complement system are significantly upregulated, acting to facilitate the movement of prion agents towards lymphoid organs during early development of the disease (Mabbott, 2004). Additionally, the membrane attack complex, which is produced through successful activation of the complement system, is closely associated with increased disease-specific neurodegeneration within the brain tissues of TSE-infected individuals, further implicating these selected genes with prion pathology (Kovacs *et al.*, 2004). For many commercial breeds the application of the National Scrapie Plan for Great Britain (DEFRA, 2001) resulted in reductions to genetic diversity, effective population size and productivity traits as rams with higher risk genotypes were removed from the breeding system, however, this did not affect each breed to the same magnitude (Dawson *et al.*, 2008; Brown *et al.*, 2014). The breed's innate resistance and high ARR allele frequency (Townsend *et al.*, 2005; Melchior *et al.*, 2011) not only helps explain the relative stability observed in the breed (Figure 4.4), but also has facilitated the recovery of a substantial breeding population and subsequent removal of threatened status in recent decades.

Interestingly, during the landscape genomics analysis to detect selection, when analysing all Ryeland simultaneously and ignoring the marginal preference for population structure forming two clusters, 4,553 significant associations remained between genotypes and environmental variables in SAMβADA even when using a highly conservative FDR of 20%. However, adding the observed population structure to the multivariate model eliminated these associations. These spurious results were likely an artefact of environmental gradients occurring in a similar geographical distribution as the Northern and Southern divide observed within Ryeland. Despite the seemingly small difference in CV error for one to two clusters, the landscape genomic analysis highlights the significance of the structure within the breed and the importance of considering population structure in such analyses.

Exclusion of Northern Ryeland and subsequent univariate analysis identified three genes derived from four genotypes. Each gene was involved with regulation of cell death, either oncotic or apoptotic, and thus implicated in cancer progression. Furthermore, both *DNAJC25* and *BZW2* are directly involved in inhibition or acceleration of the same cancer, HCC (Jin *et al.*, 2019; Liu *et*

al., 2012). HCC is widespread across domestic animals, however, it is most prevalent in ruminants – especially sheep (Cullen and Popp, 2002). Indications of selection against this disease extends to the aforementioned complement system, detected in the XPEHH analysis. The components of the complement system are primarily of hepatic origin, and component *C7* is directly associated with the hepatic nodule growth and risk for development of HCC (de Lima *et al.*, 2018; Mercer *et al.*, 2018).

HCC is relatively common in sheep, accounting for 20 percent of all tumours identified in slaughtered sheep, however, studies of gene dependencies and molecular mechanisms are largely comprise of humans or mouse models (Gholami *et al.*, 2006; Nourani and Karimi, 2007). Progression and susceptibility to HCC is characterised by high genetic heterogeneity, whereby polymorphisms within multiple unlinked genes produce similar phenotypes, thus, making the identification of causative links difficult (Dragani, 2010). Nevertheless, some of the strongest signals arising from SAMβADA and XPEHH analyses are implicated in HCC progression, suggesting Southern Ryeland have undergone selective pressures related to this prevalent hepatic cancer.

A significant risk factor associated with the development of HCC is the presence of a liver fluke infection (Xia *et al.*, 2015; Machicado *et al.*, 2016), a possible mechanism is the production of the by-product severin, which inhibits apoptotic processes and advances HCC in human cell lines (Chen *et al.*, 2013; Xia *et al.*, 2015). The common liver fluke (*Fasciola hepatica*) is a parasitic trematode that infects the livers of both sheep and cattle; freshwater snails are intermediate hosts for the parasite, thus infection rates in grazing ruminants is greatest near wet or waterlogged grass (Machicado *et al.*, 2016). Additionally, in the UK, western regions are generally exposed to a greater number of rain days and higher monthly/annual rainfall (Met Office, 2017), this correlation retained within the specific subset of data extracted across farms (Table S4.3). Longitude as a proxy for rainfall, liver fluke exposure and increased risk of HCC thereby provides a potentially causative pathway for the selective pressures observed through landscape genomic analysis in this study. Anecdotally, Ryeland sheep have not been reported to have a prevalence of liver related issues, including liver fluke infections (J. Donovan and I. Lloyd, personal communications, 2nd May 2019), therefore, these results could suggest increased resistance to some hepatic diseases. It is important that this link is further investigated as climate models predict increasingly milder winters causing serious epidemics of *F. hepatica* in Wales by 2050 (Fox *et al.*, 2011); this will have direct implications for not only the Ryeland population but British sheep and production, nationally.

Landscape genomics yielded interesting associations to diseases also linked to the XPEHH analysis free of geographical variables, however, the application of a landscape approach

in a livestock breed with a history such as Ryeland may not be entirely suitable. Although population structure is accounted for with SAMβADA, signatures of selection associated with environmental variables can still be masked by complex demography. Recent bottlenecks within the breed may have removed much of the low frequency variation present in locally adapted individuals. Perhaps the main contributing factor to the lack of environmental and genotypic associations is due to the increasingly mobile trade and sale of sheep between farmers. Individuals are exchanged across large geographical distances disrupting fine-scale local adaptive selection often present in systems with isolation by distance such as wild populations and traditional livestock management (Stucki *et al.*, 2017). Additionally, the transition from a commercial to hobbyist breed alters the magnitude and direction of artificial selection applied by each farmer on their own stock. Often detached from economic gain, motivation for selection differs between farmers based on personal values and aims - some favouring fleece quality, meat production, hardiness or even tameness (Garforth, 2015). Furthermore, ethical background and husbandry practices have been shown to affect the connectivity of flocks in modern livestock practices (Berthouly *et al.*, 2009). Divergent preferences are evidenced by the establishment of coloured flocks, and regular ongoing genotyping of individuals for coat colour variation within the breed society (Kelham *et al.*, 2013). The Northern subdivision in these data, at least in part, may be due to the individual's preference for favouring finer fleece production over other traits; definitive tests could include sampling closely related flocks in the region that are not under selection for fleece quality. This is likely just an example of what is systemic practice within hobbyist breeds and the variation between farmers' priorities. Not only can farmer-driven selection potentially exacerbate population structure within a breed, but it can also confound environmental selection and local adaptation, particularly if (neutrally) traits that are deemed desirable by farmers is a preference shared within a confined geographical region (Abebe *et al.*, 2020). In order to control for such effect, management practice of the flock and selective aims of the farmers could be surveyed, ensuring only sheep under comparable artificial selection are analysed for geographical and environmental adaptation. To this end, the XPEHH results are to be interpreted with more weight than the landscape genomics approach in this particular study in identifying selection within Ryeland, as the latter is susceptible to multiple confounding effects and false positives (Rellstab *et al.*, 2015; Nadeau *et al.*, 2016). The follow up of post-hoc validation in additional datasets or experimental evidence would confirm the presence or absence of the identified loci as selectively advantageous. Nonetheless, the considerable overlap of biological functions and pathways through both methodologies indicates a partial level of validation of the approach.

4.5.4 CONCLUSION

This is the first SNP array study characterising Ryeland sheep. Despite the modern mobility of livestock, this study highlights the importance of maintaining high connectivity within Ryeland – the small flock size and dissimilar selection priorities render the breed vulnerable to genetic drift. Historical demographic trajectories have been most similar to ancient feral breeds such as Boreray and Soay, consistently with small effective population sizes but a relatively low and stable rate of decline. Multiple signatures of selection associated with biological functions and processes were identified, matching both previously stated qualitative and quantitative observations. Although phenotypically broad such as "thriving despite low feed", initial indications of selective sweeps centred around various metabolic processes were numerous. Additionally, the breed's innate immunity to scrapie could be enhanced through the selection of genotypes within the complement system. However, it would be invaluable to genotype these samples for the prion protein gene and investigate the frequency of 'resistant' polymorphisms. Unexpectedly, genes related to HCC were prominent in GO analyses with possible associations between liver fluke abundance; investigation into the rates of HCC and liver cancer within the breed could provide insight into the nature of these associations and possible resistances. If selection signals are functionally validated, breeding programs could be initiated focussing on genotypes involved in the HCC resistance pathways. This could be initially validated within the Ryeland sheep population, but potentially has far reaching implications, particularly within the Welsh commercial sheep industry which is predicted to be facing significant pressure from increased liver fluke abundance and temporal persistence in the coming decades.

4.6 ACKNOWLEDGEMENTS

Thanks to the Ryeland Flock Book Society (RFBS) for the sampling of animals analysed in this chapter, and specially to Ifan Lloyd and John Donovan (RFBS Genetics Subcommittee), and Simon Donovan for coordinating sampling and the useful discussions. Thanks to the ClimGen Project for generating and providing the raw genotyping data used in this chapter.

Chapter Five

Traits associated with fertility maintain heterozygosity within the feral Chillingham cattle

---

5.1 ABSTRACT

Measurements of the genetic health of small populations have often been based on genome-wide or neutral marker levels of variation which are used as proxies for quantifying inbreeding depression and genetic drift. However, the availability of high-quality sequence data and well annotated reference genomes has resulted in a greater understanding of genomically localised functional genetic diversity and its role in long-term survival. Whole-genome resequencing data of 10 Chillingham feral cattle (*Bos taurus*) from a highly inbred, small, yet viable herd situated in Northumberland, England, was used comparatively with 10 Angus cattle individuals from the 1000 Bull Genomes Project to investigate the distribution and function of variation across the genome. Regions identical-by-descent (IBD) detected as runs of homozygosity (RoH) with individual- and population-level coverage were used to estimate inbreeding (F_{RoH}) and recurrence of RoH, respectively. On average, 91% of the Chillingham genome was within a RoH, with 77.2% of SNPs included in a RoH across all 10 samples. Contrastingly, regions of peak heterozygosity (HPW) were also identified, forming 511 discrete windows with high clustering. HPW in Chillingham had elevated counts of loci with a significant excess of heterozygosity, potentially resulting from balancing selection. These regions were enriched for fertility quantitative trait loci (QTL) as well as milk QTLs, however, they were devoid of immunity QTLs. The major histocompatibility complex (MHC) had high-levels of homozygosity, did not overlap HPW, with variation only present around BOLA-NC1, a gene linked with maternal immunity in pregnant cows. These findings emphasise the extreme homozygosity that can persist in inbred populations while maintenance of functional variation may be prioritised towards fertility rather than immunity.

5.2 INTRODUCTION

Conservation efforts and management strategies have often been informed through genetic diversity of genome-wide markers within a population (Reed and Frankham, 2003; Spielman *et al.*, 2004). This is particularly the case regarding endangered species with a low genetic diversity and the assumption that it is an indicative measure of, firstly, strong genetic drift increasing mutational load, through the fixation of mildly deleterious alleles (Lande, 1994), and, secondly, mating of related individuals leading to inbreeding depression, through the increased expression of recessive deleterious alleles in homozygous individuals (Charlesworth and Willis, 2009). Higher levels of inbreeding are frequently associated with lower reproductive fitness,

reduced evolutionary potential and an elevated extinction risk (Spielman *et al.*, 2004). Furthermore, reductions of genetic diversity – in particular allelic richness and heterozygosity – can effectively predict 'threatened' populations under the IUCN conservation rankings (Willoughby *et al.*, 2015).

Quantifying inbreeding depression can be challenging as it is more likely that strongly deleterious mutations are purged from small population (Hedrick and Garcia-Dorado, 2016). Purging selection is the process in which deleterious alleles are reduced to lower frequency in the population, potentially to the extent of removal. This usually occurs during a severe or extended bottleneck or as a result of inbreeding depression and the heightened expression of recessive highly deleterious alleles from consanguineous mating (Crnokrak and Barrett, 2002). Whilst strong purging events may result in the removal of linked variation and thus affect genome-wide estimates of genetic diversity, it has also been correlated to improved fitness in species including birds, ibex, foxes and cattle (Crnokrak and Barrett, 2002; Laws and Jamieson, 2011; Purfield *et al.*, 2012; Robinson *et al.*, 2016; Grossen *et al.*, 2020).

Contrastingly, balancing selection results in the persistence of variation in surrounding loci through mechanisms including heterozygote advantage, negative frequency-dependent selection and pleiotropy (Hedrick, 2011; Key *et al.*, 2014; Croze *et al.*, 2016). Under strong enough selection, balanced loci may maintain intermediate allele frequencies despite exposure to genetic drift (Piertney and Oliver, 2006). While these mechanisms can influence any genomic region, some of the most polymorphic genes in vertebrates belong to the major histocompatibility complex (MHC) and are the most prominent genes classically investigated for balancing selection (Hedrick, 1994). The MHC is a multigene family that are central to the vertebrate immune system. Predominantly, the MHC genes encode cell surface-expressed glycoproteins that trigger the activation of the immune response by presenting self- and pathogen-derived antigens to T-cells. Despite evolutionary divergence in architectural and sequence-level variation of the MHC complex, homologous gene function and key residues are often highly conserved (Piertney and Oliver, 2006). This can extend across broad taxa such as avian and mammalian class I peptide-loading complex (Hinz *et al.*, 2014) and T-cell CD8 receptors across chicken, swine and bovine (Liu *et al.*, 2016). Variability across MHC genes is correlated with the diversity of cell-surface receptors observed in T-cells, with each receptor binding to a restricted set of peptides (Hedrick *et al.*, 2001). Due to codominance observed across many MHC genes, increased diversity therefore confers immunity to a wider array of pathogens, thus, heterozygosity is often maintained in these regions through balancing selection (Hedrick *et al.*, 2001; Piertney and Oliver, 2006; Giovambattista *et al.*, 2013). Theoretically, while the persistence of a small population with high

levels of inbreeding depression is possible through mechanisms such as purging and balancing selection. In practice, in natural systems the erosion of genetic diversity often elevates extinction risk to the extent that stochastic events (e.g., disease, food availability, demographic events and natural catastrophes) drive the population to extinction (Spielman *et al.*, 2004; Willoughby *et al.*, 2015).

This study focusses on an inbred, isolated *Bos taurus* cattle breed which has undergone a long-term and numerically extensive bottleneck, but that seemingly exhibits the genomic signatures of inbreeding and low-level variability yet retain pronounced levels of fitness. Chillingham cattle are a feral breed inhabiting an area of 1.34 km^2 at Chillingham Park in Northumberland, UK, and have remained a closed herd for at least the past 300 years (Visscher et al., 2001; Hall and Bunce, 2019). Some records indicate that the breed's isolation may have begun as early as the 13^{th} century (Hall and Hall, 1988). In 1947, the herd went through a severe bottleneck, with 13 individuals remaining (5 males; 8 females), however, this has recovered to 108 individuals as of February 2019, with an estimated carrying capacity of the site at 120 (Visscher *et al.*, 2001; Hudson *et al.*, 2012; Hall and Bunce, 2019). Analysis of 25 microsatellites loci revealed 24 to be homozygous (Visscher *et al.*, 2001) and a sample of 8 individuals presented the same mitochondrial DNA haplotype (Hudson *et al.*, 2012). At least one recorded genetic bottleneck coupled with an absence of gene flow and prolonged inbreeding has seemingly resulted in near complete homozygosity across the genome (Visscher *et al.*, 2001; Hudson *et al.*, 2012; Williams *et al.*, 2016). More recent studies using high density (~770 k loci) SNP array data estimate Chillingham present 2.6% heterozygous genotypes with an inbreeding coefficient (F_{IS}) of 0.924 (Williams *et al.*, 2016). From these data, a subset of loci corresponding to the Bovine medium density SNP arrays (~54 k loci) showed that the Chillingham displayed the highest F_{IS} and lowest heterozygosity relative to the global metapopulation – an order of magnitude lower than the majority of 55 other cattle breeds (Orozco-terWengel *et al.*, 2015). The high degree of homozygosity has been attributed to, inbreeding, genetic drift and widespread purging selection (Williams *et al.*, 2016). Furthermore, an autocorrelation analysis indicated that the minimal variation that is observed across the Chillingham genome forms a non-random distribution (Williams *et al.*, 2016), providing potential targets of regional balancing selection. Currently, there is limited published data on the MHC in Chillingham, despite a hypothetical hotspot for balancing selection and the critical role the region has in the recognition of pathogens and parasites – often associated with the stochastic threats that small, inbred populations are vulnerable to. Visscher *et al.* (2001) observed an absence of heterozygosity across three microsatellites located around the MHC region, further confirmed by Ellis and Hammond (2014) who cite complete homozygosity in all class I and class II loci.

Chillingham cattle have been closely monitored for 7 decades and remain relatively unmanaged with no human-mediated breeding system or forced sex ratio (Visscher et al., 2001; Hall, 2006). Supplementary feeding only occurs in winter months, where up to two thirds of the herd's diet is replaced with the provided hay (Towers et al., 2017). Medical intervention is sparse, with even bovine tuberculosis screening permissibly replaced by the autopsy of at least one animal a year (Hall and Bunce, 2019). Despite a lack of management and the apparent minimal variation both genome-wide and at the MHC regions, the breed remains viable, showing no obvious signs of inbreeding depression or susceptibility to disease (Hall and Hall, 1988; Visscher *et al.*, 2001; Hall, 2006). Genetic load and the effect of deleterious alleles may remain subtle and incredibly challenging to detect, although few notable congenital conditions are seemingly present among the population (Hall and Bunce, 2019). There are a few known persistent medical abnormalities present in the Chillingham population, including increased levels of molar defects (Ingham, 2002), testicular hypoplasia (Hall *et al.*, 2005), and subfertile males with poor semen quality (Hall and Bunce, 2019). The latter two occur commonly within inbred commercial breeds and exhibit a partially causative relationship (Steffen, 1997).

The availability of a high-quality *Bos taurus* reference genome and high-throughput whole-genome resequencing allows more precision when detecting variable sites and genomic segments that are identical-by-descent (IBD; Kardos *et al.*, 2017). Inbreeding results in the offspring receiving lengths of identical haplotypes from related parents, which have ultimately been acquired from a single common ancestral source. IBD segments are detectable as extended regions of homozygosity, known as runs of homozygosity (RoH; Marras *et al.*, 2015). Inbreeding coefficient can be derived from proportion of the autosomes covered by RoH with the timeframe correlated to the length of RoH, in that longer RoH are indicative of more recent inbreeding events as recombination is yet to shuffle haplotypes (Browning and Browning, 2013). Furthermore, the population-level of inbreeding can be assessed through the consistency at which loci appear within RoH across individuals.

This study aims to utilize whole-genome resequencing data to identify SNPs to determine levels of variation within Chillingham relative to a reference breed. Angus was chosen as the reference breed due to the breed's relatively high relatedness to Chillingham (Orozco-terWengel *et al.*, 2015) and ease of availability and high coverage through the 1000 Bull Genomes Project (Daetwyler *et al.*, 2014; Hayes and Daetwyler, 2019). In this study, the validity of previous array-based conclusions on inbreeding and variability is assessed. Additionally, regions of peak heterozygosity will be identified, and clustering analysed and contrasted with regions of high homozygosity, determining recurrence at a population-wide level. The possible selection

mechanisms with underlying causative effects in peak heterozygosity windows will be investigated, with a particular focus on balancing selection. The functional phenotypic consequences of these regions will be assessed with quantitative trait loci, including traits related to fertility, immunity, and different aspects of production. Contrastingly, the MHC region will be scrutinised for the apparent lack of heterozygosity. Finally, a brief analysis of demographic history to contextualize the results arising from a small, inbred population.

5.3 METHODOLOGY

5.3.1 SAMPLING, RESEQUENCING, AND ALIGNMENT.

Blood samples were retrieved post-mortem from ten feral Chillingham (CIL) white cattle individuals from Chillingham Park, Northumberland, England, with equal sampling of male and female. Novogene Co., Ltd. (https://en.novogene.com/) was commissioned to carry out DNA extraction, library preparation and Illumina sequencing, producing standard 150 bp long paired-end sequences from 500 bp fragments with 200 bp gap. Raw reads were first examined using FastQC v0.11.8 (Andrews *et al.*, 2015). The data were then processed in accordance with guidelines from the 1000 Bull Genomes Project (Hayes and Daetwyler, 2019) up until variant calling. Raw reads were filtered using TRIMMOMATIC V0.38 (Bolger *et al.*, 2014) using five sequential parameters: all reads were trimmed at 5' and then 3' until phred quality scores exceeded 20 (LEADING:20 TRAILING:20); remaining base calls were discarded from the point that a sliding window of three bases had an average quality of less than 15 (SLIDINGWINDOW:3:15). Average quality of the trimmed read had to exceed 20 and length had to exceed 35 bases (AVGQUAL:20 MINLEN:35).

Ten Angus (AAN) individuals were retrieved from the 1000 Bull Genomes Project (Daetwyler *et al.*, 2014) to generate a dataset of another archetypical British breed against which the Chillingham could be compared. Angus were chosen due to their close phylogenetic relationship to Chillingham based on SNP array data from Orozco-terWengel *et al.*, (2015). Samples were retrieved from the NCBI Sequence Read Archive (Leinonen *et al.*, 2011); project accession code SRP039339; individual accession codes SRX547877, SRX527500 - SRX527508. Data were in SRA format, downloaded with *prefetch* and converted to BAM format using *sam-dump* from the SRATOOLKIT V2.9.3 (Leinonen *et al.*, 2011). As these sequences were aligned to an outdated reference genome (UMD3.1 bovine assembly [RefSeq:GCF_000003055.5]) they were

first reverted to trimmed FASTQ format with SAMTOOLS V1.5 (Li *et al.*, 2009) and processed in tandem with Chillingham samples.

Remaining trimmed reads were aligned to the *Bos taurus* ARS-UCD1.2 reference genome (Rosen *et al.*, 2018) with the addition of Btau5.0.1 Y chromosome (Bellott *et al.*, 2014). Alignment was carried out using BWA V0.7.15 *mem* command (Li and Durbin, 2009) with optional parameters to mark shorter split hits as secondary and to assign read groups for downstream base recalibration. Aligned files were converted to BAM files and indexed using SAMTOOLS V1.5 (Li *et al.*, 2009). PCR and optical duplicates were marked using PICARD V2.18.14 (Broad Institute, 2019), optical duplicate pixel distance was set to 2500 due to sequencing being arrayed flowcell data.

Base quality score recalibration was carried out with the GENOME ANALYSIS TOOLKIT (GATK) V3.8.1 (McKenna *et al.*, 2010) BaseRecalibrator function, using default settings apart from a modified bqsrBAQGapOpenPenalty of 45 which has been shown to work better for *Bos* spp. (Hayes and Daetwyler, 2019). The list of 'known variants' to be used as a training dataset was acquired from the 1000 Bull Genomes Project (Hayes and Daetwyler, 2019) and combines variants from dbSNP build 150 (Sayers *et al.*, 2019) and 1000 Bull Taurus-Indicus Run6 (Hayes and Daetwyler, 2019). Variants originally mapped to the UMD3.1 reference genome coordinates were realigned to ARS-UCD1.2. Two passes of base recalibration were carried out before convergence between empirical and reported qualities was achieved.

Variant calling was performed using GATK HaplotypeCaller and both breeds were genotyped using GATK GenotypeGVCFs. Only autosomal biallelic SNPs were retained and then filtered for QD > 2.0, FS < 60, ReadPosRankSum > -8.0, SOR < 3.0, mapping quality (MQ) > 40 and Z-score from Wilcoxon rank sum test of alt vs. ref read mapping qualities (MQRankSum) > -12.5. SNP genotypes were set to 'no call' if individual read depth (DP) < 6, DP > 20 (approximately twice the mean read depth) or phred-scaled genotype quality (GQ) < 20.

Additionally, sites where all called genotypes were homozygous for either alt or ref allele were removed. Variants that were heterozygous in every individual in either Chillingham or Angus were removed to avoid mapping errors arising from genome duplication. SNPs were retained if they had a minor allele count ≥ 2, a call rate ≥ 70% across all individuals and a call rate ≥ 70% across Chillingham. Missingness, heterozygosity and inbreeding coefficients (F_{IS}) were calculated for each individual using VCFTOOLS V0.1.15 (Danecek *et al.*, 2011). Genetic relatedness using IDB between individuals within each breed was calculated in PLINK.

5.3.2 ANNOTATION

Variants were annotated using Ensembl's VARIANT EFFECT PREDICTOR (VEP) v98 (McLaren *et al.*, 2016) based off of the ARS-UCD 1.2 cattle reference genome annotations. Sites were classified as synonymous if annotated as either stop retained variant or synonymous variant, and non-synonymous if annotated with stop lost, start lost, stop gained or missense variant, all of these categories exclusively occur in coding regions.

5.3.3 RUNS OF HOMOZYGOSITY

Genomic stretches with identical alleles on both chromosomes – Runs of Homozygosity (RoH) – were detected across the 20 individuals using the package BCFTOOLS/ROH (Narasimhan *et al.*, 2016). Detecting RoH in genotype data often relies on approaches using either static or sliding windows (e.g. **chapter four**). These methodologies have been shown to consistently overestimate the size of RoH and have false positive rates exceeding 10% (Narasimhan *et al.*, 2016), further reflected in less accurate estimations of RoH derived inbreeding coefficients (Forutan *et al.*, 2018). The package BCFTOOLS/ROH (Narasimhan *et al.*, 2016) instead applies a hidden Markov model to genotype likelihoods and phasing information generated by GATK HaplotypeCaller. The transition probabilities within the model are influenced by the incorporated genetic map and allele frequencies across samples and the derived recombination rates between markers. Extended RoH are unlikely to arise by chance and are indicative of autozygous tracts across the population (Marras *et al.*, 2015). Overall, the method is more effective at identifying regions of autozygosity within populations rather than regions identical by state within an individual (Narasimhan *et al.*, 2016).

RoH were summarised as in **chapter four**. Additionally, to assess breed-wide conservancy of homozygous regions, the frequency at which each SNP was included within a RoH across both Angus and Chillingham was calculated.

5.3.4 HETEROZYGOSITY PEAK ANALYSIS

Heterozygosity peak analysis was carried out for each individual, calculating the fraction of heterozygous genotypes for called sites within a window of 100 kb (Robinson *et al.*, 2016). This was extended to genome level using sliding windows with a step size of 10 kb, windows with 10 or fewer SNPs were discarded to reduce skewing by small sample size. To identify the most extreme windows, or "peaks", a null distribution of expected window heterozygosity scores was first built for each individual using 100,000 windows randomly sampled, with replacement across the genome. All test windows that fell outside of the upper 5% of the distribution were isolated as

outliers for the given individual, and windows that were defined as outliers in at least five individuals across the breed were considered as heterozygosity peak windows (HPW). Overlapping peak windows were merged to assess their relative clustering. Using this method three different types of comparisons were carried out: (i) Chillingham individuals in Chillingham-defined HPW (CIL_{HPW}) as the test region of focus; (ii) Angus individuals in Angus-defined HPW (AAN_{HPW}) providing a baseline for within-population HPW; and (iii) Angus individuals in Chillingham-defined HPW ($AAN_{CIL\text{-}HPW}$) providing a baseline for the specific HPW identified in Chillingham.

To test whether balancing selection was contributing to the maintenance of HPW, Hardy-Weinberg equilibrium (HWE) was calculated for each site and compared between HPW and the remainder of the genome. For each biallelic genotype combination possible in a population of 10 individuals, significant excess of heterozygotes and deviation from HWE was estimated using Pearson's Chi-squared test with a simulated *p*-value of 0.05 based on 100,000 replicates. Thereby producing a matrix of varying proportions of individuals either homozygous for the reference allele, heterozygous or homozygous for the alternative allele and an assigned *p*-value (for example: 1/8/1 = 0.0043; 2/7/1 = 0.03921; 5/0/5 = 1.00). This significance was then assigned to each SNP with a matching population-wide genotype frequency. An indication and quantification of any over-representation in HPW of sites with a significant excess of heterozygosity was summarised for CIL_{HPW}, AAN_{HPW} and $AAN_{CIL\text{-}HPW}$ with:

$$H = \frac{h_{HPW}/h_{genome}}{n_{HPW}/n_{genome}}$$

where h is the number of sites that are not in HWE and are displaying a significant excess of heterozygosity and n is the number of SNPs – either within *HPW*, or across the *genome*. A value of H = 1 indicates that the proportion of SNPs with a heterozygosity excess is the same for HPW and the whole genome, thus, higher values of H indicate an increasing presence of heterozygosity excess within HPW relative to the rest of the genome and may be indicative of balancing selection. A null distribution was generated using 1,000 bootstrapped replicates of the H score sampled with a quantity of random genomic SNPs equal to n_{HPW}. *p*-values were estimated as the proportion of replicates with H scores greater than the experimental value.

Annotations produced by VEP were extracted for exonic variable sites occurring within HPW for each breed. Potentially deleterious non-synonymous variants were predicted with the SORTING INTOLERANT FROM TOLERANT (SIFT) V5.2.2 algorithm (Sim *et al.*, 2012). SIFT predicts the probable impact of the amino acid substitution from a non-synonymous variant by comparing protein homology, under the assumption that variants disrupting regions that are highly conserved across

taxa are more likely to be deleterious. A SIFT score and qualitative tolerance is applied to each variant derived from the normalised probability that the alternative residue is tolerated, with probabilities below 0.05 deemed as deleterious.

Genome wide and within HPW linkage disequilibrium (LD) was calculated through the measurement of r^2 between loci for each breed using VCFTOOLS v0.1.15 (Danecek *et al.*, 2011). LD was calculated for all pairs of SNPs within a maximum distance of 7,500 kb. The uncharacteristically large span was selected due to the assumption that covariance of loci within the Chillingham genome may extend beyond typical levels for the species based on observed inbreeding and previous reports of LD (Williams *et al.*, 2016). Due to the extensive file size (~1 TB) of data generated from pairwise comparisons of LD in the Angus genome, a random 1% sample was extracted from the r^2 calculations for the genome wide AAN dataset, resulting in ~360 million SNP pairwise comparisons. LD decay was calculated by grouping r^2 measurements into bins based on the physical map distance between the paired SNPs. A total of 150 bins spanning 50 kb each were categorised across the 7,500 kb distance. A distribution of the difference between the binned means for genomic and HPW r^2 (Δr^2) for each breed was compared with Kolmogorov-Smirnov tests.

5.3.5 QUANTITATIVE TRAIT LOCI

To assess the potential phenotypic impact of HPW, quantitative trait loci (QTLs) were collated from BOVINEMINE v1.6 (Shamimuzzaman *et al.*, 2020) and tested for enrichment within HPW. QTLs spanning more than a single nucleotide were removed, resulting in 105,918 entries for 390 unique traits used for further analysis. For each trait, significance was defined using a Fisher's exact test using a 2x2 contingency table – the first variable concerning the counts of QTLs matching/mismatching the given trait, the second variable identifying the presence within/outside of CIL_{HPW}. The test was made more conservative by removing one count from the QTLs matching the trait in CIL_{HPW} and adding one count to QTLs matching the trait in the rest of the database, similarly to the enrichment methodology implemented in DATABASE FOR ANNOTATION, VISUALIZATION AND INTEGRATED DISCOVERY v6.8 (Huang *et al.*, 2009). An alpha value of 0.01, with Bonferroni multiple sample correction was applied to filtered significance of enriched terms.

5.3.6 MAJOR HISTOCOMPATIBILITY COMPLEX

The bovine MHC class I, class II and class III regions were isolated from chromosome 23. The class II region is further divided into two subregions – IIa and IIb – due to a separation of approximately 18 Mb and 15cM (van Eijk *et al.*, 1995). Class IIa and IIb were treated as independent regions and classes in this study as the analysis is focussed on variation at a genomic

level rather than class-level functional effect. MHC genes from Behl *et al.* (2012) were identified in the annotations for the updated ARS-UCD 1.2 reference genome. Heterozygosity, SNP density and HWE in the MHC regions for CIL compared with genomic levels and AAN. To capture adjacent upstream and downstream effects, a buffer of 50 kb either side of the most peripheral genes were incorporated into the region.

5.3.7 Demographic history

A mappability mask from the ARS-UCD1.2 reference genome was created with SNPable Regions (Li *et al.*, 2009) to determine regions on which short reads can be uniquely mapped using all overlapping 100-mers. Within population effective population size (N_e) and cross-coalescence rates between populations were estimated across time with the Multiple Sequentially Markovian coalescent (MSMC2) approach (Schiffels and Durbin, 2014; Malaspinas *et al.*, 2016). Starting with base quality score recalibrated reads produced with GATK (see above; prior to variant calling), two haploid samples per individual from all 20 samples were used for this analysis. Individual VCF and mask files were generated with SAMTOOLS V1.5 (Li *et al.*, 2009) and the bamCaller.py script from the MSMC tool repository (https://github.com/stschiff/msmc-tools), only retaining sites with MQ and base quality of at least 20. MSMC2 was run within Chillingham, within Angus and between Chillingham and Angus, each time comparing the haploid genomes of all individuals of the respective groups under default parameters. From the MSMC2 output, the cross-coalescence rates between populations were calculated on a combined set of time interval boundaries with combineCrossCoal.py from the MSMC tool repository. Coalescent-scaled units were converted to biological units assuming a generation time of 5 years and a substitution rate of 1.25 $x10^{-8}$ per site, per generation (Gautier *et al.*, 2016). To estimate divergence time, the cross-coalescence rate was used, defined as the most contemporary time point at which the cross-coalescence rate was at or above 0.5. Similarly, the range for the divergence time estimate was calculated using 0.25 and 0.75 as the lower and upper bounds, respectively. The initial time point at which the cross-coalescence rate is at or above 0.01 was interpreted as the cessation of isolation between breeds.

5.4 Results

5.4.1 Resequencing and alignment

Following the trimming and removal of low-quality reads from 10 CIL and 10 AAN individuals, 99.85% of paired-end reads (2.4 billion from CIL and 2.5 billion from AAN) were

mapped to the ARS-UCD1.2 reference genome. Of these, 98.52% were properly paired identifiable by correct orientation and with acceptable separation between pairs. PCR/optical duplicates represented 12.8% (11.8-13.9%) of reads in CIL and 4.8% (1.2-8.9%) in AAN. Average read length was consistent within each breed, with lengths of 147 for CIL and 95 for AAN. Genome-wide sequencing depth was above 10-fold coverage across all CIL, with mean depth per individual ranging from 10.8 to 13.8, a single individual in AAN had a sequencing depth exceeding 10-fold coverage with the breed ranging from 7.7 to 10.5 (Table S5.1 /Figure S5.1). Base quality score was reduced from an average of 36.0 to 27.8 after successful recalibration and convergence of empirical and reported scores.

5.4.2 VARIANT ANALYSIS

Variant filtering resulted in 3,447,119 quality-controlled biallelic SNPs across both breeds. When considering only sites that are polymorphic within populations, Chillingham (367,619 variants) had approximately 10% the number of variable sites as Angus (3,244,980 variants) and with 301,213 shared between the two breeds (Figure 5.1). The majority of variants for CIL (87.5%) and AAN (89.5%) had been previously identified across *Bos* spp. However, a total of 12,645 and 308,072 novel SNPs that are unique to the breed and not within the training dataset were identified in CIL and AAN, respectively. After quality control, missingness of called sites averaged 22.8 ±10.7% across individuals and ranged between 7.3% to 53.0% (Table S5.2). Missingness was higher overall in AAN (30.7 ±8.7%) than CIL (14.8 ±4.9%). Observed heterozygosity averaged 0.490 ±0.067 in AAN and 0.044 ±0.006 in CIL with approximately 0.475 and 0.051 heterozygous sites per 1,000 base pairs in individuals of the respective breeds (Table S5.2). F_{IS} derived from

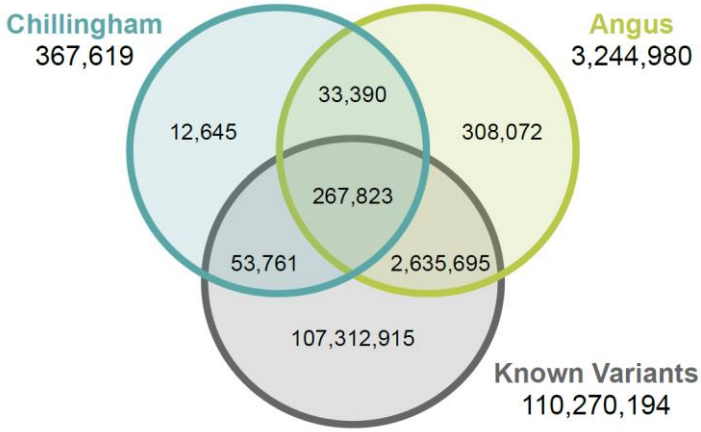


Figure 5.1. SNP segregating between Chillingham, Angus and previously identified "known variants" (defined as variants present in dbSNP build 150 and 1000 Bull Taurus-Indicus Run6). Only biallelic autosomal SNPs that are polymorphic within their respective populations are shown for Chillingham and Angus.

observed and expected homozygosities indicated higher levels of inbreeding in CIL (0.862 ±0.021) than AAN (-0.533 ±0.215). This was in-line with IBD measures, where the average for CIL was 0.889 (±0.025) compared to AAN 0.363 (±0.061; Table S5.5), indicating the samples in CIL and potentially the population as a whole are highly related.

5.4.3 Runs of Homozygosity

Angus and Chillingham significantly differed on all RoH-derived statistics (Table 4.2). Chillingham had approximately a fifth the number of RoH (791 ±49) that were observed in Angus (3,728 ±1,152), however, RoH in CIL were longer; the mean length and number of SNPs occurring within each RoH for CIL was 32-fold and 28-fold higher, respectively (Figure S5.2). Inbreeding coefficients (F_{RoH}) were consistent across Chillingham (0.91 ± 0.01) whilst also indicating that over 90% of the genome is included in RoH. Levels of autozygosity in Angus were more variable but overall lower than in Chillingham (0.13 ±0.12). All Chillingham individuals had at least 74 occurrences of RoH that exceeded 8 Mb (Figure S5.3). In Angus only 10 RoH segments larger than 8 Mb were detected, 6 of which occurred in AAN_1, with seven individuals lacking these extended regions.

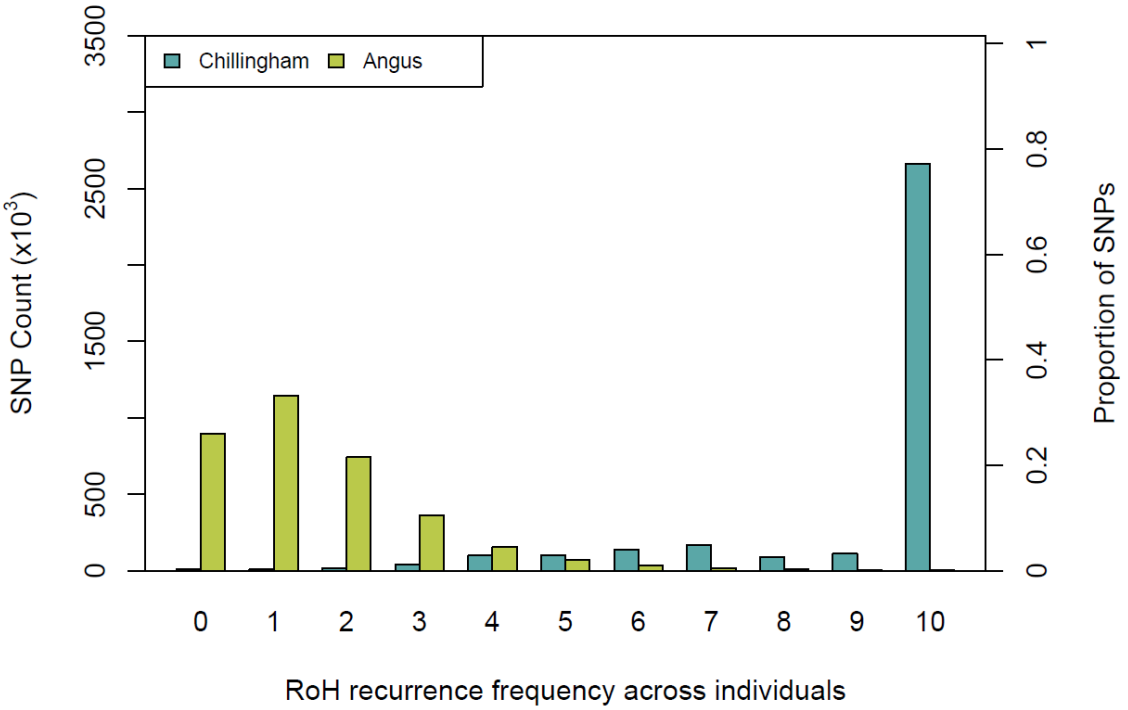


Figure 5.2. Runs of Homozygosty (RoH) recurrence within breed for Angus (n=10) and Chillingham (n=10) cattle. Each SNP was analysed for the recurrence of inclusion within a RoH for each given breed. Ranging from 0, the SNP was absent from any RoH, to 10, the SNP occurred in a RoH for all 10 individuals of the breed. The two y-axes are different representations of the same value - Proportional values were derived from the full dataset of 3,447,119 identified SNPs.

All 3,447,119 SNPs were assessed for the frequency of inclusion with a RoH independently for each breed. Most SNPs were included in a RoH for at least one individual in both breeds, totalling 99.7% and 73.9% within Chillingham and Angus, respectively (Figure 5.2). In Chillingham, 77.2% of SNPs occurred within a RoH across all ten individuals assessed, contrasting the 0.1% of SNPs that fulfilled the same criteria in Angus. RoH identified within Angus were often uniquely found in the given individual of the breed, with approximately a third (33.2%) of autosomal SNPs appearing in a RoH for only one tenth individuals tested.

Table 5.1. Runs of Homozygosity (RoH) derived statistics in Angus and Chillingham. One standard deviation is shown where mean values are reported. Emboldened values display breed means. Independent group t-tests were calculated between each breed for each statistic – all comparisons were significant.

	ID	Number of RoH	Length of RoH (Mb)	SNPs within RoH	Number of RoH >8 Mb	Fraction of RoH >8 Mb	F_{RoH}
Chillingham (n = 10)	PW_2	685	3.31 ±5.42	4582 ±7633	89	0.13	0.91
	PW_3	761	2.94 ±4.18	4074 ±5883	82	0.11	0.90
	PW_4	797	2.87 ±4.16	3986 ±5982	85	0.11	0.92
	PW_5	833	2.74 ±4.01	3803 ±5653	75	0.09	0.92
	PW_6	858	2.63 ±3.82	3643 ±5347	86	0.10	0.91
	PW_7	755	3.05 ±4.30	4228 ±6105	93	0.12	0.92
	PW_8	810	2.75 ±4.02	3820 ±5710	81	0.10	0.89
	PW_9	823	2.75 ±3.83	3790 ±5528	74	0.09	0.91
	PW_10	775	2.98 ±4.29	4146 ±5995	81	0.10	0.93
	PW_11	813	2.77 ±4.21	3822 ±6027	77	0.09	0.90
		791 ±49	**2.87 ±4.22**	**3974 ±5983**	**82 ±6**	**0.10 ±0.01**	**0.91 ± 0.01**
Angus (n = 10)	AAN_1	5636	0.14 ±0.52	205 ±731	6	1.06E-3	0.31
	AAN_2	2033	0.04 ±0.06	77 ±78	0	0	0.03
	AAN_3	2980	0.05 ±0.06	84 ±82	0	0	0.05
	AAN_4	3023	0.05 ±0.06	87 ±91	0	0	0.06
	AAN_5	3515	0.05 ±0.06	87 ±92	0	0	0.07
	AAN_6	3240	0.04 ±0.06	79 ±78	0	0	0.06
	AAN_7	3277	0.05 ±0.06	88 ±92	0	0	0.06
	AAN_8	3377	0.04 ±0.06	80 ±77	0	0	0.06
	AAN_9	5026	0.15 ±0.54	226 ±749	3	5.97E-4	0.31
	AAN_10	5172	0.15 ±0.50	218 ±660	1	1.93E-4	0.31
		3728 ±1152	**0.09 ±0.35**	**140 ±475**	**1 ±2**	**1.85E-4 ±3.61E-4**	**0.13 ±0.12**
t-test (t; df; *p*)		-8.05 9.0 < 0.001	58.51 7931.7 < 0.001	56.96 7930.1 < 0.001	40.09 10.9 < 0.001	25.04 9.0 < 0.001	20.04 9.1 < 0.001

5.4.4 HETEROZYGOSITY PEAK ANALYSIS

Heterozygosity per base pair was calculated for 248,434 windows spanning 100 kb each. A total of 1.7% windows were comprised of fewer than 10 SNPs and were subsequently removed (~4,102 ±1,588 windows removed per individual). The average fraction of genotypes that were heterozygous across all autosomal 100 kb windows was 0.039 (±0.145) across CIL and 0.472 (±0.272) across AAN (Table S5.3). All individuals had a similar number of windows exceeding the 5% upper null distribution (12,092 ±156), with no difference between breeds. Retaining outlying windows that appeared in at least 5 individuals of each breed resulted in 9,602 and 11,234 HPW for Chillingham and Angus, respectively (Figure 5.3). CIL_{HPW} were more clustered, with 94.7% overlapping at least one other HPW. Merging resulted in 511 discontinuous segments, which on average were made up of 18.8 (±20.8) original adjacencies with a mean extended window size of 296 kb (±230 kb). Fewer HPW (84.7%) were merged in AAN_{HPW}, resulting in 1,722 segments, combining an average of 6.5 (±7.0) windows with an extended length of 170 kb (± 90 kb). Considering the ARS-UCD1.2 reference genome approximates a mappable genome length of 2.5 Gb for chromosomal scaffolds, HPW spanned 6.1% (151,430 kb) and 11.7% (292,350 kb) of the overall genome for CIL and AAN, respectively. Within-population polymorphic SNP density was higher in CIL_{HPW} with an observed density of 0.643 SNPs/kb compared to overall genome density in CIL of 0.148 SNPs/kb. This trend is reversed in AAN_{HPW}, with 0.338 SNPs/kb within HPW and a genome density of 1.304 SNPs/kb – possible as HPW is a measure of heterozygosity of the SNPs that are present in the window, rather than the frequency of SNPs. A higher density of SNPs was observed in $AAN_{CIL\text{-}HPW}$ (0.992 SNPs/kb) than CIL_{HPW}. Observed heterozygosity across HPW was not significantly different from genome-wide measures for either AAN_{HPW} (t-test = -1.41; df =

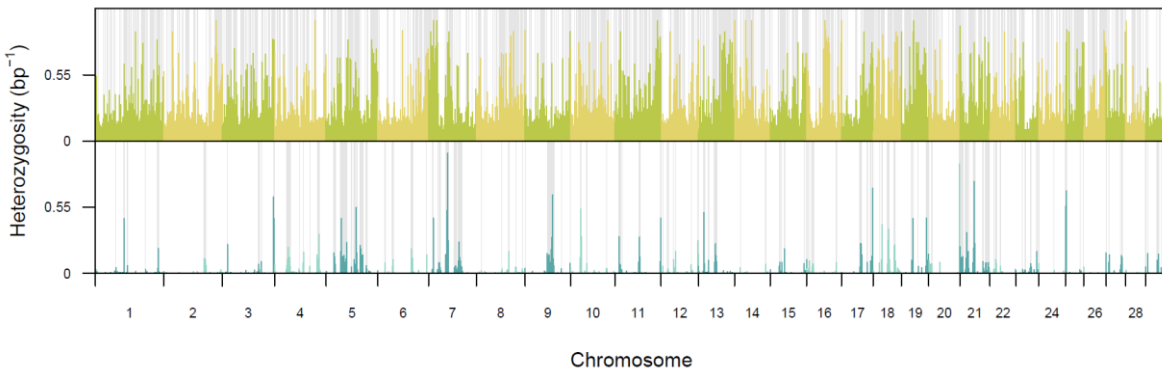


Figure 5.3. Heterozygosity peak analysis for Angus (n=10; top) and Chillingham (n=10; bottom) cattle. Each vertical, coloured line represents the median breed score of a 100 kb sliding window across the genome and the heterozygosity per base pair observed within. Windows with fewer than 10 SNPs were excluded. Grey regions show heterozygosity peak windows – identified as outlying windows in at least 5 individuals for each given breed.

17.6; *p*-value = 0.177) or $AAN_{CIL\text{-}HPW}$ (t-test = 1.43; df = 17.9; *p*-value = 0.171). Contrastingly, observed heterozygosity was approximately 8-fold higher in CIL_{HPW} compared to genome measures (t-test = -23.38; df = 9.5; *p*-value < 0.001; Table S5.2).

Overall, 9.6% of the 367,619 SNPs found in CIL and 7.8% of the ~3.2 million SNPs found in AAN showed a significant excess of heterozygotes. If balancing selection is a predominant driving force in maintaining heterozygosity within HPW there would be a larger proportion of sites with a significant excess of heterozygotes observed in these regions compared to the rest of the genome, thus the corresponding H score is expected to exceed 1. H scores for Angus were 0.890 (*p*-value = 1) and 0.891 (*p*-value = 1) for AAN_{HPW} and $AAN_{CIL\text{-}HPW}$, respectively, indicating a slight scarcity of sites with an excess of heterozygosity within these HPW relative to the genome as a whole. HPW in Chillingham were significantly enriched for SNPs with an observed heterozygosity excess (H = 2.654; *p*-value < 0.001). CIL_{HPW} contained 26.5% of the total number of SNPs, however, these regions contained 70.3% of all the SNPs displaying excess heterozygosity across the genome.

Annotation using VEP identified synonymous and non-synonymous polymorphic SNPs within HPW. In CIL_{HPW} half were found to be non-synonymous (364) and half were synonymous (363) out of a total of 727. Angus had lower proportions of non-synonymous variants, accounting for 42.9% (473 of 1103) in AAN_{HPW} (Fisher's exact test: *p*-value = 0.00294), but there was no significant difference in the proportion observed in $AAN_{CIL\text{-}HPW}$ (48%; 555 of 1155; Fisher's exact

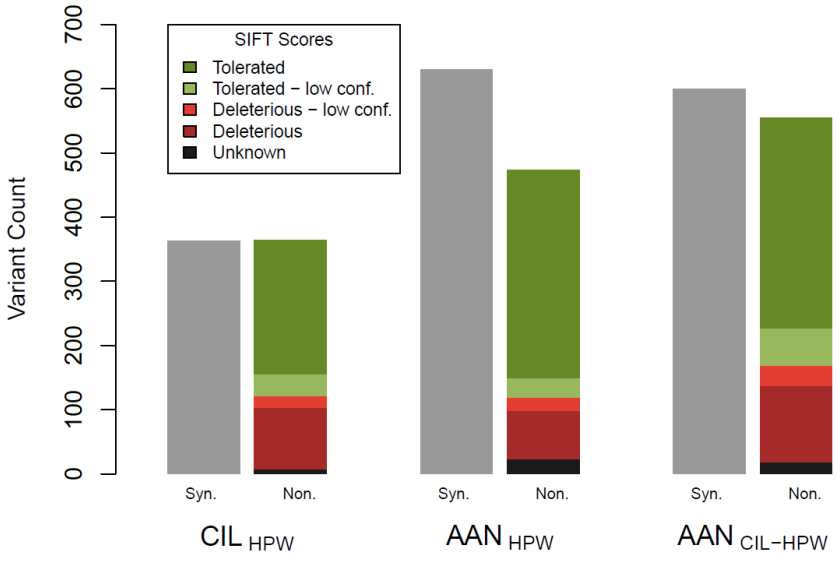


Figure 5.4. Variant effect prediction and SIFT tolerances for exonic variable sites within heterozygosity peak windows (HPW) for Chillingham (CIL) and Angus (AAN) *Bos taurus* breeds. SIFT scoring requires amino acid alterations so is only applicable to non-synonymous substitutions.

test: *p*-value = 0.395) (Figure 5.4). SIFT scores predicted deleterious effects with high confidence for 26.4%, 15.9% and 21.3% of non-synonymous variants for CIL_{HPW}, AAN_{HPW} and $AAN_{CIL\text{-}HPW}$, respectively (Figure 5.4).

Across all divisions of data for LD estimations, the minimum number of r^2 calculations contribution to a single bin was always in excess of 158,000. Median values for each bin indicated complete LD (r^2 = 1.00) spans 1,800 and 2,250 kb for CIL and CIL_{HPW}, respectively (Figure 5.5). The less conservative upper quartile boundary suggests full LD in Chillingham may extend as far as 4,300 kb for the genome or 5,500 kb across HPW (Figure S5.4). Weaker LD is observed in Angus, with a maximum median r^2 = 0.44, occurring in AAN_{HPW} between 0 - 50 kb (Figure 5.5). Full LD extended no further than 100 kb when considering the interquartile range (Figure S5.4). All Δr^2 distributions, calculated from the difference between genomic r^2 and HPW r^2 estimates, significantly differed from each other (*p*-value < 0.001; $D[CIL_{HPW}|AAN_{HPW}]$ = 0.47; $D[CIL_{HPW}|AAN_{CIL\text{-}HPW}]$ = 0.97; $D[AAN_{HPW}|AAN_{CIL\text{-}HPW}]$ = 1.00). A bimodal Δr^2 distribution was observed for CIL_{HPW} with a mode of 0.097 and a second, lower peak at 0.039. AAN_{HPW} and $AAN_{CIL\text{-}HPW}$ distributions had modes of 0.054 and -0.003, respectively (Figure S5.5).

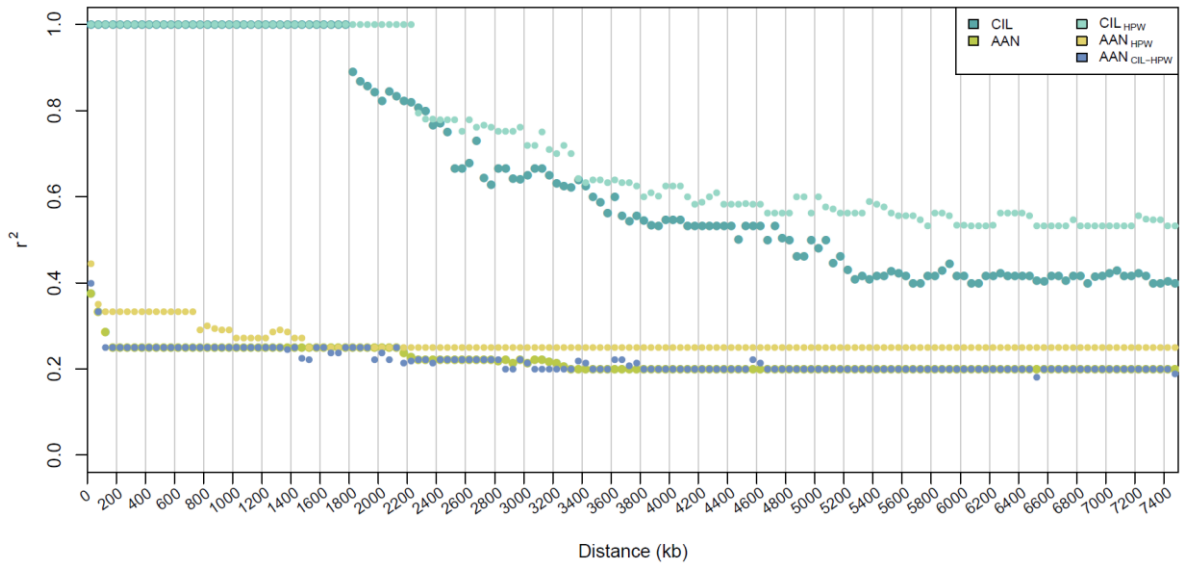


Figure 5.5. Decay of linkage disequilibrium measured as r^2 between pairwise SNPs as a function of physical distance. Grouped into bins spanning 50 kb. Median value displayed (full interquartile range plotted in Figure S5.4).

5.4.5 QUANTITATIVE TRAIT LOCI

A total of 8,465 QTLs (8.0% of database QTLs) were within CIL_{HPW}, resulting in significant enrichment of 13 traits (Table S5.4), with eleven of the traits consisting of at least 300 individual QTLs (Table 5.2). The trait "Texture", which describes the physical properties of meat, appeared

as a significantly enriched trait, however, only eight QTLs were associated with this trait across the database, six of which are within a completely linked tight window of 709 base pairs and the seventh within 100 kb of a single chromosome. This was a potentially erroneous result skewed by the heavy linkage. When this window is treated as a single locus, the "Texture" trait yields a non-significant result. Additionally, the terms "Inhibin level" and "Cheese fat recovery" were individually isolated to single chromosomes with QTLs spanning only 11.5 and 3.8 Mb, respectively (Figure S5.6). Thus, despite significance arising from a greater number of QTLs compared to the "Texture" trait, the QTLs within each region are likely to be tightly linked (Figure 5.5) and all three should cautiously interpreted. The remaining 10 significantly enriched traits were each comprised of QTLs within CIL_{HPW} across at least six chromosomes (Table 5.2). Overall, seven traits were associated with various properties of milk, with enrichment odds ratios ranging from 1.38 to 5.76. The remaining five traits were associated with fertility and were all within the top six most highly enriched terms, with odds ratios between 5.13 to 12.80 (Table 5.2).

Table 5.2. Quantitative trait loci showing significant enrichment in Chillingham's heterozygosity peak windows. Percent in CIL_{HPW} represents the percentage of QTL associated to a particular trait from the database that appeared within CIL_{HPW}. Significance detailed in Table S5.4.

		QTLs in trait		Percent	Enrichment	
	Trait	**Database**	**CIL_{HPW}**	**in CIL_{HPW}**	**odds ratio**	**nChr**
Meat	Texture*	8	7	87.5%	34.56	1
Fertility	Non-return rate	2350	1159	49.3%	12.80	6
Fertility	Interval from first to last insemination	424	220	51.9%	12.60	7
Fertility	Calving ease	3202	1386	43.3%	10.30	17
Fertility	Interval to first oestrus after calving	1031	457	44.3%	9.59	10
Milk	Cheese fat recovery*	36	13	36.1%	5.76	1
Fertility	Inhibin level*	297	92	31.0%	5.13	1
Milk	Butyric acid content	841	211	25.1%	3.90	7
Milk	Caproic acid content	610	97	15.9%	2.16	7
Milk	Fat percentage	6973	819	11.7%	1.59	12
Milk	Unglycosylated kappa-casein percentage	2514	294	11.7%	1.54	12
Milk	Glycosylated kappa-casein percentage	2719	295	10.8%	1.41	10
Milk	Kappa-casein percentage	4710	497	10.6%	1.38	16
	All traits (incl. non-significant; n = 390)	**105918**	**8465**	**8.0%**		

Asterisks denote traits with potentially erroneous significance owing to the high linkage between most QTLs and/or the low number of QTLs associated with the trait.

5.4.6 MAJOR HISTOCOMPATIBILITY COMPLEX

The MHC class I, IIa, IIb and III regions span a combined length of ~1.9 Mb containing 94 SNPs in CIL, resulting in an approximate mean density of 0.048 SNPs/kb, a 14-fold and 223-fold smaller than Chillingham's genome-wide and Angus' MHC region SNP densities,

respectively. The windowed scores derived from the heterozygosity peak analysis in CIL show low levels of variation across MHC classes with an average of 0.0326 (±0. 0763) $x10^{-3}$ and a maximum of 0.600 $x10^{-3}$. None of the MHC regions overlapped CIL_{HPW}. Despite overall low levels of variability, the majority of SNPs (71 of 94; 75.5%) are within the class I gene, *BOLA-NC1* (Chr23:28,548,781-28,553,089), or within 50 kb upstream and downstream of the genic region (Figure 5.6b). There was a significant excess of heterozygosity in 31 SNPs across the MHC, 22 of which were within the 50 kb of *BOLA-NC1*. The gene was not overlapped by any RoH (Figure 5.6) and was the only region of the MHC where no Chillingham individuals displayed a RoH (Figure S5.7).

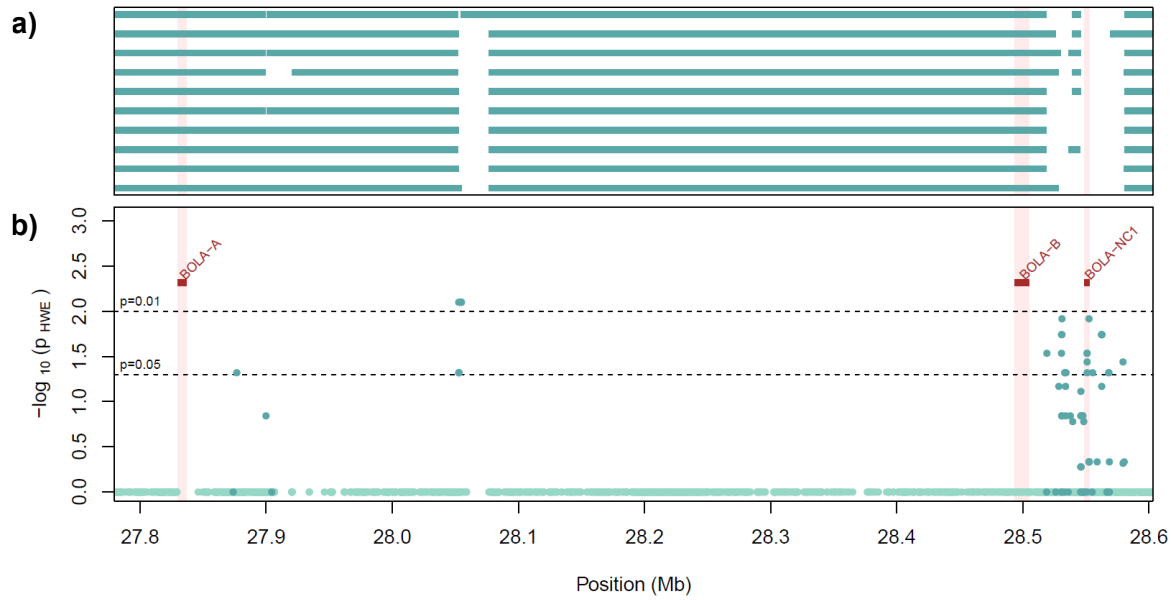


Figure 5.6. Class I bovine MHC region on chromosome 23. Red regions denote the span of class I genes described in Behl *et al.* (2012). a) Runs of homozygosity in Chillingham cattle genomes, each horizontal level represents a unique individual. b) Hardy-Weinberg equilibrium excess heterozygosity *p*-values for polymorphic SNPs within Chillingham (dark blue) and monomorphic SNPs within Chillingham that are polymorphic within Angus (light blue). MHC class IIa, class IIb and class III regions are depicted in Figure S5.7.

5.4.7 Demographic history

With MSMC2 modelling, demographic history for Chillingham and Angus was inferred up to approximately 84,000 and 504,000 years before present (YA), respectively (Figure 5.7). Overall declines in N_e were observed across both breeds, with a constant decline in Angus from prior to the second Pleistocene glacial period to the recent past (Figure 5.7a). A dramatic reduction in N_e was observed in Chillingham ~ 44,000 YA during the last glacial period, later increasing ~22,000 YA during the last glacial maximum (LGM) which occurred between 19,000 – 26,500 YA (Clark *et*

al., 2009). Decline in N_e is also observed in both breeds ~10,000 YA. Estimates of the contemporary N_e are 35.9 in Chillingham and 1479.0 in Angus, ranging between 9.7 – 39.8 and 233.3 – 5377.0 over the past 100 years for each breed, respectively. While Angus are typically analysed as regional purebred populations, this is broadly in line with other estimates, including American Angus with N_e estimates of 445 (Márquez *et al.*, 2010) and 654 (Saatchi *et al.*, 2011).

Divergence time between Chillingham and Angus was derived from MSMC2 outputs (Figure 5.7b). The breeds increasingly become separated from ~10,000 YA. Cross-coalescence

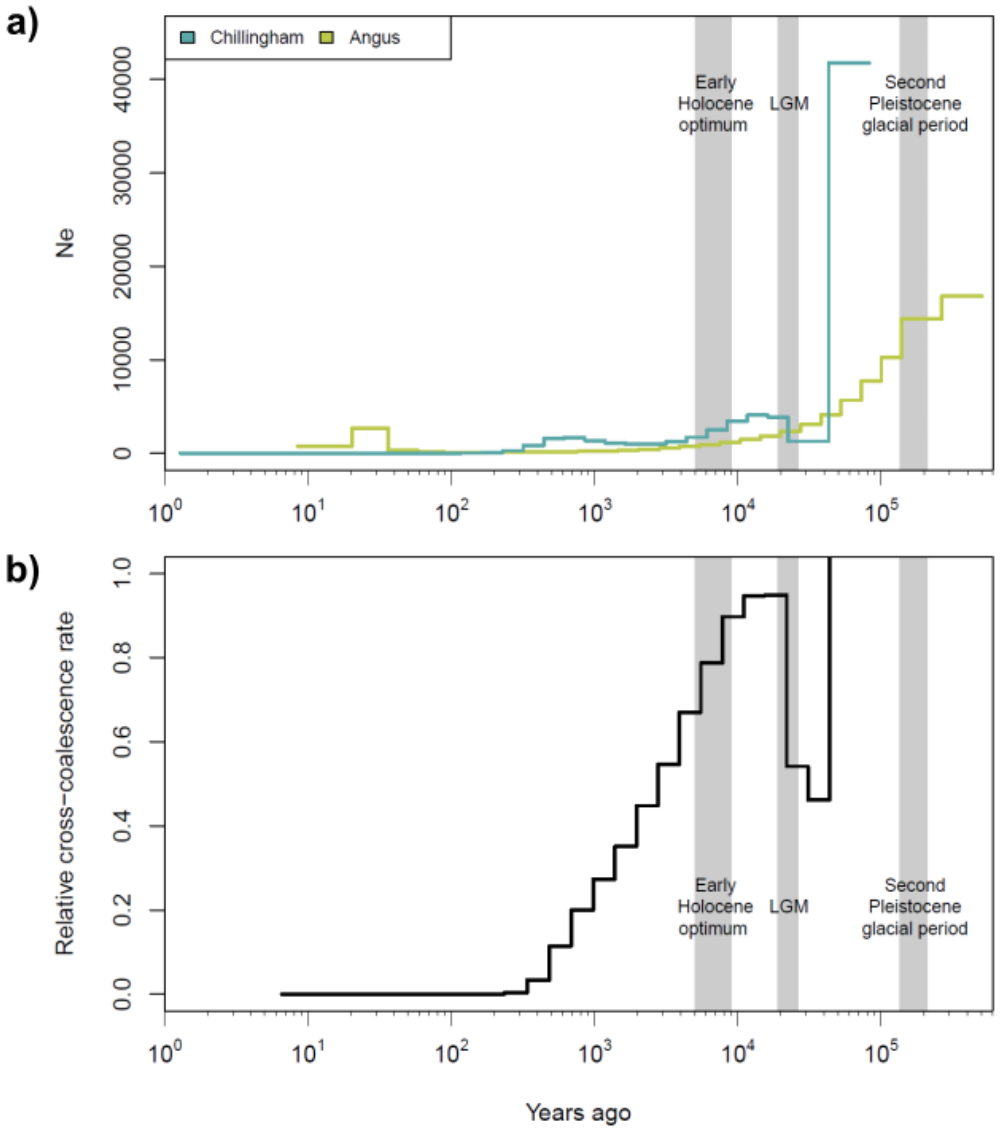


Figure 5.7. Coalescent-based estimations of demographic history of Chillingham and Angus cattle breeds using MSMC2. Ten resequenced genomes were used for each breed. Axes were scaled using a generation time of 5 years and a mutation rate of 1.25 x10^{-8}. x-axes of each plot are on a log10 scale. The grey shaded regions from left to right represent the early Holocene optimum, the Last Glacial Maximum (LGM) and the second Pleistocene glacial period. a) Effective population size (N_e) inferences calculated separately for each breed. b) Inferred relative cross-coalescence rate between Chillingham and Angus.

rates fell below 0.5 at ~2,800 YA suggesting a split between lineages alongside reduced gene flow was beginning at this time or within an estimated range of approximately 1,000 – 5,600 YA. The cross-coalescence rate declined from 0.034 to < 0.003 between 339 YA to 235 YA, remaining below 0.003 from 235 YA to present – suggesting the effective cessation of gene flow during that period and until the present. Interestingly, cross-coalescence rates decreased to 0.46 approximately 44,000 YA. This is followed by an increase to 0.95 during the LGM, this mirrors the pattern of N_e increase observed in Chillingham at this time.

5.5 DISCUSSION

Overall, Chillingham displayed high levels of inbreeding and limited genome-wide variation, with F_{IS} of 0.862 and a SNP count an order of magnitude lower than Angus. This largely is consistent with previous studies that noted Chillingham's unusually high homozygosity (Visscher *et al.*, 2001; Hudson *et al.*, 2012; Orozco-terWengel *et al.*, 2015; Williams *et al.*, 2016). An important distinction is that high density SNP array data indicated a higher inbreeding coefficient of 0.924 (Orozco-terWengel *et al.*, 2015; Williams *et al.*, 2016). This slight disparity likely encapsulates the effect of ascertainment bias in the array – particularly noticeable in Chillingham as the breed has been subject to isolation and genetic drift for over 300 years. Far fewer variants were identified in this study (0.475 and 0.051 heterozygous sites/kb for AAN and CIL, respectively) compared to the first phase of the 1000 Bull Genomes Project, where an average of 1.44 heterozygous sites/kb are identified across 234 individuals (Daetwyler *et al.*, 2014). This is expected in Chillingham due to the observed lack of variation, however, it could be argued that Angus should be more reflective of the extensive dataset considering the presence of the Angus breed in the previous study. This discrepancy is probably largely explained by the increased sample size in the previous study and the inclusion of 5 different breeds, however, the more conservative filtering implemented in the later stages of variant calling in the present study would have reduced the number of variable sites identified.

The method to detect autozygosity estimated an inbreeding coefficient closer to previous studies for CIL (F_{RoH} = 0.91), reflective of the extensive coverage of RoH across the genome. The distribution of RoH was skewed towards longer RoH in CIL, with an average length of 2.87 Mb, 32-fold higher than AAN (Figure S5.3). Long RoH arise from more recent inbreeding events as recombination has not had sufficient time to shuffle autozygous regions (Browning and Browning, 2013). Furthermore, overall rates of recombination within a population increase with N_e allowing the shuffling of smaller autozygous tracts. The reverse implying high levels of recent inbreeding

and with the low N_e observed in Chillingham has resulted in extended stretches of population-wide autozygosity. This is reflected in the complete linkage of loci 1,800 kb apart and the decay to $r^2 = 0.5$ after 5,000 kb, values that are 350 kb further and linkage that is 0.25 higher than previous estimations, respectively (Williams *et al.*, 2016). Linkage disequilibrium declined to $r^2 = 0.25$ within 150 kb, a more typical rate observed across cattle breeds (Orozco-terWengel *et al.*, 2015). The high linkage observed in Chillingham was anticipated to be artefactually inflated by the genome-wide scarcity of variation that may allow recombination events to go undetected, however, LD within CIL_{HPW} – exclusively regions with enriched variability – indicated median r^2 values of one at even greater distances than genome-wide estimates (Figure 5.5). Complete linkage was frequently observed between regions of distinct HPW (e.g. 1:67.0 – 74.3 Mb; 4:28.8 – 36.6 Mb; 7:21.3 – 24.1 Mb) despite being separated by RoH and over 2 Mb distance. This provides further validation to LD and F_{RoH} estimates and indicates that selective forces may even be acting across multiple HPW simultaneously (Figure 5.5; Figure S5.5).

Research on Scandinavian wolves also note the presence of similarly extended RoH with some runs spanning the length of whole chromosomes, however, despite a founding population of 2 individuals, high inbreeding and semi-isolation F_{RoH} did not exceed 0.54 due to the relatively short bottleneck duration of 6-7 generations (Kardos *et al.*, 2018). Interestingly, Chillingham displayed high fidelity RoH, with 77.2% of SNPs occurring in RoH for all ten individuals. Conservative population-wide filtering in the heterozygosity peak analysis largely recapitulated these results – albeit with a focus on maintenance of variation rather than a lack of – indicating recurring signals of variation across individuals. The higher rate of overlap in windows forming CIL_{HPW} compared to AAN_{HPW} suggests variation across the genome is more spatially localised and occurs within specific regions, in agreement with the autocorrelation analysis carried out by Williams *et al.*, (2016). Additionally, as SNP arrays are often developed to identify roughly equidistant markers across the genome, the clustered distribution of variation in Chillingham alongside the identification of novel variants contributes to the aforementioned ascertainment bias of the SNP arrays.

The maintenance of variation in specific clusters may be facilitated by balancing selection. The H scores indicated a disproportionate abundance of loci with excess heterozygosity in CIL_{HPW} comparatively to the rest of the genome. Because the distribution of Chillingham's HPW is not likely to represent the null distribution of HPWs in the breed, such windows were also estimated in Angus. One analysis represented the null distribution of HPW along the genome in a less inbred traditional cattle British breed (AAN_{HPW}), while the other analysis looked at whether the genomic regions of the Chillingham HPWs presented any unexpected patterns of variation in the less

inbred traditional cattle British breed ($AAN_{CIL\text{-}HPW}$). In both instances, there was a scarcity of excess heterozygosity in these regions with H scores below 1, despite representing methodological- and genomic-replicated regions of CIL_{HPW}, respectively. Additionally, CIL_{HPW} represented the regions with the highest proportion of non-synonymous mutations among exonic variable sites (Figure 5.4; Figure S5.8). The San Nicolas Channel Island foxes provide a comparable population, where an N_e of 64 has persisted over 500 generations, including a strong bottleneck 30 generations ago. Similarly, to Chillingham, the foxes show a 'saw-toothed' pattern of variability across the genome, with long RoH interspersed with regions of high heterozygosity (i.e., Figure 5.3). While this is attributed fully to the demographic history in the foxes (i.e. long term low population size but ancestral variation not yet removed by drift), they do not show evidence of significant excess heterozygosity within HPW, indicating Chillingham have experienced balancing selection alongside demographic effects to maintain the high heterozygosity of HPW (Robinson *et al.*, 2016). CIL_{HPW} potentially contain an abundance of alleles conferring a heterozygous advantage or deleterious recessive effect. Interestingly, SIFT predictions show elevated proportions of deleterious variants of the non-synonymous sites across CIL_{HPW} compared to AAN_{HPW}, providing insight into the evolutionary costs of regions associated with strong balancing selection; these could arise through hitchhiking of moderately deleterious alleles with selected loci, or indicate purging selection is ineffective or less beneficial in HPW (Lenz *et al.*, 2016). Furthermore, $AAN_{CIL\text{-}HPW}$ had a similar synonymous to non-synonymous ratio to CIL_{HPW} while also retaining a higher proportion of deleterious variants than AAN_{HPW}, suggesting that the regions identified as CIL_{HPW} may be functionally important across other breeds.

5.5.1 QUANTITATIVE TRAIT LOCI

CIL_{HPW} therefore provided regions of interest, to investigate potential causative genes, QTLs and processes associated with balancing selection. It is important to note that QTLs in cattle are often investigated with a focus on commercial breeds and traits, therefore, there is a selection bias towards meat, carcass, and milk traits (Ma *et al.*, 2019). The top four QTL terms enriched within heterozygosity peak windows in Chillingham are related to fertility, while notably, no immunity traits were identified as significant enriched throughout this analysis. Due to the typically additive nature of many QTLs and the strong artificial selection modern commercial breeds are subjected to, significant enrichment of traits was not expected in an unmanaged feral breed such as Chillingham.

Non-return rate and interval from first to last insemination are closely linked traits – the former is defined as the proportion of cows that must be inseminated a second time within a particular number of days after an the initial insemination took place (often ranging between 28

to 90 days) and the latter refers to the time between the first attempted and the final successful inseminations (Madrid-Bury *et al.*, 2005). These traits are reliant on both the conception rate immediately following insemination and gestation post-conception, which in turn are comprised of further complexity (e.g. ova availability, parental contribution to developmental potential of the conceptus, spermatozoa characteristics and quantity; den Daas, 1992), however, non-return rate and interval from first to last insemination are typically used to evaluate male fertility (Koops *et al.*, 1995). These QTLs may represent the site of genetic compensatory mechanisms to counteract the apparent subfertility observed in Chillingham bull semen (Hall and Bunce, 2019). Furthermore, the lack of a human-mediated mating system in the Chillingham herd has resulted in highly contested mating of cows in oestrous (Hall, 1989). Thus, male fertility and conception at first attempt is likely paramount to an individual bull's reproductive success.

Interval to first oestrus after calving was significantly enriched in CIL_{HPW} which could be reflective of the aseasonality of mating within the breed, Chillingham cows have been observed conceiving as early as 40 days following calving (Hall and Hall, 1988). Seasonality is usually considered beneficial for offspring survival as calving occurs when food is plentiful, however, Chillingham are unusual for ungulates in that they reproduce all year round (Hall, 1989; Hall, 2006). The viability of this strategy is possibly facilitated by the historically maintained low population size of the herd, an estimated carrying capacity of 120 individuals in the 1.34 km^2 protected estate and the supplementary winter feed proving an abundance of available nutrients all year round (Visscher *et al.*, 2001; Hall and Bunce, 2019). The reverse was observed in the now-eradicated feral cattle on Amsterdam Island (France, southern Indian Ocean), where seasonal rutting occurred between January and March, followed by extended post-partum anoestrus hypothesised to be causatively related to their poor nutrition (Berteaux and Micol, 1992). Furthermore, seasonal mating results in a dominant male receiving a large proportion of female mates (Hall and Hall, 1988; Hall, 1989; Berteaux and Micol, 1992), whereas the aseasonality such as that observed in Chillingham allows more individual bulls mating opportunities, potentially mitigating the effect of increased inbreeding and reduced N_e.

Perhaps the most important significantly enriched QTL trait to herd fertility is calving ease. Despite lack of human intervention and elevated herd inbreeding only 1.5% of calvings fail due to the death of the cow (Hall, 1989), with fertility and viability absent of diminishment (Hall and Hall, 1988; Visscher *et al.*, 2001). Overall, these results indicate that fertility is one of the primary driving forces in maintaining variation in the few remaining clusters of heterozygosity across the Chillingham genome.

The identification of enriched QTL traits may provide useful preliminary information for further investigation, however, direction functional inferences in diverged populations may lose power as population specific effects of pleiotropy, epistasis and gene expression may modulate the relative phenotypic effect of each given QTL (McKay and Latta, 2002). Predicting the relative accumulation of deleterious alleles with the Chillingham genome and CIL_{HPW} indicates the potential antagonism between balancing and purging selection and provides a basis for future research to investigate the specific deleterious alleles that are maintained or fixed within Chillingham. Similarly to QTLs, the effects of these variants are inferred by trends observed across wider taxa, but may vary within Chillingham (Sim *et al.*, 2012). The assumption that deviations from a more commonly observed variant implies a deleterious effect (the SIFT methodology relies on such assumptions) may be erroneous when applied to a population under unique environmental conditions compared to conspecifics. Furthermore, epigenetic modifications such as methylation can regulate expression and provide a mechanism for rapid adaptation that would not be detectable with the current analysis (Sevane *et al.*, 2019). Transcriptomics would contextualise the functional affect of the accumulation of putatively deleterious mutations and determine if mechanisms such as functional rescue or duplication-divergence events are mitigating the overall negative effect of these variants.

5.5.2 MAJOR HISTOCOMPATIBILITY COMPLEX

Despite no overlap with CIL_{HPW}, it remains important to investigate the MHC for three primary factors concerning these regions: widespread evidence of balancing selection (e.g. Giovambattista *et al.*, 2013; Lenz *et al.*, 2016; Kloch *et al.*, 2018), the contradictory suggestion of homozygosity within these regions for Chillingham (Visscher *et al.*, 2001; Williams *et al.*, 2016), and the wider importance of immunity in the survival of inbred populations (Acevedo-Whitehouse and Cunningham, 2006; Ellis and Hammond, 2014).

The whole-genome data presented are primarily in-line with previous observations suggesting limited variability across MHC regions in Chillingham (Visscher *et al.*, 2001; Ellis and Hammond, 2014). Heterozygosity and within-population SNP density far below genome-wide averages for the breed, with complete homozygosity across most MHC genes and the entire class IIb region. A similar ruminant system is the feral population of Soay sheep (*Ovis aries*) on the island of St. Kilda in Scotland, however, unlike Chillingham, the sheep exhibit long-term balancing selection on class II MHC genes with haplotypes impacting male fertility and female lifespan (Charbonnel and Pemberton, 2005; Huang *et al.*, 2022). The main force maintaining this MHC

diversity is thought to be driven by the nematode parasite *Teladorsagia circumcincta* (Charbonnel and Pemberton, 2005), to which Chillingham seemingly lack an analogue (Visscher *et al.*, 2001; Hall and Bunce, 2019). Contrastingly, subpopulations of species such as the Eurasian beaver (*Castor fiber*), great crested newt (*Triturus cristatus*) and moose (*Alces alces*) have remained viable for hundreds to thousands of years despite a complete lack or considerable reduction in variation across the MHC (Mikko and Andersson, 1995; Babik *et al.*, 2005; Babik *et al.*, 2009; Radwan *et al.*, 2010). While Soay sheep may not mirror Chillingham cattle at the MHC loci, the disassortative mating observed in the sheep population may provide a behavioural mechanism to the maintenance of HPW observed elsewhere in the Chillingham genome (Huang *et al.*, 2022) acting to maximise diversity and minimise inbreeding despite high levels of drift.

Despite the near complete monomorphism across MHC in Chillingham, the variation that was observed was not uniformly distributed, but was instead clustered around the class I non-classical MHC gene, *BOLA-NC1*. Typically, class I classical MHC molecules present antigens to $CD8^+$ T cells, whereas class I non-classical MHC molecules activate or inhibit stimuli in natural killer (NK) cells (Halenius *et al.*, 2015). Contrary to observations in Chillingham, non-classical MHC genes often have limited polymorphism relative to the variation observed in classical MHC genes (Birch *et al.*, 2008; Shu *et al.*, 2014). In cattle, *BOLA-NC1* is prone to alternative splicing at exon 5, which encodes a transmembrane domain; the inclusion or exclusion of exon 5 forms a membrane-bound or soluble protein, respectively (Davies *et al.*, 2006). The membrane-bound isoform is suspected to play a role in pathogen subversion of the immune response whilst the soluble isoform is an immunosuppressive factor which induces apoptosis in activated $CD8^+$ T cells (Davies *et al.*, 2006; Birch *et al.*, 2008). Spatio-temporal studies of expression have also implicated *BOLA-NC1* in maternal immunity, due to upregulation during the first and third trimester in pregnant cows (Shu *et al.*, 2014). It is possible that the variability observed in this gene is essential to fertility and immunity, maintaining multi-functionality through strong balancing selection.

Interestingly, Chillingham exhibit population-wide homozygosity for a transitional mutation at Chr23:28,549,793 resulting in premature stop codon in two out of the three transcripts encoded by *BOLA-NC1* gene. This occurs in exon 5, potentially disrupting the production of the membrane-bound isoform (Davies *et al.*, 2006). It has been shown in MHC regions of humans that transcript expression is heavily modulated by allele-specific variants occurring in upstream and downstream elements, such as promotors, enhancers and loci altering epigenetic factor regulation (Gensterblum-Miller *et al.*, 2018). The increased variation observed in SNPs adjacent to the genic region may therefore be involved in regulating expression of multiple transcripts, rather than through the previously identified mechanism of alternative splicing (Davies *et al.*, 2006).

Multiple MHC genes exhibit overdominance, as heterozygotes can present a greater diversity of peptide antigens to T lymphocytes , thus initiating the recognition of a greater variety of foreign antigens through the adaptive immune response (Hedrick, 1994). This can lead to balancing selection and select for heterozygous individuals through mechanisms such as MHC-disassortative mating (Huang *et al.*, 2022). The lack of heterozygosity within Chillingham MHC regions without apparent deterioration of immune functionality could be explained gene duplication and divergence of exons encoding peptide binding sites, thus maintaining both protein level variability and region-wide homozygosity (Dearborn *et al.*, 2016).

A potential concern with identifying variants is the high copy number variation of genes within the MHC region and thus the alignment of adjacent copies to the same location on the reference genome resulting in potentially spurious heterozygosity calls. This effect was minimised with the various conservative filtering steps implemented prior to variant calling, including alignment with GATK HaplotypeCaller, base quality score recalibration and limiting maximum read depth. The SNP distribution in Angus indicates that this method successfully and relatively uniformly identified SNPs across MHC regions, therefore suggesting that the variation observed in Chillingham is not artefactual. One exception lies within the class IIa region (Chr23:25,579,666-25,658,473), where no variable loci passed filtering in a 79 kb window for either breed.

5.5.3 Demographic history

The trends observed from the coalescent-based estimations of N_e (Figure 5.7a) highlight the importance of caution when inferring results directly from single MSMC2 analyses. If taken in isolation, it seems apparent the ancestral taurine population were exposed to first a bottleneck and secondly population expansion occurring approximately 44,000 and 22,000 YA, respectively. Supplementing the N_e estimations with cross-coalescence rates calculated using additional *B. taurus* data allows inferences to be made whilst considering population connectivity, admixture and introgression. It has been shown that deviations from panmixia and the introduction of population structure can result in spurious signals of population size changes in such coalescent-based estimations (Orozco-terWengel, 2016; Mazet *et al.*, 2016). Biases can therefore arise towards the interpretation of N_e increases within populations experiencing hybridisation and introgression regardless of the presence of any actual growth in N_e (Orozco-terWengel and Bruford, 2014; Bosse *et al.*, 2014). While this effect is reduced with cross-coalescent models such as MSMC2, the relative cross-coalescent rates to estimate population separation can disguise both post-split and archaic admixture between populations (Wang *et al.*, 2020). Furthermore, no attempt was made to include only neutral sites in this demographic analysis, consequently, selective or other demographic processes may influence the N_e inferences – for example, an

introgression event may erroneously be detected as a population expansion as heterozygosity increases and longer autozygous tracks are disrupted, or positive select and the associated regional loss of variation may be interpreted as a bottleneck. While the use of neutral sites is preferable, neutral allele frequencies can also be influenced through processes such as background selection, therefore the ideal sites for demographic inferences are likely neutral sites in regions of high recombination when an accurate recombination map is available (Marchi *et al.*, 2021).

The comparison of demography sourced from the two extant populations of Chillingham and Angus imply drastically different trends beyond ~20,000 YA (Figure 5.7). As divergence and breed formation is far more contemporary there is an expectation of a shared demographic history at such an ancient time point. The disparity observed may be due to the extensive homozygosity present in the Chillingham samples, as similar spurious signals emerge in MSMC2 analyses on other small, inbred populations (e.g. Robinson *et al.*, 2021; Zhang *et al.*, 2021).

Both demographic trends show a marked decline in N_e from ~10,000 YA, this is associated with the domestication of *B. taurus* cattle by Neolithic humans which led to the gradual reduction in gene flow and increased inbreeding (Bruford *et al.*, 2003; Ajmone-Marsan *et al.*, 2010; MacHugh *et al.*, 2017). The cross-coalescence rate estimates drops below 0.5 indicating the potential development of population structure within the ancestral population 2,800 YA (range of 1,000 – 5,600 YA; Figure 5.7b). Regarding taurine cattle this would be an expected post-domestication range for divergence between two European breeds with examples of divergence between European *B. taurus* and taurine in the domestication centre and Africa occurring approximately 6,000 – 8,000 YA (Chen *et al.*, 2018). The relatively high cross-coalescent rates during the 13th century (0.200 – 0.273) and maintained at slightly lower levels for 400 years, refutes the earliest hypothesis that Chillingham have been completely isolated for 700 years (Hall and Hall, 1988). Nonetheless, observations suggest that gene flow between the two breeds finally ceased between 235 – 339 YA (Figure 5.7b), capturing the known history of Chillingham to have been an isolated and inbred breed for at least 300 years (Hudson *et al.*, 2012; Williams *et al.*, 2016).

Purging selection is more efficacious during extended, consistently small populations rather than sudden, short-lived bottlenecks (Ehiobu *et al.*, 1989; Day *et al.*, 2003; Robinson *et al.*, 2016). The maintenance of a low effective population size for the past ~60 generations, despite the severe bottleneck in the 1940s, supports the possibility of purging selection allowing for the long-term persistence of the herd. Furthermore, the quantitative variation (here measured through enrichment of QTLs) observed in Chillingham may underrepresent the total additive variation of the associated traits. While additive genetic variation is removed by a population bottleneck, the

same demographic event can convert non-additive (dominance and/or epistatic) gene interactions to additive variance through increased frequency of recessive alleles, thus potentially allowing for maintenance or even improving fitness (Dlugosch and Parker, 2008; Heerwaarden *et al.*, 2008).

5.5.4 CONCLUSION

It is important to contextualise the results presented in this study, one such limitation is the comparative theme with Angus. Both breeds are genetically distinct (Orozco-terWengel *et al.*, 2015) and present unique and novel variation (Figure 5.1), with Angus representing one archetypical British breed, however, the Angus samples were bulls obtained from a commercial setting from lineages exposed to strong artificial selection (Daetwyler *et al.*, 2014). Angus was chosen as a reference due to the breed's middling genetic variation among taurine breeds, the close phylogenetic relationship with Chillingham (Orozco-terWengel *et al.*, 2015), but also because of the readily available high coverage short read sequence data available (Hayes and Daetwyler, 2019). Angus provided a baseline of estimated genome-wide variation, while the conservative approach to filtering data both prior to variant calling and during latter analyses allow more reliable intra-population conclusions regarding Chillingham. Nonetheless, the analyses carried out and conclusions drawn from this study rely heavily on appropriate reference individuals; using ten individuals from a single, commercial breed under strong artificial selection may not be ideal but was unfortunately necessary due to limited resources in obtaining new resequencing data and computational processing time. The introduction of additional, non-commercial breeds would allow the disentangling of biases arising from artificial selection. Specifically, two British ancient native breeds closely related to Chillingham, namely Welsh White Park or Vaynol breeds, may provide biological replicates of the current study.

The conservation of native and ancient breeds is important to maintain biodiversity, heritage, and to provide genetic resources representative of regional adaptation in livestock (Taberlet *et al.*, 2011; FAO, 1992; Hoffmann *et al.*, 2017). Chillingham provides an excellent example of the viability of conserving a locally adapted breed in a semi-wild habitat despite severe bottlenecks and minimal management. Furthermore, although the Chillingham estate is not a rewilding project, ability of the herd to fill niche of extinct megaherbivores such as aurochs (*Bos primigenious*) provides potential ecological function and benefit to the ecosystem (Hall and Bunce, 2019).

Highly variable regions in the Chillingham genome that were putatively under balancing selection are associated primarily with genes affecting fertility rather than immunity. The lack of

immunity-related QTLs enriched within HPW as well as the almost ubiquitously homozygous MHC regions provide further questions into the historical pathogenic burden of the breed and the adaptive potential to respond to emerging diseases. Future work could focus on identifying copy number variation as a source of intra population variation, additionally, investigating epigenetic markers and the effect on short term expression profiles could be of interest. Examining other systems provide a valuable comparison for conserved strategies of small, inbred populations such as Chillingham – this can be approached from multiple aspects including closely related breeds (e.g. Vaynol and Welsh White Park) (Orozco-terWengel *et al.*, 2015), feral breeds in analogous systems (e.g. Amsterdam Island cattle) (Berteaux and Micol, 1992), and due to the highly-conserved nature of MHC and other immunity-related genes, other mammalian systems with low effective population size and high homozygosity (e.g. Soay sheep, San Nicolas Channel Island foxes and Scandinavian wolves) (Robinson *et al.*, 2016; Kardos *et al.*, 2018; Huang *et al.*, 2020).

5.6 Acknowledgements

A special thanks to the Stephen Hall and the Chillingham Wild Cattle Association for providing the blood samples for resequencing, useful discussions, and the excellent reading resources on Chillingham cattle.

Chapter Six

General Discussion

6.1 BACKGROUND

It is estimated that 1,458 livestock breeds (17%) are currently at risk of extinction, with insufficient data on a further 58% of them (FAO, 2015). Cattle and sheep have seen the greatest number of breed extinction events, 184 (of 1,408; 13%) and 160 (of 1,542; 10%), respectively; only 20% of cattle and 26% of sheep breeds are currently classified as "not at risk" of extinction (FAO, 2015). As demand to feed the growing human population increases, so too will the productivity requirements of the livestock sector (Thornton, 2010). The area covered by farmland and pastures has not increased since 1991 (O'Mara, 2012), thus intensification of current practices will be required to meet demand.

Artificial selection is a fundamental process in improving herd productivity as individuals with increasingly productive phenotypes are selectively bred each generation. The efficacy in controlling mate pairing substantially improved with the development of artificial insemination; however, this is often to the detriment of genetic diversity (Taberlet *et al.*, 2011). Holstein cattle, for example, have a contemporary global effective population size of 50, with all commercial bulls (available in the United States of America) descended from two ancestral individuals born in 1880 (Yue *et al.*, 2015). While Holstein is perhaps an extreme example, not only are rare and native breeds experiencing greater extinction risks due to marginalisation from more productive breeds (Villalobos Cortés *et al.*, 2009), efforts to increase the financial viability of autochthonous breeds can encourage indiscriminate crossbreeding with more productive breeds, resulting in loss of genetic variation (FAO, 2015). Local adaptation may sometimes be viewed as secondary to production, however, resilience to disease and harsh environmental conditions can reduce expenditure (antibiotics, feed, water and climate-controlled housing), while also maintaining sufficient productivity (Flori *et al.*, 2012; Strandén *et al.*, 2019; He *et al.*, 2020).

Anthropogenic emissions of greenhouse gases are accelerating substantial shifts in global climate and ecology (FAO *et al.*, 2018). Changes in the variability and magnitude of temperature, humidity, water supply, extreme weather events, and disease burden are a few examples of both direct and indirect threats to the food supply chain and wider agricultural sector (Rojas-Downing *et al.*, 2017; Godde *et al.*, 2021). Local livestock breeds are a reservoir of genetic variation and locally adaptive traits; characterising the underlying genetic basis to local adaptation, climate resilience, and perseverance of small populations is therefore invaluable as environmental shifts hinder the productivity of our most populous and relied upon livestock breeds (Ajmone-Marsan *et al.*, 2010; Porto-Neto *et al.*, 2014).

6.2 COMPLETION OF AIMS

In **chapter two**, SNP array data was analysed for 3,196 individuals (*Bos taurus*, *Bos indicus* and hybrids) across 180 breeds extracted from a total of ten published studies, making this one of the most inclusive collations of cattle to date. The aims included characterising genetic diversity, population structure, and demographic history of extant cattle, with particular interest on testing alternative hypotheses about the domestication events in the two species. While this is directly informative of the cattle population and extant variation, due to animal husbandry, strong associations form between the demographic processes of livestock and development of our own species. Thus, by addressing the timing of domestication within taurine cattle (*Bos taurus*), inferences can also be made about the spread of pastoralism and Neolithic culture within humans.

Across most breeds, observed heterozygosities were significantly lower than expected values, a mostly predictable result due to the increased fragmentation of cattle into distinct and isolated populations since breed formation (Taberlet *et al.*, 2011). Additionally, taurine had significantly higher heterozygosity than indicine breeds and while initially thought to be due to the more recent domestication and expansion of indicine, recent studies have highlighted the significant ascertainment bias of the SNP arrays used for the analyses (Utsunomiya *et al.*, 2019). The identification of 'high frequency' SNPs was determined almost exclusively from sequence diversity in European taurine (Angus, Holstein, Jersey, Limousin, and Norwegian Red) with only a single indicine breed (Brahman; The Bovine HapMap Consortium *et al.*, 2009). Divergence, genetic drift, and differential selection can result in significant underestimations of genetic diversity in those breeds not within the ascertainment panel, even within conspecifics (i.e. African taurine; McTavish and Hillis, 2015). Oversampling of highly polymorphic loci within populations can result in lower F_{ST} between populations, underestimating their relative divergence; contrastingly, a subpopulation biased sampling can inflate F_{ST} estimates (Albrechtsen *et al.*, 2010). Multiple steps were implemented to tackle the recurring challenge of ascertainment bias throughout this thesis, including pruning markers based on linkage disequilibrium (Malomane *et al.*, 2018), using biological replicates (Rougemont *et al.*, 2016), using haplotype-based methods (Sabeti *et al.*, 2007), focussing on more closely related breeds (**chapters three, four and five**; McTavish and Hillis, 2015), using less bias SNP arrays (**chapter four**; Kijas *et al.*, 2012), and whole-genome resequencing data (**chapter five**; Albrechtsen *et al.*, 2010). Nonetheless, it is evident that ascertainment bias effected many estimates throughout this project; therefore, inferences made – particularly of genetic diversity between distantly related breeds – should be taken cautiously.

The greatest reduction in ascertainment bias and finest-scale investigation into breed-specific variation was within, **chapter five**, which introduces Chillingham cattle, a feral breed of *Bos taurus* that has been isolated from other cattle breeds for at least 300 years. The herd provides an interesting study of a small, inbred, and isolated mammalian population that despite multiple demographic bottlenecks remains viable. Determining the distribution and magnitude of genetic variation across the Chillingham genome was of particular interest, as well as the selective mechanisms maintaining variation and the phenotypic consequences. This contrasts previous chapters, in that the focus is more on a single isolated population, with little substantial risk or influence from future or past climate change, respectively. Instead, as marginalisation of rare and native breeds continue, Chillingham provides a model for the management of endangered populations and the potential role as ecosystem engineers or even for rewilding projects.

Chillingham had previously been investigated with microsatellite markers, mitochondrial DNA, and SNP array data all of which have highlighted the apparent lack of genetic variation of the breed (Visscher *et al.*, 2001; Hudson *et al.*, 2012; Williams *et al.*, 2016). While SNP arrays allow for affordable and consistent standards between studies, evident by the vast datasets concatenated for **chapter two** and **chapter four**, ascertainment bias is an issue in most marker sets, particularly when the population in question is distantly related to the ascertainment panel. This is initially evident in Chillingham as inbreeding coefficients from SNP arrays (F_{IS} = 0.924; Williams *et al.*, 2016) seem to overestimate inbreeding compared with results from **chapter five** (F_{IS} = 0.862), and was a key limitation into investigating indicine influence in **chapter two** and **chapter three**. Whole-genome resequencing (WGRS) can alleviate some issues presented by ascertainment bias in a population that has been isolated for multiple generations by identifying the unique variation that would be missed from analysing only previously defined 'variable' markers (Albrechtsen *et al.*, 2010). In total, 12.5% of the SNPs identified in the WGRS in Chillingham, weren't present in the dbSNP and the 1000 Bull databases, highlighting the potential missed variation of other methods (Sayers *et al.*, 2019; Hayes and Daetwyler, 2019). This effect is exacerbated in Chillingham due to the non-uniform distribution of variable sites along the genome, with small regions of relatively dense variation interspersed with vast RoH (defined as heterozygosity peak windows; HPW; 0.643 SNPs kb^{-1}). Runs of homozygosity cover 91% of the Chillingham genome (F_{RoH} = 0.91), this exceeds observations even in commercial cattle that have experienced strong artificial selection, including Angus (F_{RoH} = 0.13) and North American Holstein (F_{RoH} = 0.10; Forutan *et al.*, 2018). The lack of variation across large runs of the breed's genome complicated linkage disequilibrium measurements as numerous recombination events in extended autozygous tracts would have occurred undetected, this restricted me from using SNeP as in previous chapters as the software requires smaller distances between SNPs to estimate

effective population sizes for more ancient time points (Barbato *et al.*, 2015). Even using MSMC2, abundant RoH disrupt local estimates of coalescent times between pairs of haplotypes (Wang *et al.*, 2020), and may be causative for the substantial, ancient decline observed in Chillingham.

One exploratory analysis frequently carried out on genetic data across multiple breeds or populations is investigating population structure used to contextualise admixture between subpopulations, shared origin, and explain ongoing as well as previous divergences. For **chapter two** and **chapter three,** population structure inferences were largely corroborated by multiple genetic sources (e.g. Decker *et al.*, 2009; Decker *et al.*, 2014; Orozco-terWengel *et al.*, 2015; Mastrangelo *et al.*, 2020) and historical accounts (Ajmone-Marsan *et al.*, 2010; Upadhyay *et al.*, 2017). Primarily, initial divisions between species (*Bos taurus* and *Bos indicus*), followed by a secondary split within taurine (European and African) and then indicine (African and Asiatic) was observed with ADMIXTURE, MDS and neighbour-net analyses. These four main groups represent relatively ancient and substantial splits between (~250,000 YA) and within (~8,000 – 10,000 YA) species, so are often clearly identifiable. Despite the addition of hybrid populations, the establishment of geographically disparate groups remained; African hybrids clustered between African taurine and indicine, while American hybrids clustered between Asiatic indicine and European taurine. Seemingly quite well supported with the data presented in this thesis, a more recent analysis including more Italian, Balkan and Anatolian cattle breeds argued for a more central, admixed relationship between hybrid populations (Mastrangelo *et al.*, 2020). Again, this highlights the important of extensive sampling to capture unique genetic variation, specifically of native breeds descended from populations along putative major migration routes from domestication centres.

Contrastingly, population structure analysis in **chapter four** refuted part of the original hypothesis that the Ryeland sheep breed would be contained to a single genetic cluster. **Chapter four** moves away from cattle and looks at Ryeland sheep. This is a particularly interesting ancient British breed that possesses phenotypes (i.e. coat quality, scrapie resistance, and resilience) with a perceived value which has been extremely temporally variable and thus resulting in a tumultuous demography. Population structure of the breed, demography, genetic diversity, inbreeding and runs of homozygosity were all characterised to get a holistic view of the breed. Surprisingly, ADMIXTURE analyses carried out on just the 60 Ryeland individuals had the lowest CV error at 2, indicating a split between the most Northerly flock and the remaining 51 individuals. This was unexpected as no observational or genetic inferences had suggested such a divide (Kelham *et al.*, 2013; Ryeland Flock Book Society, 2019). Nonetheless, possible causative differences were present: the separated flock had produced numerous prizewinning stock; had

been anecdotally noted for exceptional coat quality; and is exposed to a colder, wetter climate than most other flocks in the dataset. Initial hypothesis considered adaptive introgression (similar to observations in **chapter three**), as well as particularly strong selection; however, notably higher F_{IS}, IBD, F_{RoH} of the Northern cluster, and a lack of shared ancestry from ADMIXTURE analysis across 99 breeds all supported the hypothesis of genetic isolation and inbreeding being the causative mechanisms. Within-breed familial structure is not uncommon (e.g., Bray *et al.*, 2009; Deniskova *et al.*, 2018), often developing as individual farmers pursue improved phenotypes of personal value (Garforth, 2015). Importantly, this issue arose due to the sampling distribution – no other single farm was selected for such a large contribution (nine individuals) – the bias occurred due to an interest in coat colour genotyping and parentage analysis for a parallel project (not presented in this thesis). The presence of a division within the highlighted the importance in considering population structure prior to inferring selection as attempting to run SAMBADA analysis without defining population structure erroneously indicated the selection of thousands of loci.

Complex models were defined and simulated under an ABC framework with the aim to gather estimates of specific parameters and events occurring in the history of cattle. Perhaps the most provocative aim of **chapter two** was to determine if either two or three independent domestication events occurred in cattle: the Fertile Crescent (*Bos taurus*), the Indus Valley (*Bos indicus*) and a debated third event around the Western Desert of Egypt (African *Bos taurus*; Grigson, 1991; Bradley *et al.*, 1996; Applegate *et al.*, 2001; Stock and Gifford-Gonzalez, 2013). The direct comparison between scenarios modelling two or three domestication events, resulting in consistent signals – multiple demographic models and biological replicates – favouring the occurrence of only two domestications. While the primary objective was to compare different demographic scenarios to one another, it is important that posterior estimates capture a similar range to the putative academic consensus for a given parameter (Wegmann *et al.*, 2010). Resolution of posterior estimates can be *very* broad, this is due in part to the necessary simplification of models and also the condensing of genetic information into summary statistics that result in inflated credibility intervals (Sunnåker *et al.*, 2013; Gray *et al.*, 2014). The largest challenge was the discrepancy of magnitude between inter- and then intra-species divergence events. Attempting to characterise an event that occurred ~250,000 YA with events occurring ~8,000 YA in the same model, confounded by uncertain estimates of generation length over such a long period, resulted in potentially inaccurate ancient parameter inferences. By focussing on a shorter historical period of ~600 years, using only recently diverged breeds in **chapter three**, it was easier to elucidate a more precise demographic history. One particularly noteworthy parameter that is difficult to directly contextualise with biological values is the mutation rate. Substitution (i.e. point mutations causing SNPs) rate is estimated to be $7x10^{-11}$ and $6.5x10^{-11}$ for

Hanwoo and Holstein breeds, respectively (Lee and Shin, 2018); however, not only is substitution rate variable across the genome (Goldman and Yang, 1994), during the process of selecting and filtering SNP arrays there is an inherent favourable selection of highly variable loci, inflating the apparent substitution rate of the data (Excoffier and Foll, 2011). As the ABC models were progressively refined, posterior distributions of substitution rate centred around ~$1x10^{-4}$ – $1x10^{-5}$ in **chapter two**, with modes as low as ~$1x10^{-3}$ for the Iberian-Creole cattle panel in **chapter three**. This is difficult (and erroneous) to interpret as a direct reflection of biological processes, but should be reflective of the substitution rate of the particular subset of data and the respective variability in the subset of markers selected (Gray *et al.*, 2014). Additionally, definitive answers were elusive for the occurrence and strength of migratory events in the ABC modelling, apparent from the low Bayes factors discriminating between models (scenarios 8, 11, and 14). Migration between Asiatic indicine and European taurine was difficult to determine and later analyses with TREEMIX indicated only a single weak migration edge between the French taurine Montbeliarde breed and the Asiatic indicine breeds; however, admixture between African taurine and African indicine appeared more certain – present in all three favoured ABC model, strong migration edges in TREEMIX, and shared population ancestry in ADMIXTURE. Sympatric occurrence of both domesticated cattle species likely facilitated strong admixture, potentially introducing many locally adaptive traits into the more recent indicine migrants (Chen *et al.*, 2018; Utsunomiya *et al.*, 2019).

Chapter three focusses on the ancestry, demography, and selection in Creole and Iberian cattle with twenty-nine new individuals sampled. The extensive dataset collated for **chapter two**, was used to provide context for admixture, adaptive introgression, and ancestral populations. This system provided an excellent *in situ* experiment for rapid climate change, with test populations of cattle (Creole) translocated from a temperate region to a tropical region (Spain and Portugal to the Neotropics during colonial times), while the putative ancestors (Iberian) remained in their ancestral temperate region. Where ABC modelling in **chapter two** relied upon previous archaeozoological and genetic data for prior estimates spanning a large timeframe, for **chapter three** multiple historical records were available alongside the timeframe of interest. These records included estimates of colonisation time, number (Rodero *et al.*, 1992), subsequent expansions, contractions (Villalobos Cortés *et al.*, 2009), as well as restocking and replacement of the Creole population from European taurine, African taurine and indicine (Willham, 1982; Felius *et al.*, 2014; Huson *et al.*, 2014). While obviously potentially error-prone, they provide good prior estimates for ABC modelling. Similarly to **chapter two**, migration patterns were difficult to discern, with the occurrence and timing of restocking from Iberian breeds (scenarios 6 and 7) inseparable with Bayes factors from scenario 2 which indicated no restocking since divergence

between Creole and Iberian Nonetheless, these three favoured scenarios all incorporated the reported population expansions post-colonisation by Creole cattle, with the colonising N_e of 84 reaching 57,278 over ~350 years. Importantly, the population contraction of Creole breeds was apparent in ABC modelling and further corroborated by SNEP profiles, which further implies an ongoing decline at least up until 13 generations ago (~65 YA). As recent reports (e.g. Ginja *et al.*, 2019; Sponenberg *et al.*, 2019) suggest further marginalisation of Creole breeds, it is possible that the contemporary N_e is lower than SNEP estimates (~70 for all breeds except Texas Longhorn). The implementation of the novel NeS method to assess the degree of change of SNEP slopes offered more detailed information about any acceleration of population decline. Furthermore, the inter-generational synchronised nature of N_e fluctuations observed in Creole cattle would be unapparent without the NeS analysis.

This thesis includes populations exposed to shifting environments, locations, and management practices over the recent past, interactions with which alter genetic variation, a key aim was to understand the selective processes, evolutionarily significant loci, and biological processes affected by external change. While there were potential indications of adaptive introgression identified through shared ancestry in **chapter two**, signatures of selection were first explicitly tested for in **chapter three**. Selective sweeps were expected to have occurred in the Creole populations, leaving behind possible genomic signatures of genetic differentiation (here measured by F_{ST}) and/or haplotype divergence (here measured by XPEHH) relative to the putative ancestral populations as advantageous alleles and thus surrounding linked variation is swept to fixation (see **chapter one** for a review). On occasion, strikingly clear signals of selection can be observed, such as polledness in sheep (Kijas *et al.*, 2012). An interesting analogous comparison was the identification of a region involved in the polled phenotype of Senepol which while it was in line with previous studies (Flori *et al.*, 2012; Medugorac *et al.*, 2012), it lacked the precision of a causative mutation or even a single genic region. Nonetheless, the relatively conservative analysis that required repeated consistent signals to pass quality control, generated 49 distinct windows of putative selection across Creole breeds. Despite broad functional implications of the selected regions, there was seemingly a particular high incidence of thermotolerance and immunity-related processes underlining the two major physiological challenges the cattle experienced during adaptation to the tropics (Huson *et al.*, 2014). A key benefit to working with cattle (in particular, *Bos taurus*) data is the extensive quantitative trait loci data available (e.g. http://www.animalgenome.org; Shamimuzzaman *et al.*, 2020), which provides very narrow windows and often specific point mutations of alleles with measurable phenotypic effects; contrastingly, most publicly available sheep QTL span several hundred or thousand base pairs. Finally, the LAMP analyses identified the origin of the selected regions, estimating any

deficiencies or excesses of proportional ancestry relative to genome-wide measures. Most regions showed little variance from Creole ancestry and likely arose from a selective sweep acting on novel or standing (but previously neutral) variation; however, fourteen regions had indicine enrichment, implying adaptive introgression has had substantial influence in the success of Creole breeds and simultaneously providing potential gene targets conferring climate resilience (Hansen, 2020). **Chapter three** takes two closely related groups (Creole and Iberian cattle) that have been exposed to different climatic conditions and makes *inferences* of environmentally driven selection, however, does not explicitly test this assumption. To some extent it is a valid - due to the considerable difference in environments and relatively instantaneous translocation of Creole cattle to a different continent that reduced further contact between the two groups (Villalobos Cortés *et al.*, 2009). However, without clear quantification of the magnitude of different environmental factors, selection signatures may be confounded by other effects, for example farmer driven selection.

The addition of environmental variables in **chapter four**, including topography, long term climate, and land usage provides a quantitative dataset from which to extract genotype associations, allowing identification of more subtle local adaptation across multiple environmental gradients. Only four genotypes from three loci were identified as having significant environmental variable associations, however, the candidate genes identified (*TMEM123*, *DNAJC25*, and *BZW2*) were all involved with apoptotic processes and cancers, in particular hepatocellular carcinoma (Ma *et al.*, 2002; Liu *et al.*, 2012; Jin *et al.*, 2019). Given that: firstly, the significant environmental associations included rainfall, humidity, and longitude (the latter often correlated with the former two across Great Britain); secondly, there is a higher incidence of liver fluke in wetter regions (Fox *et al.*, 2011; Machicado *et al.*, 2016); and finally, liver fluke infection increases the risk of developing hepatocellular carcinoma (Xia *et al.*, 2015; Machicado *et al.*, 2016) – there is a compelling narrative that Ryeland display significant local adaptation in mitigating the effects of liver fluke infection. Such claims would require specific environmental and pathological data collection to be validated. Additionally, as previously discussed expanded on in **chapter three**, the ease of moving livestock, frequent shifts in desired traits, and demographic fluctuations all act to disguise or disrupt clear signatures of selection. This methodology is more applicable to established sedentary populations naturally expanding and exhibiting isolation by distance (e.g. Stucki *et al.*, 2017; Capblancq *et al.*, 2018; Duruz *et al.*, 2019). Furthermore, signatures of selection were isolated at a breed level through comparisons with other British, European, and Asian breeds. XPEHH analysis identified 262 genes consistently under selection against British breeds, allowing characterisation of some unique traits in Ryeland. The resulting ontological pathways (e.g. biosynthesis of antibiotics and various metabolic pathways) seem to support the breed's

noted resilience, both to general disease burden and poor feed quality (Youatt, 1837). While the KEGG pathway for 'prion disease' further supported Ryeland's previously observed resistance to scrapie (Gordon, 1966; Townsend *et al.*, 2005; Melchior *et al.*, 2011). As XPEHH requires comparative breeds as reference populations, candidate loci, genes, and ontological pathways identified are of potential value to the society, as it further emphasises the unique qualities of Ryeland in a competitive livestock market that increasingly marginalises non-commercial animals.

In contrast to **chapter three** and **chapter four**, which characterised positive selection through haplotype- and population differentiation-based methods, **chapter five** attempts to characterise balancing selection as the force maintaining genetic variation in the Chillingham genome. When interpreting selective signals and the functional effects, SNP arrays rely on linkage disequilibrium over large distances to identify candidate genes. While this is useful, particularly in regions with extended haplotypes (e.g. Colombian and Senepol breed clusters on chromosome 20; Figure 3.9) that may incorporate a multitude of functional genes, it is challenging to determine a single causative marker or genes due to an average spacing between SNPs of 50 – 100 kb after filtering (The Bovine HapMap Consortium *et al.*, 2009; Kijas *et al.*, 2012). Contrastingly, utilising WGRS allowed for identification of variation at a greater density, reducing the requirement for inferences across long distances. Paired with methodologies such as SIFT (Sim *et al.*, 2012), both the modified protein sequence and the tolerability of each mutation was predicted to give an indication of possible genetic load or adaptive potential. In line with the observations in Chillingham, it seems commonplace in cattle for balancing selection to be accompanied by an increase in deleterious alleles (Fasquelle *et al.*, 2009; Kadri *et al.*, 2014); it seems probable that the efficacy of purging selection is reduced when neighbouring loci are exposed to high levels of balancing selection (Lenz *et al.*, 2016).

6.3 FUTURE DIRECTIONS

This thesis included datasets ranging from expansive global panels to a comparative study of a single population with ten samples, across both sheep and two species of cattle. Each chapter was largely successful in addressing contextualised questions related to the given population(s); however, there is both additional work that would supplement the current research and improvements that would be beneficial when considering similar analyses.

Expanding the analyses by incorporating individual phenotypic data could be beneficial in quantifying the impact and adaptation to different environments. Genome-wide association studies (GWAS) and genomic breeding values (GEBVs) could be a better method of identifying subtle or polygenic environmentally adaptive traits (Hansen, 2020). Furthermore, the inclusion of thorough pedigree data can help understand confounding demographic factors, heritability of phenotypes, and mating systems. Unfortunately, quantitative genetics require substantial sampling, usually hundreds to thousands of individuals, therefore, funding requirements are often only met for commercial breeds, with a focus on profitable traits such as productivity and fertility, rather than environmental resilience (Porto-Neto *et al.*, 2014). Similar research provided the excellent QTL databases utilised in this thesis, highlighting the value of such a resource and the potential benefits if QTLs conferring adaptation to climatic conditions could be identified with the same precision.

Differences in environmental conditions throughout the thesis were often derived or assumed rather than directly measured (e.g. Iberian vs. central and southern American climates in **chapter three**; or, environmental variables surrounding farms potentially not representative of long-term exposure in **chapter four**). A full climate-controlled study across several generations of multiple livestock species with test groups exposed to harsher climatic conditions would provide quantitative evidence of important loci under selection. Clearly, funding is limited for such studies and enforcing such stresses on livestock indefinitely is incredibly unethical; however, smaller scale studies are already quantifying the impact of factors such as heat stress on dairy cattle production and metabolism (Al-Qaisi *et al.*, 2020; Hou *et al.*, 2021). Expanding to incorporate individuals carrying the candidate alleles identified throughout this thesis and similar studies would provide a measurable benefit for each locus. *In vitro* cell line cultures could provide initial functional testing of the candidate genes and resultant protein expression identified within this thesis (Alves *et al.*, 2019). Introduction of loci could then be achieved through crossbreeding, for example Mariasegaram *et al.* (2007) describe heat tolerance in Senepol X Holstein/Charolais/Angus due to the introduced slick hair coat loci. Contrastingly, instead of traditional introgressive methods, a more targeted approach with gene editing tools (e.g. CRISPR-Cas9) could improve efficiency, reduce outbreeding depression, and retain productivity traits in commercial breeds (Hansen, 2020). The previous chapters highlight potentially high impact genes, including *GDNF* for the slick phenotype (**chapter three**) as well as interactions between hepatocellular carcinoma (HCC) and *DNAJC25*, *BZW2*, and *C7* (**chapter four**). Specifically, on the latter system within Ryeland sheep, an attempt was made to collate medical information and questionnaire data with the society and farmers. Unfortunately, this side project never came to fruition, but collection of quantitative medical data about the parasitic load of liver fluke and

prevalence of liver cancers (especially HCC) could provide a layer of validation of the genomic findings.

The analysis of Chillingham data in **chapter five** only scratches the surface of an incredibly interesting system. Future analyses have been briefly mentioned, but some key directions include the attempt to characterise positive selection through asymptotic McDonald-Kreitman tests (Haller and Messer, 2017). Impeded by a lack of reliable ancestral allelic data, the sequencing of closely related conspecifics (Welsh White Park and Vaynol) or ancient ancestral samples would provide a better reference. As there was a keen interest in describing the HPW that persist in Chillingham, it was natural to consider the major histocompatibility complex (MHC). Visscher *et al.* (2001) identified only a single allele in a microsatellite adjacent to the MHC in Chillingham – while an interesting find, it was unexpected that with WGRS identified zero HPW across regions renown for high polymorphism (Hedrick, 1994). While there were numerous polymorphic sites identified in the MHC regions of Angus, detection of just SNPS from WGRS may have bypassed multiple regions rich in copy number variants (CNV). Filtering read data using a maximum read depth removes short reads derived from CNV that erroneously map to repeats or paralogous genes, the addition of long read sequencing would improve the ability to detect CNV and thus any immunity-driven selection present in Chillingham (Zorc *et al.*, 2019; Hu *et al.*, 2020). Diversity and variability of CNV in cattle is breed-specific, nonetheless, seven genomic regions including those dense with MHC loci and olfactory receptor genes are persistent across cattle (Hu *et al.*, 2020). Furthermore, the specific regions of HPW in Chillingham had similar metrics (e.g. synonymous to non-synonymous ratios) when isolated in Angus – application to other breeds may validate these regions as frequently influenced by balancing selection. It would be interesting if future functional studies could link any of the identified deleterious variation with known congenital conditions (testicular hypoplasia, molar defects, subfertile males, and poor semen quality) in Chillingham (Ingham, 2002; Hall *et al.*, 2005; Hall and Bunce, 2019). Overall, in addition to the analyses already carried out this would provide a reasonably comprehensive description of Chillingham and more importantly – a framework for which other similar populations can be assessed by. Including additional omics technologies would provide insight into the dynamic, short-term regulatory and adaptive processes of individuals in response to external conditions (Zampiga *et al.*, 2018). Transcriptomics, metabolomics, and proteomics could all be used to investigate real time responses through the expression levels of genes, metabolites, and proteins, respectively. While epigenetic modification of the genome for regulatory functions – studied through epigenomics – is increasingly considered as a steppingstone to promoting permanent, heritable genetic change (Sevane *et al.*, 2019).

For Creole cattle, Ryeland sheep, and Chillingham cattle the previous genetic knowledge was particularly limited, thus these studies have provided an important basis for future genetic work on these breeds. Nonetheless, the questions addressed within this thesis can be applied to a broader context. Domestic species, both plants and animals are widely facing challenges associated with elevated inbreeding, strong artificial selection which may work antagonistically to fertility and health, and replacement of traditional breeds in lieu of more productive breeds. Understanding the consequence and mitigation strategies of rapidly declining and sustained low N_e will improve the survival both threatened livestock breeds and endangered wild populations that are increasingly fragmented and isolated into inbred subpopulations. Similarly, the signatures of selection and adaptive introgression identified in this thesis describe the specific challenges faced by the analysed population. While the loci identified (e.g. the slick genotype) may be directly applicable to conspecifics or more conserved regions (e.g. MHC) informative to broader taxa, the characterisation of the adaptive process and disentangling introgressive signals can be applied to many divergent systems. Climate change is rapidly altering environments that have been relatively constant for millennia and due to fragmentation of natural ecosystems, migration or dispersal will no longer be a viable survival strategy for many natural populations, therefore, understanding rapid selection, local adaptation and adaptive introgression may be essential for the persistence of endangered species. Livestock continue to provide an excellent evolutionary model with detailed historical records and significant genetic resources; future research on livestock populations for both food security and wider scientific application should not be understated.

Bibliography

Abdela, N. and Jilo, K. (2016). Impact of Climate Change on Livestock Health: A Review. *Global Veterinaria* **16**:419–424.

Abebe, A.S. *et al.* (2020). Breeding practices and trait preferences of smallholder farmers for indigenous sheep in the northwest highlands of Ethiopia: Inputs to design a breeding program. *PLoS ONE* **15**:1–18.

Acevedo-Whitehouse, K. and Cunningham, A.A. (2006). Is MHC enough for understanding wildlife immunogenetics? *Trends in Ecology and Evolution* **21**:433–438.

Achilli, A. *et al.* (2008). Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Current Biology* **18**:157–158.

Adly, M.A. *et al.* (2006). Analysis of the expression pattern of glial cell line-derived neurotrophic factor, neurturin, their cognate receptors GFRalpha-1 and GFRalpha-2, and a common signal transduction element c-Ret in the human scalp skin. *Journal of cutaneous pathology* **33**:799–808.

Ajmone-Marsan, P. *et al.* (2009). New World cattle show ancestry from multiple independent domestication events. *BMC Genomics* **5**:18644–18649.

Ajmone-Marsan, P. *et al.* (2010). On the origin of cattle: How aurochs became cattle and colonized the world. *Evolutionary Anthropology: Issues, News, and Reviews* **19**:148–157.

Akey, J.M. (2009). Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Research* **19**:711–722.

Al-Qaisi, M. *et al.* (2020). Validating a heat stress model: The effects of an electric heat blanket and nutritional plane on lactating dairy cows. *Journal of Dairy Science* **103**:5550–5560.

Alba, J. de (1987). Criollo Cattle of Latin America. *FAO Animal Production and Health Paper* **66**:17–39.

Alberto, F.J. *et al.* (2018). Convergent genomic signatures of domestication in sheep and goats. *Nature Communications* **9**:813.

Albrechtsen, A. *et al.* (2010). Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. *Molecular Biology and Evolution* **27**:2534–2547.

Alexander, D.H. *et al.* (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**:1655–1664.

Alexandratos, N. and Bruinsma, J. (2012). World agriculture towards 2030/2050: the 2012 revision. *ESA Working paper No. 12-03. Rome, FAO.* **12**:146.

Allison, A.C. (1956). The sickle-cell and haemoglobin C genes in some African populations. *Annals of Human Genetics* **21**:67–89.

Almathen, F. *et al.* (2016). Ancient and modern DNA reveal dynamics of domestication and cross-continental dispersal of the dromedary. *Proceedings of the National Academy of Sciences* **113**:6707–6712.

Alves, J.M. *et al.* (2019). Parallel adaptation of rabbit populations to myxoma virus. *Science* **363**:1319–1326.

Andrews, S. *et al.* (2015). FastQC. A quality control tool for high throughput sequence data. Babraham Bioinformatics. *Babraham Institute* **1**:1.

Applegate, A. *et al.* (2001). The North Tumuli of the Nabta Late Neolithic Ceremonial Complex. *Holocene Settlement of the Egyptian Sahara*:468–488.

Aucamp, P.J. (2003). Eighteen questions and answers about the effects of the depletion of the ozone layer on humans and the environment. *Photochemical and Photobiological Sciences* **2**:ix–xxiv.

Babik, W. *et al.* (2005). Sequence diversity of the MHC DRB gene in the Eurasian beaver (Castor fiber). *Molecular Ecology* **14**:4249–4257.

Babik, W. *et al.* (2009). Long-term survival of a urodele amphibian despite depleted major histocompatibility complex variation. *Molecular Ecology* **18**:769–781.

Bagath, M. *et al.* (2019). The impact of heat stress on the immune system in dairy cattle: A review. *Research in Veterinary Science* **126**:94–102.

Bahbahani, H. *et al.* (2017). Signatures of selection for environmental adaptation and zebu × taurine hybrid fitness in East African Shorthorn Zebu. *Frontiers in Genetics* **8**:68.

Barbato, M. *et al.* (2017). Genomic signatures of adaptive introgression from European mouflon into domestic sheep. *Scientific Reports* **7**:7623.

Barbato, M. *et al.* (2015). SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Frontiers in Genetics* **6**:109.

Beauchemin, K.A. *et al.* (2008). Nutritional management for enteric methane abatement: A review. *Australian Journal of Experimental Agriculture* **48**:21–27.

Behl, J.D. *et al.* (2012). The Major Histocompatibility Complex in Bovines: A Review. *ISRN Veterinary Science* **2012**:1–12.

Beja-Pereira, A. *et al.* (2003). Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nature Genetics* **35**:311–313.

Beja-Pereira, A. *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.

Bellott, D.W. *et al.* (2014). Mammalian y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**:494–499.

Benchaar, C. *et al.* (2001). Evaluation of dietary strategies to reduce methane production in ruminants: A modelling approach. *Canadian Journal of Animal Science* **81**:563–574.

Benjamini, Y. and Hochberg, Y. (1966). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Source Journal of the Royal Statistical Society. Series C (Applied Statistics)* **15**:216–233.

Bernabucci, U. *et al.* (2010). Metabolic and hormonal acclimation to heat stress in domesticated ruminants. *Animal The Animal Consortium* **4**:1167–1183.

Berteaux, D. and Micol, T. (1992). Population studies and reproduction of the feral cattle (Bos taurus) of Amsterdam Island, Indian Ocean. *Journal of Zoology* **228**:265–276.

Berthouly, C. *et al.* (2009). How does farmer connectivity influence livestock genetic structure? A case-study in a Vietnamese goat population. *Molecular Ecology* **18**:3980–3991.

Beynon, S.E. *et al.* (2015). Population structure and history of the Welsh sheep breeds determined by whole genome genotyping. *BMC Genetics* **16**:1–14.

Bhatia, G. *et al.* (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research* **23**:1514–1521.

Bhimte, A. *et al.* (2018). Endocrine changes in livestock during heat and cold stress. *Journal of Pharmacognosy and Phytochemistry* **7**:127–132.

Bieber, A. *et al.* (2020). Comparison of performance and fitness traits in German Angler, Swedish Red and Swedish Polled with Holstein dairy cattle breeds under organic production. *Animal* **14**:609–616.

Birch, J. *et al.* (2008). Genomic location and characterisation of nonclassical MHC class I genes in cattle. *Immunogenetics* **60**:267–273.

Boadi, D. *et al.* (2004). Mitigation strategies to reduce enteric methane emissions from dairy cows: Update review. *Canadian Journal of Animal Science* **84**:319–335.

Van Boeckel, T.P. *et al.* (2017). Reducing antimicrobial use in food animals. *Science* **357**:1350–1352.

Bolger, A.M. *et al.* (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120.

Bollongino, R. *et al.* (2012). Modern taurine cattle descended from small number of near-eastern founders. *Molecular Biology and Evolution* **29**:2101–2104.

Bonfiglio, S. *et al.* (2012). Origin and spread of Bos taurus: New clues from mitochondrial genomes belonging to haplogroup T1. *PLoS ONE* **7**:1–10.

Bonhomme, M. *et al.* (2010). Detecting selection in population trees: The Lewontin and Krakauer test extended. *Genetics* **186**:241–262.

Bosse, M. *et al.* (2014). Untangling the hybrid nature of modern pig genomes: a mosaic derived from biogeographically distinct and highly divergent *Sus scrofa* populations. *Molecular Ecology* **23**:4089–4102.

Bradley, D.G. *et al.* (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.

Bray, T.C. *et al.* (2009). The population genetic effects of ancestry and admixture in a subdivided cattle breed. *Animal Genetics* **40**:393–400.

Bro-Jørgensen, M.H. *et al.* (2018). Ancient DNA analysis of Scandinavian medieval drinking horns and the horn of the last aurochs bull. *Journal of Archaeological Science* **99**:47–54.

Broad Institute (2019). Picard Toolkit. *GitHub Repository*:http://broadinstitute.github.io/picard/.

Brown, C. *et al.* (2014). Assessment of inbreeding resulting from selection for scrapie resistance: A model for rare sheep breeds. *Veterinary Record* **175**:624.

Browning, B.L. and Browning, S.R. (2013). Detecting identity by descent and estimating genotype error rates in sequence data. *American Journal of Human Genetics* **93**:840–851.

Browning, S.R. and Browning, B.L. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics* **81**:1084–1097.

Bruce, M.E. *et al.* (1997). Transmissions to mice indicate that 'new variant' CJD is caused by the BSE agent. *Nature* **389**:498–501.

Bruford, M.W. *et al.* (2003). DNA markers reveal the complexity of livestock domestication. *Nature Reviews Genetics* **4**:900–910.

Buggiotti, L. *et al.* (2021). Demographic History, Adaptation, and NRAP Convergent Evolution at Amino Acid Residue 100 in the World Northernmost Cattle from Siberia. *Molecular Biology and Evolution* **38**:3093–3110.

Cadzow, M. *et al.* (2014). A bioinformatics workflow for detecting signatures of selection in genomic data. *Frontiers in Genetics* **5**:1–8.

Cao, Y.H. *et al.* (2021). Historical Introgression from Wild Relatives Enhanced Climatic Adaptation and Resistance to Pneumonia in Sheep. *Molecular Biology and Evolution* **38**:838–855.

Capblancq, T. *et al.* (2018). Evaluation of redundancy analysis to identify signatures of local adaptation. *Molecular Ecology Resources* **18**:1223–1233.

Chang, C.C. *et al.* (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **4**:1–16.

Charbonnel, N. and Pemberton, J. (2005). A long-term genetic survey of an ungulate population reveals balancing selection acting on MHC through spatial and temporal fluctuations in selection. *Heredity* **95**:377–388.

Charlesworth, D. and Willis, J.H. (2009). The genetics of inbreeding depression. *Nature Reviews Genetics* **10**:783–796.

Chen, F. *et al.* (2015). Nuclear Export of Smads by RanBP3L Regulates Bone Morphogenetic Protein Signaling and Mesenchymal Stem Cell Differentiation. *Molecular and Cellular Biology* **35**:1700–1711.

Chen, N. *et al.* (2018). Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. *Nature Communications* **9**:1–13.

Chen, X. *et al.* (2013). Molecular Characterization of Severin from Clonorchis sinensis Excretory/Secretory Products and Its Potential Anti-apoptotic Role in Hepatocarcinoma PLC Cells. *PLoS Neglected Tropical Diseases* **7**:e2606.

Cheng, D.-D. *et al.* (2017). Downregulation of BZW2 inhibits osteosarcoma cell growth by inactivating the Akt/mTOR signaling pathway. *Oncology Reports* **38**:2116.

Chessa, B. *et al.* (2009). Revealing the History of Sheep Domestication Using Retrovirus Integrations. *Science* **324**:532–536.

Clark, P.U. *et al.* (2009). The Last Glacial Maximum. *Science* **325**:710–714.

Clutton-Brock, J. (1989). A Natural History of Domesticated Mammals. *The Beagle : Records of the Museums and Art Galleries of the Northern Territory* **6**:246–247.

Copernicus Land Monitoring Service and European Environment Agency (2019). European Union Digital Elevation Mapping v1.1.

Cornuet, J.-M. *et al.* (2008). Inferring population history with DIYABC: a user-friendly approach to Approximate Bayesian Computation. **00**:1–7.

Cothran, G. *et al.* (2011). European domestic horses originated in two holocene refugia. *PLoS ONE* **6**.

Crnokrak, P. and Barrett, S.C.H. (2002). Perspective: Purging the genetic load: A review of the experimental evidence. *Evolution* **56**:2347–2358.

Croze, M. *et al.* (2016). Balancing selection on immunity genes: review of the current literature and new analysis in Drosophila melanogaster. *Zoology* **119**:322–329.

Cullen, J.M. and Popp, J.A. (2002). Tumors of the Liver and Gall Bladder. In: *Tumor in Domestic Animals*. 4th ed. Iowa State Press, Blackwell Publishing Co., New York, pp. 483–499.

den Daas, N. (1992). Laboratory assessment of semen characteristics. *Animal Reproduction Science* **28**:87–94.

Daetwyler, H.D. (2021). 10 years of the 1000 Bull Genomes Project: impact on cattle breeding. *Easter Bush Research Seminar*.

Daetwyler, H.D. *et al.* (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics* **46**:858–865.

Danecek, P. *et al.* (2011). The variant call format and VCFtools. *Bioinformatics* **27**:2156–2158.

Darwin, C. and Wallace, A. (1858). On the tendency of species to form varieties: and on the perpetuation of varieties and species by natural means of selection. *Journal of the Proceedings of the Linnean Society of London* **3**:45–62.

Davies, C.J. *et al.* (2006). Evidence for Expression of Both Classical and Non-Classical Major Histocompatibility Complex Class I Genes in Bovine Trophoblast Cells. *American Journal of*

Reproductive Immunology **55**:188–200.

Dawson, M. *et al.* (2008). Progress and limits of PrP gene selection policy. *Veterinary Research* **39**:1–12.

Day, S.B. *et al.* (2003). The influence of variable rates of inbreeding on fitness, environmental responsiveness, and evolutionary potential. *Evolution* **57**:1314–1324.

Dearborn, D.C. *et al.* (2016). Gene duplication and divergence produce divergent MHC genotypes without disassortative mating. *Molecular ecology* **25**:4355–4367.

Decker, J.E. *et al.* (2009). Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proceedings of the National Academy of Sciences* **106**:18644–18649.

Decker, J.E. *et al.* (2014). Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle. *PLoS Genetics* **10**:e1004254.

DEFRA (2021). Farming Statistics - Livestock Populations at 1 December 2013 , United Kingdom.

DEFRA (2001). National Scrapie Plan for Great Britain. **1**:1–28.

Deng, X. *et al.* (2013). Genome wide association study (GWAS) of chagas cardiomyopathy in trypanosoma cruzi seropositive subjects. *PLoS ONE* **8**:e79629.

Deniskova, T.E. *et al.* (2018). Population structure and genetic diversity of 25 Russian sheep breeds based on whole-genome genotyping. *Genetics Selection Evolution* **50**:29.

Diamond, J. (2002). Evolution, consequences and future of plant and animal domestication. *Nature* **418**:700–707.

Dlugosch, K.M. and Parker, I.M. (2008). Founding events in species invasions: Genetic variation, adaptive evolution, and the role of multiple introductions. *Molecular Ecology* **17**:431–449.

Dragani, T.A. (2010). Risk of HCC: Genetic heterogeneity and complex genetics. *Journal of Hepatology* **52**:252–257.

Duforet-Frebourg, N. *et al.* (2016). Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data. *Molecular Biology and Evolution* **33**:1082–1093.

Durinck, S. *et al.* (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**:3439–3440.

Duruz, S. *et al.* (2019). Rapid identification and interpretation of gene–environment associations

using the new R.SamBada landscape genomics pipeline. *Molecular Ecology Resources* **19**:1355–1365.

Easterling, W.E. *et al.* (2007). Food, fibre and forest products. *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*:273–313.

Edwards, C.J. *et al.* (2010). A Complete Mitochondrial Genome Sequence from a Mesolithic Wild Aurochs (Bos primigenius). **5**:0–8.

Egea, R. *et al.* (2008). Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic acids research* **36**:157–162.

Ehiobu, N.G. *et al.* (1989). Effect of rate of inbreeding on inbreeding depression in Drosophila melanogaster. *Theoretical and Applied Genetics* **77**:123–127.

van Eijk, M.J.T. *et al.* (1995). Genetic mapping of BoLA-A, CYP21, DRB3, DYA, and PRL on BTA23. *Mammalian Genome* **6**:151–152.

Eisler, M.C. *et al.* (2014). Agriculture: Steps to sustainable livestock. *Nature* **507**:32–34.

Ekarius, C. (2008). Storey's illustrated breed guide for sheep goats cattle and pigs. **5**:319.

Ellis, S.A. and Hammond, J.A. (2014). The functional significance of cattle major histocompatibility complex class i genetic diversity. *Annual Review of Animal Biosciences* **2**:285–306.

EPA (2020). Global Mitigation of Non-CO2 Greenhouse Gases: 2010-2030. *Electrochemical Engineering* **9**:290–369.

Excoffier, L. *et al.* (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics* **9**:e1003987.

Excoffier, L. and Foll, M. (2011). fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**:1332–1334.

Excoffier, L. and Lischer, H.E.L. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**:564–567.

FAO (1992). *Guidelines on in Vivo Conservation of Animal Genetic Resources*. Rome, Italy.

FAO (2011). *Mapping Supply and Demand for Animal Source Foods to 2030.* Animal Pro. Food and

Agriculture Organization of the United Nations, Rome, Italy.

FAO (2015). The Second Report on the State of the World's Animal Genetic Resources for Food and Agriculture.

FAO (2009). *The State of Food and Agriculture: Livestock in the Balance*. Food and Agriculture Organization of the United Nations, Rome, Italy.

FAO *et al.* (2018). *Food Security and Nutrition in the World the State of Building Climate Resilience for Food Security and Nutrition*.

Fasquelle, C. *et al.* (2009). Balancing selection of a frame-shift mutation in the MRC2 gene accounts for the outbreak of the crooked tail syndrome in Belgian blue cattle. *PLoS Genetics* **5**:e1000666.

Fay, J.C. and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**:1405–1413.

Felius, M. *et al.* (2011). On the breeds of cattle-Historic and current classifications. *Diversity* **3**:660–692.

Felius, M. *et al.* (2014). On the history of cattle genetic resources. *Diversity* **6**:705–750.

De Filippo, C. *et al.* (2016). Recent Selection Changes in Human Genes under Long-Term Balancing Selection. *Molecular Biology and Evolution* **33**:1435–1447.

Flori, L. *et al.* (2012). A quasi-exclusive European ancestry in the Senepol tropical cattle breed highlights the importance of the slick locus in tropical adaptation Caramelli, D. (ed.). *PLoS ONE* **7**:1–10.

Foley, J.A. *et al.* (2011). Solutions for a cultivated planet. *Nature* **478**:337–342.

Forutan, M. *et al.* (2018). Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle. *BMC Genomics* **19**:1–12.

Fox, N.J. *et al.* (2011). Predicting impacts of climate change on fasciola hepatica risk. *PLoS ONE* **6**:19–21.

Francis, R.M. (2017). pophelper: an R package and web app to analyse and visualize population structure. *Molecular Ecology Resources* **17**:27–32.

Frantz, L.A.F. *et al.* (2015). Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. *Nature Genetics* **47**:1141–1148.

Fujita, T. (2002). Evolution of the lectin - Complement pathway and its role in innate immunity. *Nature Reviews Immunology* **2**:346–353.

Gao, H. *et al.* (2019). BZW2 gene knockdown induces cell growth inhibition, G1 arrest and apoptosis in muscle-invasive bladder cancers: A microarray pathway analysis. *Journal of Cellular and Molecular Medicine* **23**:3905–3915.

Garforth, C. (2015). Livestock keepers' reasons for doing and not doing things which governments, vets and scientists would like them to do. *Zoonoses and Public Health* **62**:29–38.

Garza-Brenner, E. *et al.* (2017). Association of SNPs in dopamine and serotonin pathway genes and their interacting genes with temperament traits in Charolais cows. *Journal of Applied Genetics* **58**:363–371.

Gautier, M. *et al.* (2016). Deciphering the Wisent Demographic and Adaptive Histories from Individual Whole-Genome Sequences. *Molecular Biology and Evolution* **33**:2801–2814.

Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics* **201**:1555–1579.

Gautier, M. *et al.* (2017). rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Molecular Ecology Resources* **17**:78–90.

Gautier, M. *et al.* (2010). Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds. *PLoS ONE* **5**:1–11.

Geibel, J. *et al.* (2021). How array design creates SNP ascertainment bias. *PLoS ONE* **16**:1–23.

Gensterblum-Miller, E. *et al.* (2018). Novel Transcriptional Activity and Extensive Allelic Imbalance in the Human MHC Region. *The Journal of Immunology* **200**:1496–1503.

Georges, M. *et al.* (2019). Harnessing genomic information for livestock improvement. *Nature Reviews Genetics* **20**:135–156.

Gerber, P.J. *et al.* (2013). *Tackling Climate Change through Livestock – A Global Assessment of Emissions and Mitigation Opportunities*. Food and Agriculture Organization of the United Nations, Rome, Italy.

Gerber, P.J. *et al.* (2008). Decision support for spatially targeted livestock policies: Diverse examples from Uganda and Thailand. *Agricultural Systems* **96**:37–51.

Gholami, M.R. *et al.* (2006). Hepatocellular Carcinoma in Sheep. *Archives of Razi Institute* **61**:53–55.

Gibson, J. *et al.* (2006). Extended tracts of homozygosity in outbred human populations. *Human Molecular Genetics* **15**:789–795.

Ginja, C. *et al.* (2019). The genetic ancestry of American Creole cattle inferred from uniparental and autosomal genetic markers. *Scientific Reports* **9**:1–16.

Giovambattista, G. *et al.* (2013). Characterization of bovine MHC DRB3 diversity in Latin American Creole cattle breeds. *Gene* **519**:150–158.

Godde, C.M. *et al.* (2021). Impacts of climate change on the livestock food supply chain; a review of the evidence. *Global Food Security* **28**:100488.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**:725–736.

Gordon, W.S. (1966). Variation in susceptibility of sheep to scrapie and genetic implications. *Report of Scrapie Seminar* **1**:53–67.

Götherström, A. *et al.* (2005). Cattle domestication in the Near East was followed by hybridization with aurochs bulls in Europe. *Proceedings of the Royal Society B: Biological Sciences* **272**:2345–2350.

Gray, M.M. *et al.* (2014). Demographic history of a recent invasion of house mice on the isolated Island of Gough. *Molecular Ecology* **23**:1923–1939.

Gregory, T.R. (2008). Artificial Selection and Domestication: Modern Lessons from Darwin's Enduring Analogy. *Evolution: Education and Outreach* **2**:5–27.

Grigson, C. (1991). An African origin for African cattle? - some archaeological evidence. *The African Archaeological Review* **9**:119–144.

Grossen, C. *et al.* (2020). Purging of highly deleterious mutations through severe bottlenecks in Alpine ibex. *Nature Communications* **11**:1001.

Grossi, G. *et al.* (2019). Livestock and climate change: Impact of livestock on climate and mitigation strategies. *Animal Frontiers* **9**:69–76.

Guillot, G. *et al.* (2014). Detecting correlation between allele frequencies and environmental variables as a signature of selection: A fast computational approach for genome-wide studies. *Spatial Statistics* **8**:145–155.

Günther, T. and Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics* **195**:205–220.

Guo, H. *et al.* (2021). Changes in rumen microbiota affect metabolites, immune responses and antioxidant enzyme activities of sheep under cold stimulation. *Animals* **11**:1–15.

Halenius, A. *et al.* (2015). Classical and non-classical MHC i molecule manipulation by human cytomegalovirus: So many targets - But how many arrows in the quiver? *Cellular and Molecular Immunology* **12**:139–153.

Hall, S.J. *et al.* (2005). Management of the Chillingham wild white cattle. *Government Veterinary Journal* **15**:4–11.

Hall, S.J.G. (1989). Chillingham cattle: social and maintenance behaviour in an ungulate that breeds all year round. *Animal Behaviour* **38**:215–225.

Hall, S.J.G. (2006). Chillingham Park and its Wild White Cattle. *Journal of the Royal Agricultural Society of England* **167**:1–8.

Hall, S.J.G. and Bunce, R.G.H. (2019). The use of cattle Bos taurus for restoring and maintaining holarctic landscapes: Conclusions from a long-term study (1946–2017) in northern England. *Ecology and Evolution* **9**:5859–5869.

Hall, S.J.G. and Hall, J.G. (1988). Inbreeding and population dynamics of the Chillingham cattle (Bos taurus). *Journal of Zoology* **216**:479–493.

Haller, B.C. and Messer, P.W. (2017). AsymptoticMK: A web-based tool for the asymptotic McDonald-Kreitman test. *G3: Genes, Genomes, Genetics* **7**:1569–1575.

Hanotte, O. (2002). African Pastoralism: Genetic Imprints of Origins and Migrations. *Science* **296**:336–339.

Hansen, P.J. (2020). Prospects for gene introgression or gene editing as a strategy for reduction of the impact of heat stress on production and reproduction in cattle. *Theriogenology* **154**:190–202.

Hayes, B.J. and Daetwyler, H.D. (2019). 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annual Review of Animal Biosciences* **7**:89–102.

He, K. *et al.* (2021). Long-Read Genome Assemblies Reveal Extraordinary Variation in the Number and Structure of MHC Loci in Birds. *Genome biology and evolution* **13**:1–13.

He, Y. *et al.* (2020). Antibiotic resistance genes from livestock waste: occurrence, dissemination, and treatment. *npj Clean Water* **3**:1–11.

Hedrick, P.W. (1994). Evolutionary genetics of the major histocompatibilty complex. *American Naturalist* **143**:945–964.

Hedrick, P.W. (2011). Population genetics of malaria resistance in humans. *Heredity* **107**:283–304.

Hedrick, P.W. and Garcia-Dorado, A. (2016). Understanding Inbreeding Depression, Purging, and Genetic Rescue. *Trends in Ecology and Evolution* **31**:940–952.

Hedrick, P.W. *et al.* (1976). Genetic Polymorphism in Heterogeneous Environments. *Annual Review of Ecology and Systematics* **7**:1–32.

Hedrick, P.W. *et al.* (2001). Parasite resistance and genetic variation in the endangered Gila topminnow. *Animal Conservation* **4**:103–109.

Heerwaarden, B. Van *et al.* (2008). Population bottlenecks increase additive genetic variance but do not break a selection limit in rain forest Drosophila. *Genetics* **179**:2135–2146.

Heinke, J. *et al.* (2020). Water Use in Global Livestock Production—Opportunities and Constraints for Increasing Water Productivity. *Water Resources Research* **56**:e2019WR026995.

Helmer, D. *et al.* (2007). The development of the exploitation of products from Capra and Ovis (meat, milk, and fleece) from the PPNB to the Early Bronze in the northern Near East. *Anthropozoologica* **42**:41–69.

Hijmans, R.J. *et al.* (2019). raster: Geographic Data Analysis and Modeling. R package. https://CRAN.R-project.org/package=raster.

Hinz, A. *et al.* (2014). Assembly and function of the major histocompatibility complex (MHC) I peptide-loading complex are conserved across higher vertebrates. *Journal of Biological Chemistry* **289**:33109–33117.

Hoffmann, A.A. *et al.* (2017). Revisiting Adaptive Potential, Population Size, and Conservation. *Trends in Ecology and Evolution* **32**:506–517.

Hou, Y. *et al.* (2021). Comparing responses of dairy cows to short-term and long-term heat stress in climate-controlled chambers. *Journal of Dairy Science* **104**:2346–2356.

Hristov, A.N. *et al.* (2013). SPECIAL TOPICS-Mitigation of methane and nitrous oxide emissions from animal operations: I. A review of enteric methane mitigation options. *Journal of Animal Science* **91**:5045–5069.

Hu, Y. *et al.* (2020). Comparative analyses of copy number variations between Bos taurus and Bos indicus. *BMC Genomics* **21**:1–11.

Huang, D.W. *et al.* (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**:44–57.

Huang, W. *et al.* (2020). A rare MHC haplotype confers selective advantage in a free-living ruminant. *bioRxiv*:1–23.

Huang, W. *et al.* (2022). Contemporary selection on MHC genes in a free-living ruminant population. *Ecology Letters* **25**:828–838.

Hudson, G. *et al.* (2012). Unique mitochondrial DNA in highly inbred feral cattle. *Mitochondrion* **12**:438–440.

Huson, D.H. and Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution* **23**:254–267.

Huson, H.J. *et al.* (2014). Genome-wide association study and ancestral origins of the slick-hair coat in tropically adapted cattle. *Frontiers in Genetics* **5**:1–12.

Ingham, B. (2002). Dental anomalies in the Chillingham Wild White Cattle. *Transactions of the Natural History Society of Northumbria* **62**:169–175.

IPCC (2006). *2006 IPCC Guidelines for National Greenhouse Gas Inventories*. 5th ed. Prepared by the National Greenhouse Gas Inventories Programme, IGES, Japan.

IPCC (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

IPCC (2018). *Global Warming of 1.5°C: An IPCC Special Report on the Impacts of Global Warming of 1.5°C above Pre-Industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change.* World Meteorological Organization, Geneva, Switzerland.

Iso-Touru, T. *et al.* (2016). Genetic diversity and genomic signatures of selection among cattle breeds from Siberia, eastern and northern Europe. *Animal Genetics* **47**:647–657.

Jensen, P. (2006). Domestication - From behaviour to genes and back again. *Applied Animal Behaviour Science* **97**:3–15.

Jiang, F. *et al.* (2021). Analysis of whole-genome re-sequencing data of ducks reveals a diverse demographic history and extensive gene flow between Southeast/South Asian and Chinese populations. *Genetics Selection Evolution* **53**:1–17.

Jin, X. *et al.* (2019). Role of the novel gene BZW2 in the development of hepatocellular carcinoma. *Journal of Cellular Physiology* **234**:16592–16600.

Jombart, T. *et al.* (2009). Genetic markers in the playground of multivariate analysis. *Heredity* **102**:330–341.

Joost, S. *et al.* (2007). A spatial analysis method (SAM) to detect candidate loci for selection: Towards a landscape genomics approach to adaptation. *Molecular Ecology* **16**:3955–3969.

Jungbluth, T. *et al.* (2001). Greenhouse gas emissions from animal houses and manure stores. In: *Nutrient Cycling in Agroecosystems*. Springer, pp. 133–145.

Kadri, N.K. *et al.* (2014). A 660-Kb Deletion with Antagonistic Effects on Fertility and Milk Production Segregates at High Frequency in Nordic Red Cattle: Additional Evidence for the Common Occurrence of Balancing Selection in Livestock. *PLoS Genetics* **10**.

Kamalakkannan, R. *et al.* (2021). Evidence for independent domestication of sheep mtDNA lineage A in India and introduction of lineage B through Arabian sea route. *Scientific Reports* **11**:1–16.

Kardos, M. *et al.* (2018). Genomic consequences of intensive inbreeding in an isolated wolf population. *Nature Ecology & Evolution* **2**:124–131.

Kardos, M. *et al.* (2017). Inferring individual inbreeding and demographic history from segments of identity by descent in Ficedula flycatcher genome sequences. *Genetics* **205**:1319–1334.

Karimi, K. *et al.* (2016). Local and global patterns of admixture and population structure in Iranian native cattle. *BMC Genetics* **17**:1–14.

Kawecki, T.J. and Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology Letters* **7**:1225–1241.

Kelham, C. *et al.* (2013). *Ryeland Coat Colour Genotyping*. MRes Thesis, Cardiff University, Cardiff.

Kemper, K.E. *et al.* (2014). Selection for complex traits leaves little or no classic signatures of selection. *BMC Genomics* **15**.

Key, F.M. *et al.* (2014). Advantageous diversity maintained by balancing selection in humans. *Current Opinion in Genetics and Development* **29**:45–51.

Kijas, J.W. *et al.* (2012). Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection Tyler-Smith, C. (ed.). *PLoS Biology* **10**:e1001258.

Kim, Y. and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**:765–777.

Kloch, A. *et al.* (2018). Signatures of balancing selection in toll-like receptor (TLRs) genes - Novel insights from a free-living rodent. *Scientific Reports* **8**:1–10.

Knapp, J.R. *et al.* (2014). Invited review: Enteric methane in dairy cattle production: Quantifying the opportunities and impact of reducing emissions. *Journal of Dairy Science* **97**:3231–3261.

Koops, W.J. *et al.* (1995). A Model for Reproductive Efficiency of Dairy Bulls. *Journal of Dairy Science* **78**:921–928.

Kovacs, G.G. *et al.* (2004). Complement activation in human prion disease. *Neurobiology of Disease* **15**:21–28.

Krol, K.M. *et al.* (2015). Genetic variation in CD38 and breastfeeding experience interact to impact infants' attention to social eye cues. *Proceedings of the National Academy of Sciences* **112**:E5434–E5442.

Lande, R. (1994). Risk of Population Extinction from Fixation of New Deleterious Mutations. *Evolution* **48**:1460.

Larson, G. *et al.* (2005). Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science* **307**:1618–1621.

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

Laws, R.J. and Jamieson, I.G. (2011). Is lack of evidence of inbreeding depression in a threatened New Zealand robin indicative of reduced genetic load? *Animal Conservation* **14**:47–55.

Lawson, D.J. *et al.* (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications* **9**:3258.

Lee, Y.-S. and Shin, D. (2018). Estimation of the Genetic Substitution Rate of Hanwoo and Holstein Cattle Using Whole Genome Sequencing Data. *Genomics & Informatics* **16**:14–20.

Legesse, G. *et al.* (2017). BOARD-invited review: Quantifying water use in ruminant production. *Journal of Animal Science* **95**:2001–2018.

Leinonen, R. *et al.* (2011). The sequence read archive. *Nucleic Acids Research* **39**:2010–2012.

Lenz, T.L. *et al.* (2016). Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary

Cost of Balancing Selection. *Molecular Biology and Evolution* **33**:2555–2564.

Li, H. *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079.

Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475**:493–496.

Li, H. and Durbin, R. (2009). Making the Leap: Maq to BWA. *Mass Genomics* **25**:1754–1760.

de Lima, R.E. *et al.* (2018). Two sides of a coin: GG genotype of C7 provides protection against fibrosis severity while showing a higher risk for hepatocellular carcinoma in patients with hepatitis C. *Human Immunology* **79**:702–707.

Liu, T. *et al.* (2012). DNAJC25 is downregulated in hepatocellular carcinoma and is a novel tumor suppressor gene. *Oncology Letters* **4**:1274–1280.

Liu, Y. *et al.* (2016). The structural basis of chicken, swine and bovine CD8αα dimers provides insight into the co-evolution with MHC i in endotherm species. *Scientific Reports* **6**:1–11.

Loftus, R.T. *et al.* (1994). Mitochondrial genetic variation in European, African and Indian cattle populations. *Animal Genetics* **25**:265–271.

M. Kimura (1968). Evolutionary Rate at the Molecular Level. *Nature* **217**:624–626.

Ma, F. *et al.* (2002). Molecular cloning of Porimin, a novel cell surface receptor mediating oncotic cell death. *Proceedings of the National Academy of Sciences* **98**:9778–9783.

Ma, L. *et al.* (2019). Symposium review: Genetics, genome-wide association study, and genetic improvement of dairy fertility traits. *Journal of Dairy Science* **102**:3735–3743.

Mabbott, N.A. (2004). The complement system in prion diseases. *Current Opinion in Immunology* **16**:587–593.

Machicado, C. *et al.* (2016). Association of Fasciola hepatica Infection with Liver Fibrosis, Cirrhosis, and Cancer: A Systematic Review. *PLoS Neglected Tropical Diseases* **10**:1–11.

MacHugh, D.E. *et al.* (2001). Genetic evidence for Near-Eastern origins of European cattle. *Nature* **410**:1088–1091.

MacHugh, D.E. *et al.* (2017). Taming the Past: Ancient DNA and the Study of Animal Domestication. *Annual Review of Animal Biosciences* **5**:329–351.

Machugh, D.E. *et al.* (1997). hDomestication and Phylogeography of Taurine and Zebu Cattle.

MacLeod, I.M. *et al.* (2013). Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. *Molecular Biology and Evolution* **30**:2209–2223.

Madrid-Bury, N. *et al.* (2005). Relationship between non-return rate and chromatin condensation of deep frozen bull spermatozoa. *Theriogenology* **64**:232–241.

Makina, S.O. *et al.* (2015). Genome-wide scan for selection signatures in six cattle breeds in South Africa. *Genetics Selection Evolution* **47**.

Malaspinas, A.S. *et al.* (2016). A genomic history of Aboriginal Australia. *Nature* **538**:207–214.

Malomane, D.K. *et al.* (2018). Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. *BMC Genomics* **19**:22.

Mannen, H. *et al.* (2004). Independent mitochondrial origin and historical genetic differentiation in North Eastern Asian cattle. *Molecular Phylogenetics and Evolution* **32**:539–544.

Marchi, N. *et al.* (2021). Demographic inference. *Current Biology* **31**:R276–R279.

Mariasegaram, M. *et al.* (2007). The slick hair coat locus maps to chromosome 20 in Senepol-derived cattle. *Animal Genetics* **38**:54–59.

Márquez, G.C. *et al.* (2010). Genetic diversity and population structure of American Red Angus cattle. *Journal of Animal Science* **88**:59–68.

Marras, G. *et al.* (2015). Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. *Animal Genetics* **46**:110–121.

Martínez, A.M. *et al.* (2012). Genetic Footprints of Iberian Cattle in America 500 Years after the Arrival of Columbus. *PLoS ONE* **7**:1–13.

Mastrangelo, S. *et al.* (2020). Refining the genetic structure and relationships of European cattle breeds through meta-analysis of worldwide genomic SNP data, focusing on Italian cattle. *Scientific Reports* **10**:1–13.

Matukumalli, L.K. *et al.* (2009). Development and characterization of a high density SNP genotyping assay for cattle. *PloS one* **4**:e5350.

Maynard Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research* **23**:23–35.

Mazet, O. *et al.* (2016). On the importance of being structured: instantaneous coalescence rates and human evolution--lessons for ancestral population size inference? *Heredity* **116**:362–71.

Mbole-Kariuki, M.N. *et al.* (2014). Genome-wide analysis reveals the ancient and recent admixture history of East African Shorthorn Zebu from Western Kenya. *Heredity* **113**:297–305.

McDonald, J.H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**:652–654.

McKay, J.K. and Latta, R.G. (2002). Adaptive population divergence: Markers, QTL and traits. *Trends in Ecology and Evolution*:285–291.

McKenna, A. *et al.* (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**:1297–1303.

McLaren, W. *et al.* (2016). The Ensembl Variant Effect Predictor. *Genome Biology* **17**:122.

McTavish, E.J. *et al.* (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.

McTavish, E.J. and Hillis, D.M. (2015). How do SNP ascertainment schemes and population demographics affect inferences about population history? *BMC Genomics* **16**:1–13.

McVean, G.A.T. and Cardin, N.J. (2005). Approximating the coalescent with recombination. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **360**:1387–93.

Meadows, J.R.S. *et al.* (2007). Five Ovine Mitochondrial Lineages Identified From Sheep Breeds of the Near East. *Genetics* **175**:1371–1379.

Medugorac, I. *et al.* (2012). Bovine Polledness – An Autosomal Dominant Trait with Allelic Heterogeneity Zhao, S. (ed.). *PLoS ONE* **7**:e39477.

Melchior, M.B. *et al.* (2011). Active surveillance for scrapie in the Netherlands. *Tijdschr Diergeneeskd* **136**:84–93.

Menck, C.F.M. and Munford, V. (2014). DNA repair diseases: What do they tell us about cancer and aging? *Genetics and Molecular Biology* **37**:220–233.

Mercer, R.C.C. *et al.* (2018). Prion Diseases. In: *The Molecular and Cellular Basis of Neurodegenerative Diseases: Underlying Mechanisms*. Cham: Springer International Publishing, pp. 23–56.

Messer, P.W. and Petrov, D.A. (2013). Frequent adaptation and the McDonald-Kreitman test. *Proceedings of the National Academy of Sciences of the United States of America* **110**:8615–8620.

Met Office (2017). UKCP09 Gridded Observation Datasets.

Mignon-Grasteau, S. *et al.* (2005). Genetics of adaptation and domestication in livestock. *Livestock Production Science* **93**:3–14.

Mikko, S. and Andersson, L. (1995). Low major histocompatibility complex class II diversity in European and North American moose. *Proceedings of the National Academy of Sciences of the United States of America* **92**:4259–4263.

Miretti, M.M. *et al.* (2004). Predominant African-derived mtDNA in Caribbean and Brazilian creole cattle is also found in Spanish cattle (Bos taurus). *Journal of Heredity* **95**:450–453.

Mosier, A. *et al.* (1998). Closing the global N2O budget: Nitrous oxide emissions through the agricultural nitrogen cycle: OECD/IPCC/IEA phase II development of IPCC guidelines for national greenhouse gas inventory methodology. In: *Nutrient Cycling in Agroecosystems*. Springer Netherlands, pp. 225–248.

Myers, S.S. *et al.* (2017). Climate Change and Global Food Systems: Potential Impacts on Food Security and Undernutrition. *Annual Review of Public Health* **38**:259–277.

Nadeau, S. *et al.* (2016). The challenge of separating signatures of local adaptation from those of isolation by distance and colonization history: The case of two white pines. *Ecology and Evolution* **6**:8649–8664.

Naji, M.M. *et al.* (2021). Investigation of ancestral alleles in the Bovinae subfamily. *BMC Genomics* **22**.

Narasimhan, V. *et al.* (2016). BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**:1749–1751.

Nardone, A. *et al.* (2010). Effects of climate changes on animal production and sustainability of livestock systems. *Livestock Science* **130**:57–69.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annual Review of Genetics* **39**:197–218.

Nielsen, R. (2004). Population genetic analysis of ascertained SNP data. *Human genomics* **1**:218–24.

Nourani, H. and Karimi, I. (2007). *Hepatocellular Carcinoma in a Sheep*.

O'Connor, S.F. *et al.* (1997). Genetic Effects on Beef Tenderness in Bos indicus Composite and Bos taurus Cattle. *Journal of Animal Science* **75**:1822–1830.

O'Mara, F.P. (2012). The role of grasslands in food security and climate change. *Annals of Botany* **110**:1263–1270.

Ohta, T. and Kimura, M. (1971). Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics* **68**:571–580.

Olschewsky, A. and Hinrichs, D. (2021). An overview of the use of genotyping techniques for assessing genetic diversity in local farm animal breeds. *Animals* **11**:e11072016.

Orozco-terWengel, P. *et al.* (2015). Revisiting demographic processes in cattle with genome-wide population genetic analysis. *Frontiers in Genetics* **6**:1–15.

Orozco-terWengel, P. (2016). The devil is in the details: the effect of population structure on demographic inference. *Heredity* **116**:349–350.

Orozco-terWengel, P.A. and Bruford, M.W. (2014). Mixed signals from hybrid genomes. *Molecular Ecology* **23**:3941–3943.

Paital, B. *et al.* (2016). Longevity of animals under reactive oxygen species stress and disease susceptibility due to global warming. *World journal of biological chemistry* **7**:110–27.

Park, S.D.E. *et al.* (2015). Genome sequencing of the extinct Eurasian wild aurochs, Bos primigenius, illuminates the phylogeography and evolution of cattle. *Genome Biology* **16**:234.

Paşaniuc, B. *et al.* (2009). Inference of locus-specific ancestry in closely related populations. *Bioinformatics* **25**:213–221.

Pasqui, M. and Di Giuseppe, E. (2019). Climate change, future warming, and adaptation in Europe. *Animal Frontiers* **9**:6–11.

Pellecchia, M. *et al.* (2007). The mystery of Etruscan origins: Novel clues from Bos taurus mitochondrial DNA. *Proceedings of the Royal Society B: Biological Sciences* **274**:1175–1179.

Pérez-Pardal, L. *et al.* (2010). Multiple paternal origins of domestic cattle revealed by Y-specific interspersed multilocus microsatellites. *Heredity* **105**:511–519.

Pickrell, J.K. and Pritchard, J.K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics* **8**:e1002967.

Piertney, S.B. and Oliver, M.K. (2006). The evolutionary ecology of the major histocompatibility complex. *Heredity* **96**:7–21.

Polley, H.W. *et al.* (2013). Climate change and North American rangelands: Trends, projections,

and implications. *Rangeland Ecology and Management* **66**:493–511.

Polsky, L. and von Keyserlingk, M.A.G. (2017). Invited review: Effects of heat stress on dairy cattle welfare. *Journal of Dairy Science* **100**:8645–8657.

Porto-Neto, L.R. *et al.* (2014). The genetic architecture of climatic adaptation of tropical cattle. *PLoS ONE* **9**:e113284.

Pritchard, J.K. and Di Rienzo, A. (2010). Adaptation - Not by sweeps alone. *Nature Reviews Genetics* **11**:665–667.

Pritchard, J.K. *et al.* (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**:945–959.

Pulina, G. *et al.* (2017). Sustainable ruminant production to help feed the planet. *Italian Journal of Animal Science* **16**:140–171.

Purcell, S. *et al.* (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**:559–575.

Purfield, D.C. *et al.* (2012). Runs of homozygosity and population history in cattle. *BMC Genetics* **13**:e14712156.

Qanbari, S. and Simianer, H. (2014). Mapping signatures of positive selection in the genome of livestock. *Livestock Science* **166**:133–143.

QGIS Development Team (2019). QGIS Geographic Information System. QGIS Association. http://www.qgis.org.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Radwan, J. *et al.* (2010). Does reduced MHC diversity decrease viability of vertebrate populations? *Biological Conservation* **143**:537–544.

Randhawa, I.A. *et al.* (2014). Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genetics* **15**:34.

Rare Breeds Survival Trust (2019). *Breed Watchlist 2019*. Rare Breed Survival Trust. https://www.rbst.org.uk/rbst-watchlist.

Reed, D.H. and Frankham, R. (2003). Correlation between fitness and genetic diversity. *Conservation Biology* **17**:230–237.

Reich, D. *et al.* (2009). Reconstructing Indian population history. *Nature* **461**:489–494.

Rellstab, C. *et al.* (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology* **24**:4348–4370.

Robinson, J.A. *et al.* (2021). Genome-wide diversity in the California condor tracks its prehistoric abundance and decline. *Current Biology* **31**:2939-2946.e5.

Robinson, J.A. *et al.* (2016). Genomic Flatlining in the Endangered Island Fox. *Current Biology* **26**:1183–1189.

Rodero, A. *et al.* (1992). Primitive andalusian livestock and their implications in the discovery of america. *Archivos de Zootecnia* **41**:383–400.

Rojas-Downing, M.M. *et al.* (2017). Climate change and livestock: Impacts, adaptation, and mitigation. *Climate Risk Management* **16**:145–163.

Roque, B.M. and Duarte, T.L. (2020). Red seaweed (Asparagopsis taxiformis) supplementation reduces enteric methane by over 80 percent in beef steers. :1–20.

Rosegrant, M.W. *et al.* (2002). *Global Water Outlook to 2025: Averting an Impending Crisis*. A 2020 vision for food, agriculture, and the environment initiative. Washington, DC: IFPRI and IWMI.

Rosen, B. *et al.* (2018). *Modernizing the Bovine Reference Genome Assembly*.

Rougemont, Q. *et al.* (2016). Reconstructing the demographic history of divergence between European river and brook lampreys using approximate Bayesian computations. *PeerJ* **4**:e1910.

Rowland, C.S. *et al.* (2017). *Land Cover Map 2015 (25m Raster, GB)*. NERC Environmental Information Data Centre.

Rubin, C.J. *et al.* (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**:587–591.

Russell, N. (2007). *Like Engend'ring Like. Heredity and Animal Breeding in Early Modern England*. Cambridge: Cambridge University Press.

Ryeland Flock Book Society (2019). *Ryeland Sheep*. http://www.ryelandfbs.com/index.html.

Saatchi, M. *et al.* (2011). Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genetics Selection Evolution* **43**.

Sabeti, P.C. *et al.* (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**:832–837.

Sabeti, P.C. *et al.* (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**:913–918.

Sabeti, P.C. *et al.* (2010). Positive natural selection in the human lineage. *Science* **312**:665–74.

Saravanan, K.A. *et al.* (2020). Selection signatures in livestock genome: A review of concepts, approaches and applications. *Livestock Science* **241**:104257.

Sayers, E.W. *et al.* (2019). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **47**:D23–D28.

Scheet, P. and Stephens, M. (2006). *A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase*.

Schibler, J. *et al.* (2014). Incorporation of aurochs into a cattle herd in Neolithic Europe: Single event or breeding? *Scientific Reports* **4**:8–13.

Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics* **46**:919–925.

Schraiber, J.G. and Akey, J.M. (2015). Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics* **16**:727–740.

Schrider, D.R. *et al.* (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics* **204**:1207–1223.

Sevane, N. *et al.* (2019). Genome-wide differential DNA methylation in tropically adapted Creole cattle and their Iberian ancestors. *Animal Genetics* **50**:15–26.

Shamimuzzaman, M. *et al.* (2020). Bovine Genome Database: New annotation tools for a new reference genome. *Nucleic Acids Research* **48**:D676–D681.

Shu, L. *et al.* (2014). *Non-Classical Major Histocompatibility Complex Class Makes a Crucial Contribution to Reproduction in the Dairy Cow*. J Reprod Dev.

Sim, N.L. *et al.* (2012). SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research* **40**:452–457.

Smith, P.G. and Bradley, R. (2003). Bovine spongiform encephalopathy (BSE) and its epidemiology. *British Medical Bulletin* **66**:185–198.

Sonesson, U. *et al.* (2009). Greenhouse Gas Emissions in Animal Feed Production. *Decision support for climate certification*:28.

Spielman, D. *et al.* (2004). Most species are not driven to extinction before genetic factors impact them. *Proceedings of the National Academy of Sciences of the United States of America* **101**:15261–15264.

Sponenberg, D.P. *et al.* (2019). Conservation strategies for local breed biodiversity. *Diversity* **11**:177.

Steffen, D. (1997). Genetic causes of bull infertility. *The Veterinary clinics of North America. Food animal practice* **13**:243–253.

Steinfeld, H. *et al.* (2006). Livestock production systems in developing countries: Status, drivers, trends. *OIE Revue Scientifique et Technique* **25**:505–516.

Stinchcombe, J.R. and Hoekstra, H.E. (2008). Combining population genomics and quantitative genetics: Finding the genes underlying ecologically important traits. *Heredity* **100**:158–170.

Stock, F. and Gifford-Gonzalez, D. (2013). Genetics and African Cattle Domestication. *African Archaeological Review 2013 30:1* **30**:51–72.

Stoffel, M.A. *et al.* (2020). Genetic architecture and lifetime dynamics of inbreeding depression in a wild mammal. *bioRxiv*.

Strandén, I. *et al.* (2019). Genomic selection strategies for breeding adaptation and production in dairy cattle under climate change. *Heredity* **123**:307–317.

Stronen, A. V. *et al.* (2019). Genomic analyses suggest adaptive differentiation of northern European native cattle breeds. *Evolutionary Applications* **12**:1096–1113.

Stucki, S. *et al.* (2017). High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources* **17**:1072–1089.

Sunnåker, M. *et al.* (2013). Approximate Bayesian Computation Wodak, S. (ed.). *PLoS Computational Biology* **9**:e1002803.

Sved, J.A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* **2**:125–141.

Sved, J.A. and Feldman, M.W. (1973). Correlation and probability methods for one and two loci. *Theoretical Population Biology* **4**:129–132.

Szpiech, Z.A. and Hernandez, R.D. (2014). Selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution* **31**:2824–2827.

Taberlet, P. *et al.* (2011). Conservation genetics of cattle, sheep, and goats. *Comptes rendus biologies* **334**:247–254.

Tajima, F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123**:585–595.

Tapio, M. *et al.* (2006). Sheep Mitochondrial DNA Variation in European, Caucasian, and Central Asian Areas. *Molecular Biology and Evolution* **23**:1776–1783.

The Bovine HapMap Consortium *et al.* (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **324**:528–532.

Thornton, P.K. (2010). Livestock production: Recent trends, future prospects. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**:2853–2867.

Thornton, P.K. *et al.* (2009). The impacts of climate change on livestock and livestock systems in developing countries: A review of what we know and what we need to know. *Agricultural Systems* **101**:113–127.

Towers, J. *et al.* (2017). An isotopic investigation into the origins and husbandry of Mid-Late Bronze Age cattle from Grimes Graves, Norfolk. *Journal of Archaeological Science: Reports* **15**:59–72.

Townsend, S.J. *et al.* (2005). PrP genotypes of rare breeds of sheep in Great Britain. *Veterinary Record* **156**:131–134.

UN (2019). *World Population Prospects 2019*. United Nations, Department of Economic and Social Affairs, Population Division.

UNHCR (2018). *The Sustainable Development Goals and Addressing Statelessness*. UN High Commissioner for Refugees.

Upadhyay, M.R. *et al.* (2017). Genetic origin, admixture and population history of aurochs (Bos primigenius) and primitive European cattle. *Heredity* **118**:169–176.

Utsunomiya, Y.T. *et al.* (2013). Detecting Loci under Recent Positive Selection in Dairy and Beef Cattle by Combining Different Genome-Wide Scan Methods. *PLoS ONE* **8**:1–11.

Utsunomiya, Y.T. *et al.* (2019). Genomic clues of the evolutionary history of Bos indicus cattle. *Animal Genetics* **50**:557–568.

Utsunomiya, Y.T. *et al.* (2015). Genomic data as the 'hitchhiker's guide' to cattle adaptation: Tracking the milestones of past selection in the bovine genome. *Frontiers in Genetics* **5**:36.

Valente-Campos, S. *et al.* (2019). Critical issues and alternatives for the establishment of chemical water quality criteria for livestock. *Regulatory Toxicology and Pharmacology* **104**:108–114.

Verdugo, M.P. *et al.* (2019). Ancient cattle genomics, origins, and rapid turnover in the Fertile Crescent. *Science* **365**:173–176.

Vigne, J. and Helmer, D. (2007). Was Milk a 'Secondary Product' in the Old World Neolithisation Process? Its Role in the Domestication of Cattle, Sheep, and Goats. *Anthropozoologica* **42**:9–42.

Villalobos Cortés, A. *et al.* (2009). History of Panama bovines and their relationships with other Iberoamerican popultions. *Archivos de Zootecnia* **118**:169–176.

de Villemereuil, P. and Gaggiotti, O.E. (2015). A new FST-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution* **6**:1248–1258.

Visscher, P.M. *et al.* (2001). A viable herd of genetically uniform cattle. *Nature* **409**:303–303.

Vitti, J.J. *et al.* (2013). Detecting Natural Selection in Genomic Data. *Annual Review of Genetics* **47**:97–120.

Voight, B.F. *et al.* (2006). A map of recent positive selection in the human genome. *PLoS Biology* **4**:0446–0458.

Wang, E.T. *et al.* (2006). Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proceedings of the National Academy of Sciences of the United States of America* **103**:135–140.

Wang, K. *et al.* (2020). Tracking human population structure through time from whole genome sequences. *PLoS Genetics* **16**:1–24.

Wegmann, D. and Excoffier, L. (2010). Bayesian inference of the demographic history of chimpanzees. *Molecular Biology and Evolution* **27**:1425–1435.

Wegmann, D. *et al.* (2010). ABCtoolbox: A versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* **11**.

Weir, B.S. and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**:1358.

Weisdorf, J.L. (2005). From Foraging To Farming: Explaining the Neolithic Revolution. *Journal of Economic surveys* **19**:561–586.

Wilkins, A.S. *et al.* (2014). The 'domestication syndrome' in mammals: A unified explanation based

on neural crest cell behavior and genetics. *Genetics* **197**:795–808.

Will, R.. *et al.* (1996). A new variant of Creutzfeldt-Jakob disease in the UK. *The Lancet* **347**:921–925.

Willham, R.L. (1982). Genetic improvement of beef cattle in the United States: cattle, people and their interaction. *Journal of animal science* **54**:659–666.

Williams, J.L. *et al.* (2016). Inbreeding and purging at the genomic Level: The Chillingham cattle reveal extensive, non-random SNP heterozygosity. *Animal Genetics* **47**:19–27.

Willoughby, J.R. *et al.* (2015). The reduction of genetic diversity in threatened vertebrates and new recommendations regarding IUCN conservation rankings. *Biological Conservation* **191**:495–503.

Wolf, J.B.W. and Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics* **18**:87–100.

Wright, S. (1939). the Distribution of Self-Sterility Alleles in Populations. *Genetics* **24**:538–552.

Wright, S. (1949). The genetical structure of populations. *Annals of eugenics* **15**:323–354.

Würsch, D. *et al.* (2016). CD38 Expression in a Subset of Memory T Cells Is Independent of Cell Cycling as a Correlate of HIV Disease Progression. *Disease Markers* **2016**:e9510756.

Xia, J. *et al.* (2015). Association between liver fluke infection and hepatobiliary pathological changes: A systematic review and meta-analysis. *PLoS ONE* **10**:1–19.

Xia, X.T. *et al.* (2021). Mitochondrial genomes from modern and ancient Turano-Mongolian cattle reveal an ancient diversity of taurine maternal lineages in East Asia. *Heredity* **126**:1000–1008.

Xu, X. *et al.* (2021). Global greenhouse gas emissions from animal-based foods are twice those of plant-based foods. *Nature Food* **2**:724–732.

Yaro, M. *et al.* (2017). Molecular identification of livestock breeds: A tool for modern conservation biology. *Biological Reviews* **92**:993–1010.

Youatt, W. (1837). *Sheep; Their Breeds, Management and Diseases.* London: Baldwin and Cradock.

Yue, X.P. *et al.* (2015). A limited number of Y chromosome lineages is present in North American Holsteins. *Journal of Dairy Science* **98**:2738–2745.

Yurchenko, A. *et al.* (2018). Genome-wide genotyping uncovers genetic profiles and history of the Russian cattle breeds. *Heredity* **120**:125–137.

Zampiga, M. *et al.* (2018). Application of omics technologies for a deeper insight into quali-quantitative production traits in broiler chickens: A review. *Journal of Animal Science and Biotechnology* **9**:1–18.

Zeder, M.A. (2008). Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proceedings of the National Academy of Sciences* **105**:11597–11604.

Zeder, M.A. (2017). Out of the fertile crescent: The dispersal of domestic livestock through Europe and Africa. In: *Human Dispersal and Species Movement: From Prehistory to the Present*. pp. 261–303.

Zeuner, F. (1963). *A History of Domestic Animals*. Hutchinson, London.

Zhang, H. *et al.* (2013). Morphological and genetic evidence for early Holocene cattle management in northeastern China. *Nature Communications* **4**:1–7.

Zhang, S. *et al.* (2021). Genetic Differentiation of Reintroduced Père David's Deer (Elaphurus davidianus) Based on Population Genomics Analysis. *Frontiers in Genetics* **12**:e705337.

Zorc, M. *et al.* (2019). The new bovine reference genome assembly provides new insight into genomic organization of the bovine major histocompatibility complex. *Journal of Central European Agriculture* **20**:1111–1115.