

Genomic variation across European cattle: contribution of gene flow

Maulik Upadhyay

Thesis committee

Promotor:

Prof. Dr M.A.M. Groenen
Professor of Animal Breeding and Genomics
Wageningen University & Research

Co-promotors:

Dr R.P.M.A. Crooijmans
Assistant Professor, Animal Breeding and Genomics
Wageningen University & Research

Prof. Dr G. Andersson
Professor, Department of Animal Breeding and Genetics
Swedish University of Agricultural Sciences, Sweden

Dr S. Mikko
Assistant Professor, Department of Animal Breeding and Genetics
Swedish University of Agricultural Sciences, Sweden

Other members (assessment committee)

Prof. Dr C.H.J. van Oers, Wageningen University & Research
Dr E. Jonas, Swedish University of Agricultural Sciences, Sweden
Prof. Dr M. Boichard, National Institute of Agricultural Research (INRA), France
Dr D.K. Aanen, Wageningen University & Research

The research presented in this doctoral thesis was conducted under the joint auspices of the Swedish University of Agricultural Sciences and the Graduate School Wageningen Institute of Animal Sciences of Wageningen University and is part of the Erasmus Mundus Joint Doctorate program “EGS-ABG”.

Genomic variation across European cattle: contribution of gene flow

Maulik Upadhyay



ACTA UNIVERSITATIS AGRICULTURAE SUECIAE
DOCTORAL THESIS No 2019:16

Thesis

submitted in fulfilment of the requirements for the joint degree of doctor between

Swedish University of Agricultural Sciences

by the authority of the Board of the Faculty of Veterinary Medicine and

Animal Science and

Wageningen University

by the authority of the Rector Magnificus, Prof. Dr. A.P.J. Mol,

in the presence of the

Thesis Committee appointed by the Academic Board of Wageningen University and

the Board of the Faculty of Veterinary Medicine and Animal Science at

the Swedish University of Agricultural Sciences

to be defended in public

on Tuesday 12 March, 2019

at 4.00 p.m. in the Aula of Wageningen University.

Maulik Upadhyay

Genomic variation across European cattle: contribution of gene flow

161 pages.

Joint PhD thesis, Swedish University of Agricultural Sciences, Uppsala, Sweden
and Wageningen University, the Netherlands (2019)

With references, with summaries in English, Dutch and Swedish.

ISBN (print version) 978-91-7760-350-4

ISBN (electronic version) 978-91-7760-351-1

ISSN 1652-6880

ISBN 978-94-6343-420-1

DOI <https://doi.org/10.18174/469250>

Abstract

Upadhyay M.R. (2019). Genomic variation across European cattle: contribution of gene-flow. Joint Ph.D. thesis, between Swedish University of Agricultural Sciences, Sweden and Wageningen University and Research, the Netherlands.

European cattle display vast phenotypic diversity which can be attributed to genomic variation such as single nucleotide polymorphisms (SNPs) and structural variations (SVs). The distribution of these genomic variations in a population is heavily influenced by different population genomic forces. In this thesis, I used genome-wide SNPs to characterize genomic variation and admixture across different European cattle populations. Broadly, I show the difference in the domestication histories for north-western and southern European cattle. I argue that this difference can be attributed to a differential pattern of genomic admixture involving wild local aurochs and zebu cattle. Genomic admixture analysis revealed shared ancestry between Balkan and Italian cattle (BAI) breeds, and zebu cattle. Moreover, I also show that southern European cattle breeds displayed shared ancestry with African taurine cattle. Using linked SNP based approaches, I inferred a common origin of the African taurine and zebu cattle ancestry in BAI cattle breeds. Furthermore, I also characterized the genomic diversity and structure in European cattle populations. I show that, on average, nucleotide diversity is higher in southern European cattle than western European (British and commercial) cattle. However, some of these southern European cattle breeds such as Romagnola and Maltese appeared to have undergone a recent bottleneck. On the other hand, Swedish native cattle breeds like Swedish Mountain cattle, despite recorded bottleneck in the past, still display significant genomic diversity. However, southern Swedish cattle breeds like Väneko and Ringamålako requires attention for conservation management as these breeds display lowest genetic diversity among all the Swedish cattle breeds. To understand the patterns of genomic variations comprehensively, I also characterized the structural variations (SVs) in the genome of European cattle. I inferred the influence of demographic changes in the distribution of SVs in the cattle genome. In addition, I also identified an SV CNV overlapping the *KIT* gene in English Longhorn cattle which has previously been associated with color-sidedness. Finally, using whole genome sequencing data, I identified various protein-coding genes and regulatory elements encompassing SVs which represents valuable resources for future studies aimed at finding the association between physiological processes and SVs in cattle.

Contents

	Page
Abstract	v
Chapter 1 General introduction	1
Chapter 2 Genetic origin, admixture and population history of aurochs (<i>Bos primigenius</i>) and primitive European cattle	19
Chapter 3 Deciphering the pattern of genetic admixture and diversity in southern European cattle using Genome-wide SNPs	36
Chapter 4 Genomic relatedness and diversity of Swedish native cattle breeds	54
Chapter 5 Distribution and functionality of copy number variation across European cattle populations	69
Chapter 6 Comparative evaluation of structural variations in taurine and indicine cattle using individual whole genome sequences	87
Chapter 7 General discussion	102
References	115
Summary	134
Samenvatting	137
Sammanfattning	140
Acknowledgements	143
Curriculum vitae	145
Training and Supervision Plan	148
Data availability and supplementary material	150
Colophon	152

Chapter 1

General introduction

1.1 Evolution of Bovinae sub-family

The mammalian sub-family Bovinae comprises several diverse species (Figure 1.1), some of which are culturally and economically very important throughout the world. The sub-family is further classified into the three major tribes: Tragelaphini, Boselaphini, and Bovini. While the first two tribes comprise of spiral, four-horned, large ox-like antelope, the Bovini tribe comprises almost all domestic and wild bovine species. The first split within the Bovini tribe occurred somewhere between 5-10 million years ago (MYA) when the subtribe Bubalina (*Bubalus* and *Syncerus spp.*) diverged from the subtribe Bovina (*Bos* and *Bison spp.*) (Hartl et al., 1988; L. Janecek et al., 1996; Ritz et al., 2000). These two subtribes have consistently shown to be forming dichotomous groups and no evidence of viable hybrid offspring has been reported from the mating involving these two subtribes (Hartl et al., 1988; L. Janecek et al., 1996; Ritz et al., 2000; Hassanin and Ropiquet, 2004; MacEachern et al., 2009; Dorian J. Garrick and Ruvinsky, 2014). Within the subtribe Bovina, divergence events involving the remaining species appeared to have occurred recently, in the last 2 MYA. As a result, the species within this sub-tribe can still produce viable offspring indicating incomplete speciation. In fact, the introgression events involving domestic cattle in the yak (*Bos Grunniens*) and wisent (*Bison Bonasus*) lineage have already been inferred using whole genome sequencing data (Soubrier et al., 2016; Medugorac et al., 2017).

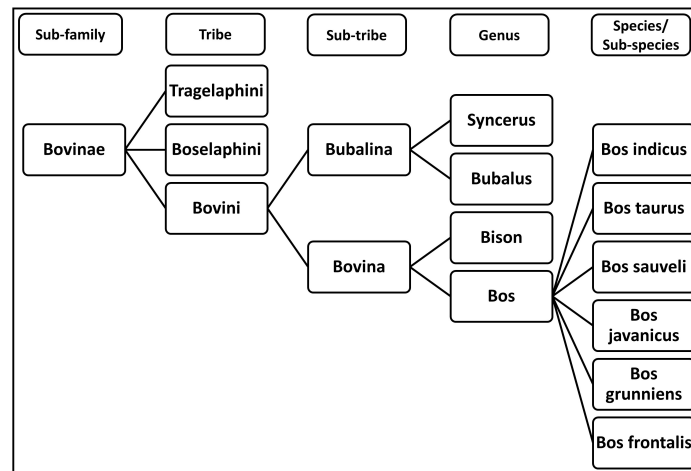


Figure 1.1: Taxonomic classification of sub-family Bovinae

Mitochondrial DNA (mtDNA) and genome-wide SNP based analyses have estimated the divergence date between the two most economically important *Bos* sub-species, *Bos indicus*, and *Bos taurus*, somewhere between 0.117 to 0.275 MYA (Loftus et al., 1994; Bradley et al., 1996; Gautier et al., 2016). The majority of the world cattle populations can be categorized under these two *Bos* sub-species with cross-breeding practices between these sub-species being widely prevalent in many parts of the world such as North America and Africa. The major morphological differences between these two sub-species are the presence of a thoracic hump, floppy rather than upright ears, and a large dewlap in *Bos indicus*. Both sub-species also display identical karyotypes with 29 autosomal pairs and a pair of sex chromosomes (X/Y). The Y-chromosome, however, is sub-metacentric in taurine and acrocentric in zebu, respectively (Kieffer and Cartwright, 1968; Jorge,

1974). Both these sub-species also display differences in physiological adaptation; while indicine cattle are very well adapted to harsh environmental conditions, most taurine cattle have been intensively selected for production related traits.

1.2 Initiation of domestication and early dispersion of cattle in Europe and Africa

The geographic origin and number of domestication events of cattle are arguably one of the most debated questions among bovine geneticists. Evidence based on archaeological and molecular data, points towards at least two centres of cattle domestication: domestication of *Bos primigenius namadicus* (Indian aurochs) in the Indus valley and domestication of *Bos primigenius primigenius* (European aurochs) in the Near East (Loftus et al., 1994; Bradley et al., 1996). The independent domestication of African aurochs has also been proposed (Grigson, 1991). However, a recent study has refuted this hypothesis (Decker et al., 2014).

The taurine lineage might have been domesticated first ~10,000 years before present (YBP) in the Near East, most likely near the regions of the upper Euphrates basin and adjacent to the uppermost Tigris basin (Helmer et al., 2005). Based on approximate Bayesian computation approach on mtDNA of ancient and modern cattle samples, it has been estimated that only about 80 female aurochs were initially domesticated (Bollongino et al., 2012; Scheu et al., 2015). Like other successful innovations, agriculture and animal husbandry also dispersed to other human populations, which can partly be attributed to migrations of early Neolithic farmers. Based on the archaeological and molecular evidence, it is possible to reconstruct the demographic events leading up to the dispersion of domestic cattle throughout Europe. Following domestication, it has been suggested (Martins et al., 2015; Hofmanová et al., 2016) that Neolithic farmers along with their livestock took at least two distinct routes (Figure 1.2) to reach mainland Europe: the Mediterranean Sea route and the Danube river route. Following these migrations, the earliest evidence of domestic cattle in Europe are reported in the form of cattle bones found at a Neolithic site in Greece which are dated ~8,500 YBP (Conolly et al., 2012). Evidence also suggests that, via the Mediterranean Sea route, farming was introduced in Corsica, the southwest of France and in eastern Spain between ~7,700–7,600 and ~7,400–7,300 YBP, respectively (De Lagrán, 2014). Via the Danube river route, domestic cattle reached central Europe and Northern Europe ~7,500 YBP and ~6,500 YBP respectively (Tresset, 2003). Indeed, studies involving Isotope analyses of organic residues of the major milk fatty acids preserved in archaeological pottery have indicated the use of milk products by European farmers from as early as ~8,000 YBP (Salque et al., 2013).

During the early dispersion of domestic taurine, the wild population of ancestral European aurochs was still prevalent across mainland Europe. In fact, the last aurochs died at the beginning of the 17th Century (Kędzierska 1959; 1965 cited by van Vuure 2005). At its peak, the aurochs were distributed all over Eurasia; the distribution ranged from the Atlantic coast of Europe to the Pacific coast of China (Wright, 2013). Aurochs remains, however, have not yet been found

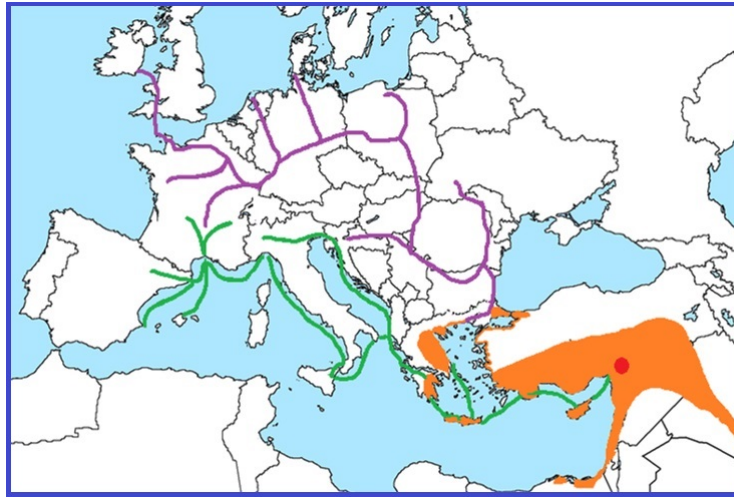


Figure 1.2: Representation of migration routes of Neolithic farmers. Red colour represents the center of domestication; the green line represents the Mediterranean Sea route, while the violet line represents the Danube river route. The figure is adapted from Feliuss et al., (2014). (map outline from D-maps.com-https://d-maps.com/carte.php?num_car=2232&lang=en)

in Ireland, making West Iberia as the westernmost range of its distribution (Wright, 2013). Due to the long history of shared geography between aurochs and domestic cattle, the possibility of inter-crossing between them cannot be ruled out. In fact, several studies have investigated this hypothesis of post-domestication contact between domestic cattle and aurochs, some of which are discussed elsewhere in the thesis.

The theory of independent cattle domestication of the now-extinct African aurochs (*Bos primigenius africanus*) is highly disputed. The supporters of the theory often point out the prevalence of mitochondrial T1 haplotypes (Bradley et al., 1996; Edwards et al., 2004) in African cattle and osteological evidence found in the western Egyptian desert dating from ~10,000 YBP (Wendorf et al., 1989) as proofs for backing their claim. Zooarchaeologists have cast their doubt on the origin of osteological evidence, and analysis of complete mtDNA sequences has shown that the T1 mtDNA haplotype is also found among Southwest Asian cattle, albeit at low frequency (Troy et al., 2001). Moreover, it has also been shown that the T1 haplogroup node is only one mutation (np 16113) away from the common mtDNA T3 haplotypes of European taurine, and hence, a near Eastern origin is very likely (Achilli et al., 2009). Uncontroversial dates for the arrival of domestic taurine has been estimated from ~7500 YBP; based on archaeological evidence, it has been suggested that it appeared first in the region around the eastern Sahara (Gifford-Gonzalez and Hanotte, 2011). Archaeological and pictorial evidence also suggest that humpless *Bos taurus* were among the first cattle to appear on the African continent, which later got replaced or admixed by the arrival of zebu cattle. Two waves of zebu arrival in Africa have been proposed: the first wave of zebu arrival is associated with the development of Swahili-Arab civilization that started taking its root from the 7th century AD, while the second wave of zebu cattle expansion is associated with the rinderpest epidemics of the 19th century. Therefore, modern African cattle are mosaics of European taurine and zebu ancestry, though breeds like N'Dama and Mutarin have a unique genetic component which has been hypothesized as a legacy of African aurochs

(Decker et al., 2014).

1.3 European cattle diversity: Domestication to Modern times

1.3.1 Cattle genetic diversity from Neolithic to Roman era

The process of domestication initiated a symbiotic human-animal relationship in which humans started providing food and shelter to livestock in exchange for animal products and services such as fur, food, and protection. The process also allowed the transition of human society from being hunter-gatherers to settled farmers. Gradually, humans started the process of selective breeding of livestock to fulfill their specific needs. Many generations of this human-controlled livestock breeding and adaptation in their respective habitat greatly influenced behavioral and physiological traits of livestock.

The shift in livestock traits, from their ancestral forms to the more derived forms as we see today, occurred gradually as some of the traits that were desirable in wild cattle became a hindrance in domestic habitats. For instance, long horns in the ancestral bovids protected potential predators to some extent, while in the domestic setting long horns are redundant and undesirable as it made the task of handling livestock difficult. Therefore, short-horned cattle emerged somewhere in Mesopotamia in the early Bronze Age and they gradually replaced long-horned cattle in Europe from 5000 BP onward (Epstein, 1971). In the late Bronze age, short-horned cattle were widely distributed across central and Northern Europe while long-horned cattle were more common in many parts of the Mediterranean area as well in the region of today's Hungary (Bórkönyi et al., 1974; Mason, 1984).

Though it is likely that breeding schemes might have existed in ancient times, the first detailed contemporary account of animal husbandry and knowledge-based selective breeding comes from ancient Roman literature. In his classic book "History of Animals" the Greek philosopher and scientist Aristotle gave accounts of large size cattle roaming about in rich pasturelands of Epirus (Balme, 1965). Skeletal remains recovered in Epirus also indicated that between 7th and 8th century BC, the region was inhabited by large size cattle with wither heights ranging from 115 to 135 cm (Kron, 2002). Large Roman cattle that had large horns and wither heights ranging from 120 to 140 cm, also inhabited the ancient Etruria region. However, soon after the fall of the Roman empire, large cattle also disappeared (Feliu et al., 2014).

1.3.2 Cattle genetic diversity in Middle ages to the present times

During the Middle Ages, the small-sized cattle became prevalent in most parts of Europe. This has been attributed to various factors such as ease in management, poor availability of nutritious diet and castration of the large size bulls. Further, the number of livestock was

greatly affected during the 14th century due to the Great Famine as well as Great Cattle Plague (Bórkönyi et al., 1974;Kron, 2002;Campbell, 2009). Following the disastrous 14th century, cattle population gradually recovered owing to cultural and technological development (Felius et al., 2014). This was also the time, when a grey coat colored long-horned cattle of Podolian origin began replacing the local breeds in several parts of Eastern Europe (Bodo et al., 2004). Two hypotheses (Bórkönyi et al., 1974;Ferdinando and Donato, 2001) have been put forward to explain the origin of Podolian cattle: 1) they arrived from the Podolian steppe of Ukraine where they were kept and bred until the 12th century, 2) they are descendants of large cattle which were kept during the Roman era.

During the 17th and 18th century, knowledge-based animal breeding started taking its root across north-western Europe. Literature related to animal husbandry and breeding became commonly available, partly due to improvement in literacy. Cattle migrations were also important aspects of animal husbandry practices during this time. Dutch cattle, due to their superiority in milk production, were exported to Germany, France, and Britain (Felius et al., 2014). However, still until the first industrial revolution of the late 18th century, the majority of the cattle diversity that existed among European cattle was due to adaptation and selection of local cattle breeds to the local circumstances rather than selection for certain traits which were desired by a broad range of consumers (Felius et al., 2014).

In Britain, the industrial revolution that began in the late 18th century provided impetus to the innovation in the field of agriculture and animal husbandry (Thomas, 2005). To meet the demands for animal-related products such as milk and beef in a growing urban population, the farmers began selecting animals based on their performance in desired production traits. For this selection process to work, the record-keeping of a herd as well as pedigree had to be of prime importance. Therefore, the concept of herdbook was introduced in animal husbandry practices. During the 1760's, the Englishman Robert Bakewell—one of the pioneers in Animal breeding—started improving cattle by selecting cows and bulls based on long horns, early growth, docility and other phenotypes (Stanley, 1995). Many English beef cattle breeds, such as Hereford and Aberdeen-Angus, were developed following the breeding success of English Longhorn cattle (Hall and Clutton-Brock, 1988). In fact, to keep the bloodline pure, dairy breeds such as Jersey and Guernsey were forbidden for cross-breeding and kept isolated from as early as 1789 (Hall and Clutton-Brock, 1988). Following these suits of success in record keeping and breeding objectives to develop systematic breeds, many western European countries adapted these techniques. In the Netherlands, the first herdbooks were established in the late 19th Century. Also, by performing cross-breeding between local cattle populations, breeds like Holstein Friesian (HF) and Meuse-Rhine-Yssel (MRY) were developed in the last two decades of the 19th Century (Felius et al., 2014).

1.3.3 Primitive and traditional cattle breeds of Europe

Although the process of domestication led to a transformation of traits that were seen frequently in the wild ancestors, the modern animal breeding practices (such as selection, herd isolation) accelerated this process of transformation of wild traits and led to the emergence of derived

traits, such as early maturity, polledness, and docility in modern cattle that might have rarely been present in its ancestral form. However, many cattle breeds of Europe still display many ancestral features such as horn shape and size, sexual dimorphism, and aggressive behaviour (van Vuure, 2005). I refer to such cattle breeds as primitive in this thesis throughout, and in the following paragraphs, give an overview of the primitive cattle breeds of Europe (Table 1.1).

Primitive Cattle breeds of Balkan and Italian regions largely fall under the category of Podolic cattle breeds. As I mentioned in the previous section, the origin of a Podolic group of cattle and their diffusion to southern Europe is highly debated among bovine geneticists. This group of cattle along with Busha represents some of the most underdeveloped taurine populations. Apart from displaying common characteristics such as long horns and grey coat colour, these Podolic cattle breeds are some of the hardiest European taurines that can be raised under extensive management (Ferdinando and Donato, 2001; Felius et al., 2014; Di Lorenzo et al., 2018). These cattle are also adapted to a wide range of environments and display high disease resistance (Bartosiewicz, 2011). Some of the sampled Podolian cattle breeds include the following: Romanian grey, Boskarin, Chianina, Maremmana, Podolica, Romagnola, and Marchigiana (Table 1.1). Busha and Maltese are two other cattle breeds that we included in the Balkan and Italian group. Busha is distributed throughout the Balkan peninsula including Bulgaria and Greece. This group of cattle is characterized by small height, red to grey coat colour and small horns (Broxham et al., 2015). Busha cattle are hypothesized to have originated from small cattle of Medieval Europe. At present, several strains of Busha exist throughout the Balkan peninsula (Broxham et al., 2015). Maltese cattle are an ancient cattle breed of Malta; it is characterized by large body size and red coat colour. Although not much is known about its origin, it is hypothesized that the origin of Maltese traces back to the prehistoric era.

Iberian cattle breeds are the group of cattle displaying a large variety of coat colours and horn morphology. Many of the Iberian cattle display ancestral characteristics such as sexual dimorphism, coat colour, and horn morphology. It has been suggested that, mostly, Iberian cattle breeds have been developed in many different types with relatively little contribution from the outside (Felius et al., 2014). However, during the 1950's, sires from exotic cattle breeds, which displayed the same coat colour as some Iberian breeds, were used in "upgrading" some of the local Iberian cattle breeds such as Alentejana and Pajuna (Felius et al., 2014).

As described in the previous section, many modern British cattle breeds such as Hereford, Longhorn, Shorthorn, were developed using modern animal breeding principles. However, several British, Scottish, and Irish cattle breeds, such as White Park, British White cattle, and Highland cattle, have been developed with minimal human interventions. Further, many British and Irish cattle breeds have individuals that display various ancestral traits (van Vuure, 2005). In this thesis, we also used genotyping data of commercial European cattle breeds such as Dutch cattle breeds and Jersey for comparative purposes. Some of the Dutch cattle breeds investigated in the study have undergone a drastic reduction in effective population size, for example, Dutch Friesian. HF and Jersey are among the most widespread cattle breeds in the world. HF originates from the Dutch provinces of North Holland and Friesland, while Jersey originates from Jersey Island. These cattle breeds are suitable for intensive farming which aims

at maximizing the overall production and economic profit.

Apart from primitive and commercial cattle breeds, we also studied various Swedish and Dutch traditional cattle breeds. Note that the term “primitive” used in this context, refers to the selection of breeds based on their ancestral phenotypes. However, no such distinction is made while using the term “traditional”, the breed defined as traditional should be native to a particular region and maintained using traditional ways. Swedish traditional cattle used in the study includes various mountain breeds from the northern and western part of Sweden and some commercial cattle breeds of southern Sweden. A large phenotypic diversity exists among these Swedish cattle breeds. For example, white coat colour and polledness are predominant traits in Swedish mountain cattle breeds, while a large number of southern Swedish cattle breeds display red coat colour and a relatively high frequency of horned individuals. These breeds also display large temporal variation regarding the foundation of breed standards and herd books. For instance, Swedish mountain cattle was recognized as a breed way back in the 19th Century, while Vaneko was recognized as a breed in the late 20th Century.

Table 1.1: Table showing information of samples genotyped in this thesis. Sampling information: First column is Breed information where, in bracket, “C” stands for commercial breed, second column is Breed code, third and fourth columns displays information about country and region of origin for the breed respectively, fifth column displays number of samples collected per breed, sixth column display present conservation status which is obtained from Domestic Animal Diversity Information System (DAD-IS) on 06th November 2018. The last column displays the types of markers used in the present thesis; note that it does not necessarily indicates the number of individuals genotyped using each type of markers. Abbreviations: ALP-Alpine, BRI-British and Irish, NLD- Dutch, JE- Jersey, IBR- Iberian, BAI-Balkan and Italy, SAN-Scandinavian, WGS-whole genome sequencing data, 777K SNP-array-bovine 777K SNP High density array- (Illumina Inc.), 150K SNP array- bovine 150K Genomic Profiler High-Density SNP array (Illumina Inc. through GeneSeek©). Note that generally the conservation status—at risk—is allotted to the population with an effective population size less than 10,000.

Breed	Code	Country of origin	Region/ species if not taurine	Sample size	Conservation status	Genetic markers used
Brown Swiss (C)	BS	Switzerland	ALP	4	Not at risk	777K SNP array
Fleckvieh (C)	FL	Switzerland	ALP	4	Not at risk	777K SNP array
Chianina	CH	Italy	BAI	3	Not at risk	777K SNP array and WGS
Maremmiana	MA	Italy	BAI	5	Not at risk	777K SNP array and WGS
Podolica	PO	Italy	BAI	1	Not at risk	777K SNP array and WGS

Maremmana x Pajuna	MP	NLD	BAI X IBR	1		777K SNP array
Busha	BU	Balkan region	BAI	6	At risk	777K SNP array and WGS
Romanian grey	RO	Romania	BAI	4	Not known	777K SNP array
Maltese	MT	Malta	BAI	4	At risk	777K SNP array and WGS
Boskarin	BK	Croatia	BAI	4	At risk	777K SNP array and WGS
Nellore	NE	Brazil	Bos indicus	4	Not at risk	777K SNP array
Aurochs	AU	Britain	Bos primigenius	1	Extinct	WGS
Angler (C)	AN	Germany	NLD	1	Not at risk	777K SNP array
Dutch Belted (C)	DB	The Netherlands	NLD	2	At risk	777K SNP array
Dutch Friesian (C)	DF	The Netherlands	NLD	4	At risk	777K SNP array
Groningen Whiteheaded (C)	GW	The Netherlands	NLD	5	Not at risk	777K SNP array
Holstein Friesian (C)	HF	The Netherlands	NLD	5	Not at risk	777K SNP array
MRY (C)	MR	The Netherlands	NLD	4	Not at risk	777K SNP array
English Longhorn	EL	England	BRI	4	At risk	777K SNP array
Galloway	GA	Scotland	BRI	5	At risk	777K SNP array
White Park	WP	England	BRI	3	At risk	777K SNP array
Highland	HL	Scotland	BRI	5	At risk	777K SNP array
Kerry Cattle	KC	Ireland	BRI	4	At risk	777K SNP array
Heck	HE	Germany	NLD	5		777K SNP array
Alentejana	AL	Portugal	IBR	2	Not at risk	777K SNP array
Arouquesa	AR	Portugal	IBR	3	At risk	777K SNP array
Cachena	CC	Portugal	IBR	3	Not at risk	777K SNP array
Caldela	CL	Portugal	IBR	1	At risk	777K SNP array
Mirandesa	MI	Portugal	IBR	2	At risk	777K SNP array

Berrenda en colorado	BC	Spain	IBR	3	At risk	777K SNP array
Berrenda en negro	BN	Spain	IBR	3	At risk	777K SNP array
Cardena	CA	Spain	IBR	5	At risk	777K SNP array
Lidia	LI	Spain	IBR	3	Not at risk	777K SNP array
Limia	LM	Spain	IBR	4	At risk	777K SNP array
Maronesa	ME	Spain	IBR	6	At risk	777K SNP array and WGS
Pajuna	PA	Spain	IBR	6	At risk	777K SNP array and WGS
Sayaguesa	SA	Spain	IBR	5	At risk	777K SNP array and WGS
Tudanca	TU	Spain	IBR	2	Not at risk	777K SNP array and WGS
Jersey (C)	JE	Jersey Island	Jersey	4	Not at risk	777K SNP array
Swedish Mountain Cattle	SMC	Sweden	SCAN	23	At risk	150K SNP array
Fjallnara cattle	FNC	Sweden	SCAN	16	At risk	150K SNP array
Swedish Polled cattle	SPC	Sweden	SCAN	3	At risk	150K SNP array
Bohus Polled	BPC	Sweden	SCAN	6	At risk	150K SNP array

1.4 Present genetic diversity status of primitive and traditional cattle

Advancement in quantitative genetics theory and techniques related to biotechnology, after the end of the second world war, led to a rapid increase in beef and dairy production in Europe. However, this rapid increase in production was brought about by using only a handful of north-western European (NWE) cattle breeds. Moreover, the effective population size for some of these NWE cattle breeds reduced to less than fifty (Gautier et al., 2007) because of intensive selection and repetitive usage of germplasm from proven sires. At the same time, industrial demand in some of the countries, where the development in animal husbandry was still in its nascent stage, led to the import of germplasm from these productive NWE breeds. As a result, the number of local cattle breeds with a long history of adaptation in their respective

environments reduced drastically (Medugorac et al., 2009). Moreover, in some European regions, where livestock was mainly used for draft purposes, the mechanization of agriculture led to a decline in effective population size in those cattle breeds. For instance, the effective population size for Andalusian black cattle breeds reduced steeply in the last decade of the twentieth century as a result of agriculture mechanization (Feliu et al., 2014). Similarly, the effective population size of Romanian grey cattle dropped from about ~ 0.2 million at the end of 19th century to just ~ 500 animals in the beginning of 21st century.

According to the FAO report (FAO, 2015), cattle are among the mammalian species with the highest number of breeds at risk. In fact, the report also provides some other worrisome statistics. For instance, of the total 1,408 global cattle breeds, the diversity status of more than 750 breeds remains unknown. Further, of the total 640 global cattle breeds with known “risk status”, 171 breeds have been classified under “at risk” category while 184 breeds are already extinct (FAO, 2015). Therefore, using genetic markers to estimate the status of genetic diversity of traditional cattle breeds is an import step towards breed conservation. One of the questions that might arise from this chapter is: what is the need of conserving primitive cattle breeds? Based on the literature that I surveyed, I give the following three broad arguments to underscore the importance of primitive cattle breeds:

- 1). Long adaptation history in their respective environments: primitive cattle breeds represent cattle populations that have a long history of adaptation in their respective indigenous environment. For instance, Italian Podolic cattle breeds such as Chianina and Maremmana are well adapted to the harsh environment, and they also display a good growth ability and resistance against parasitic diseases (Sargentini et al., 2010).
- 2). The abundance of rare alleles: It has been postulated that, because some Balkan cattle breeds such as Busha have large effective population sizes for a very long time, they might have conserved an abundance of rare alleles, some of which are lost alleles in production cattle breeds (Medugorac et al., 2009). Therefore, diversity in traditional cattle breeds represents gene pool which may play an important role to fulfil the needs of future generations.
- 3). Heritage values and unique products: In many instances, primitive cattle breeds are linked to socio-cultural values of local tradition. Moreover, the products obtained from local breeds might have some additional value that could distinguish them from commercial breeds.

1.5 Measures of genetic variation/diversity

Genetic variation can be measured as the differences in two DNA sequences sampled randomly from a panmictic population or any other well-defined population. Therefore, it can refer to variation within a population or a genome. Further, variation within an individual genome can also capture variation in a population as the haplotypes of an individual are a sample of the haplotypes segregating in a population. Two important sources of variation are de novo mutations and recombination. Genetic variation arises depending on the consequences of mutations in a genome. For instance, sometimes mutations can lead to single base pair substitution which is

called single nucleotide polymorphism (SNP) when the frequency in the population has reached a minor threshold typically more than 1%. Recombination generally does not create any de novo mutation, but rather it creates new combinations of alleles by reshuffling the genetic materials between homologous chromosomes during meiosis.

Heterozygosity is among the first parameter that often has been used by researchers to represent genetic variation in a natural population (Beja-Pereira et al., 2003; Cymbron et al., 2005). The term heterozygosity refers to the state of having two distinct alleles at a locus. The overall heterozygosity in a genome gives insight on genetic structure and demographic history of a population. For instance, reduced heterozygosity can indicate low genetic variability which can be the result of selection or a demographic process that severely reduced the population size (i.e., Bottleneck). As selection only acts on specific genetic segments, which depends on its contribution to the overall fitness of the individuals, its effect on heterozygosity would be local compared to genetic drift which would affect the entire genome.

Another parameter, which not only measures heterozygosity in a population but also provides additional information about the factors that generated it, is called runs of homozygosity (ROH). ROH are segments of identical haplotypes in an individual that are identical by descent (IBD). Inbreeding and selection are the most common causes that result in ROH within a genome. Another cause being non-random association between alleles, the phenomenon also known as linkage disequilibrium (LD). More often, ROH due to ancestral LD are much smaller in size compared to recent inbreeding as in the latter case, haplotypes have not had enough time to break-down due to recombination. Therefore, varying length of ROH provide insight into the level of inbreeding and demographic history of a population (Bosse et al., 2012).

1.6 Gene flow and genetic variation/diversity

Typically, the term migration in genetics refers to “gene flow” which is defined as the movement of alleles from one population to another. It is also an important factor affecting genetic variation. It reduces the genetic variation between previously isolated populations. This reduction in variability, however, depends on the rate and duration of gene flow. At the genomic level, gene flow followed by recombination makes the chromosomes of admixed populations mosaics of chromosomal blocks from different admixing populations (Lawson et al., 2012). Further, other population genetic forces such as selection or/and drift would determine the dispersal of introgressed segments in a population (Bosse et al., 2014).

As the events involving gene flow usually reduce allele frequency differences between the populations, several statistical approaches have been developed to classify individuals into “K” different clusters based on genetic similarity. The maximum-likelihood based approaches as implemented in the software-ADMIXTURE (Alexander et al., 2009) and STRUCTURE (Pritchard et al., 2000)-estimates underlying global admixture coefficients for each of the user-defined ancestral populations. These methods assume independence of markers. Therefore, it is important to filter SNPs based on a threshold of squared Pearson coefficient of correlation (r^2) estimate of

LD before performing the analysis.

Another way of estimating admixture events using independent SNP markers is by measuring the shared drift between populations (Patterson et al., 2012). These measures are the extension of Wright's F statistics which measures the population differentiation based on allele frequencies. Shared-drift based measures assume the null hypothesis that a tree-like fashion relates populations under investigation, i.e., they evolved independently after divergence (Figure 1.3A). Therefore, the branch lengths in the population phylogeny correspond to the amount of drift that has occurred after the divergence. The alternative model, in addition to branches, extends the phylogeny by allowing edges that represent migration events (Figure 1.3B and 1.3C). In other words, in the case of gene flow events, there will be an allele frequency correlation between source and admixed populations. However, the significant drift in either/or both admixing and source population after admixture can distort the correlation in allele frequencies. The shared-drift based measures calculated based on allele frequencies of three and four population are known as f_3 and f_4 tests, respectively. The algorithm implemented in the tool Treemix (Pickrell and Pritchard, 2012) is another interesting approach which assumes independence in allele frequencies between populations, and by modelling their relationships as bifurcating tree, it infers the migration events among sets of populations.

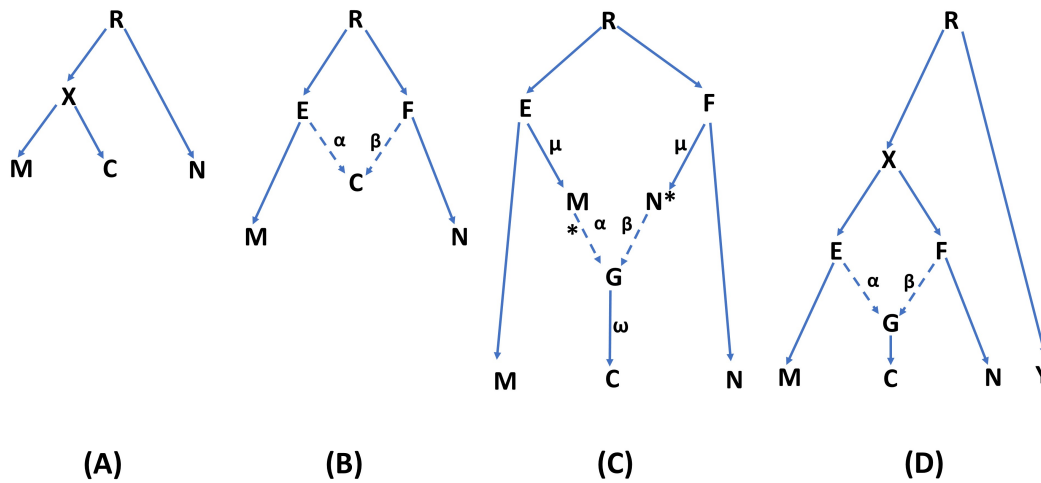


Figure 1.3: .Different demographic models: (A). Present day population M, N, and C evolved independently without any significant admixture. (B). Population C formed as a result of interbreeding between the population E and F that are ancestors of modern populations M and N respectively. (C). After receiving gene flow from population M and population N, population G undergoes a significant genetic drift. Note that "α" and "β" represents a proportion of gene flow, while "ω" represents the amount of genetic drift. The figure is adapted from Patterson et al., (2012).

Almost all the previously defined algorithms use SNPs individually and assume independence

between successive markers. However, the advent of cost-effective high throughput technologies has resulted in array-based approaches which can genotype thousands or hundreds of thousands of closely positioned markers, and the analysis of such data can, sometimes, violate this assumption of independence. Haplotype-based analyses can harness the information from such closely linked data, leading to improvement in the inference of population structure. One such algorithm is implemented in a suite of a program called *fineStructure* (Lawson et al., 2012). The algorithm reconstructs each “recipient” haplotype as a mosaic of haplotypic blocks of all the other “donor” haplotypes in the dataset using a Hidden Markov Model method as introduced by Li and Stephens (Li and Stephens, 2003). Essentially, this reconstruction results in the co-ancestry matrix wherein each value corresponds to the shared ancestry between any two haplotypes in the dataset. Later, the co-ancestry matrix is used by *fineStructure* to assign individuals into population using a Markov chain Monte Carlo (MCMC) algorithm.

1.7 Brief review on studies of genetic admixture in primitive European cattle

Because primitive cattle breeds show many ancestral phenotypes, the hypothesis of post-domestication contact between aurochs and the ancestors of these cattle has been proposed (Achilli et al., 2008; Bonfiglio et al., 2010). Many studies, using uniparental markers such as mtDNA and Y-chromosome haplotypes, investigated the hypothesis of post-domestication gene flow in European cattle (Götherström et al., 2005; Achilli et al., 2008; Bollongino et al., 2008; Bollongino et al., 2012). For instance, gene flow between Italian domestic cattle and Italian aurochs has been proposed based on the observation that Italian aurochs also carried mitochondrial T3 haplogroups which is the most common haplogroup among European taurine (Beja-Pereira et al., 2006). Additionally, Italian cattle breeds such as Romagnola and Chianina also displayed low frequency of several novel mtDNA haplogroups such as Q and R which also has been proposed as a legacy of local Italian aurochs (Achilli et al., 2008; Bonfiglio et al., 2010). On the other hand, Götherström et al., (2005) observed a high frequency of Y1 haplogroups in (Y-chromosomal markers) aurochs samples retrieved from north-western Europe and, since Y1 haplogroup is also the most common haplogroup among north-western European domestic cattle, they proposed that gene flow between aurochs and domestic cattle of north-western Europe might have occurred after domestication. Later, (Bollongino et al., 2008) refuted this hypothesis by showing that there was no difference in the frequencies of Y1 and Y2 haplogroups among the aurochs samples retrieved from north-western Europe. It should be noted that until now mitochondrial haplogroup of the British aurochs, i.e., P-haplogroup, has only been found in one to two individuals of modern European taurine (Achilli et al., 2008).

Previous studies have also hypothesized that gene flow has occurred between south-eastern European cattle and non-European cattle (zebu and African taurine) (Beja-Pereira et al., 2003; Cymbron et al., 2005; Ginja et al., 2010; Decker et al., 2014). For instance, studies using genome-wide SNPs and microsatellite markers have shown that African taurine ancestry is a common feature of Iberian cattle (Decker et al., 2014). On the other hand, a gradient of zebu ancestry from

southern Europe to Northern Europe cattle also had been proposed (McTavish et al., 2013). However, a recent study refuted this hypothesis and instead proposed that only a handful of cattle breeds, especially from Italy, carry zebu ancestry in their genome (Decker et al., 2014). The majority of these studies focused only on major breeds from Iberia, Italy, and North-western Europe and lacked in genotypes from other Eastern European regions that are close to the center of domestication.

1.8 Structural variation and its contribution to cattle diversity

Structural variation (SVs) is a term that includes various genomic alterations (Layer et al., 2014) such as insertions, deletions, duplications, inversions, translocations, or other complex rearrangements of large genomic segments (Figure 1.4). Though SVs are not very common, they may have a great impact on gene structure and function (Bickhart and Liu, 2014). Therefore, SVs are an important source of genetic and phenotypic variation between individuals.

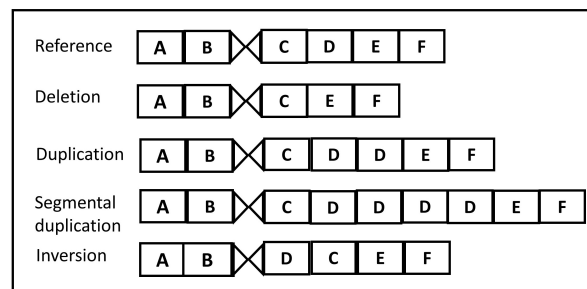


Figure 1.4: Some examples of structural variation in genome sequences

The advancement in the methodologies of genome sequencing has not only accelerated the discovery and genotyping of SVs but have also increased our understanding of its type and formation. Based on the effect of overall genome size, SVs can be categorized in two types: Balanced (translocation and inversion) and unbalanced (insertion, deletion, and duplication) (Bickhart and Liu, 2014). The unbalanced class of SVs can also be called copy number variation (CNV). CNV encompasses a large proportion of cattle and human genomes. For example, in the human genome CNV is estimated to have covered between 4.8-9.5%, while in cattle, it covered approx. 3% of the genome (Zarrei et al., 2015; Bickhart et al., 2016). The methods to identify CNV can either make use of whole genome sequencing data or array probe signal intensities. These methods are also more refined compared to the methods aimed at identifying balanced SVs as the sequence breakpoints in balanced SVs are difficult to pinpoint (Bickhart and Liu, 2014).

CNV play an important role depending on where in the genome it is present. Genic CNVs, for instance, can influence phenotypes of an organism through at least three different mechanisms: change in gene dosage, exposure to recessive alleles and expression regulation changes.

Further, it has been shown that CNVs, if present in the regulatory elements of developmental genes, can also change the phenotypic expression (Spielmann and Klopocki, 2013). Moreover, duplication of the genic region may lead to another gene copy acquiring a novel functional role (*neofunctionalization*), or the gene's functional role may get divided between these paralogs (*sub-functionalization*), thereby contributing to the genome evolution. However, genes that are conserved across species or genes that are essential for multiple biological pathways are predicted to be sensitive to CNVs affecting gene expression (Schuster-Bockler et al., 2010). Moreover, different types of repeat regions in a genome contribute to the formation of CNVs. For instance, CNVs are reported to be associated with segmental duplications in mammalian genomes (Sharp et al., 2006). These repetitive regions in the genome facilitate the formation of CNV through mechanisms such as non-allelic homologous recombination (Warburton et al., 2008). In fact, a recent study has shown that such repetitive regions in the genome are five times more likely to harbour CNVs when compared to germline CNVs (Monlong et al., 2018).

The availability of the Bovine50K and BovineHD 777K SNP arrays has revolutionized the field of bovine genomics. Extensive use of such arrays has led to the identification of many CNVs in the bovine genome (Fadista et al., 2010; Bickhart et al., 2012; Bickhart et al., 2016; Sasaki et al., 2016; Wang et al., 2016). As a result, a complex landscape of CNVs in the bovine genome has emerged. It has been shown that some gene families such as an Olfactory receptor (OR) and genes that play a role in the immune system harbour an abundance of CNVs. Because both these complex gene families serve important functions associated with a sense of smell and ability to resist pathogens, respectively, the evolutionary selection pressure might have played an important role in generating and maintaining variable copy numbers. In cattle, like sheep and pig (Moller et al., 1996; Han et al., 2015), SVs affecting coat colour have also been reported (Durkin et al., 2012; Brenig et al., 2013).

1.9 Identification of structural variations

1.9.1 SNP-array based identification

SNP array platforms typically target biallelic SNPs by including two types of probes, usually coded as A and B, for every single SNP. The resulting hybridization between targeted DNA fragments and probes generates hybridization intensity, which can be used to determine SNP genotypes (Wang et al., 2007). For instance, SVs involving deletions and duplications decrease or increase the total signal intensity, respectively. Apart from the signal intensity, other genomic factors such as GC content around the targeted sites or population allele frequencies can also be included in models to increase the accuracy of identification of SVs. In principle, these methods can only identify SVs involving deletions or duplications. Further, these methods cannot reliably identify break-points around SVs. Some examples of computation programs that can identify SVs based on signal intensity data of SNP array include PennCNV (Wang et al., 2007) and QuantiSNP (Colella et al., 2007).

1.9.2 WGS-based identification of SVs

The methods used to identify SVs from whole genome sequence (WGS) data can be categorized in four classes: Read-pair (RP), Split-read (SR), Read depth (RD) and assembly-based methods.

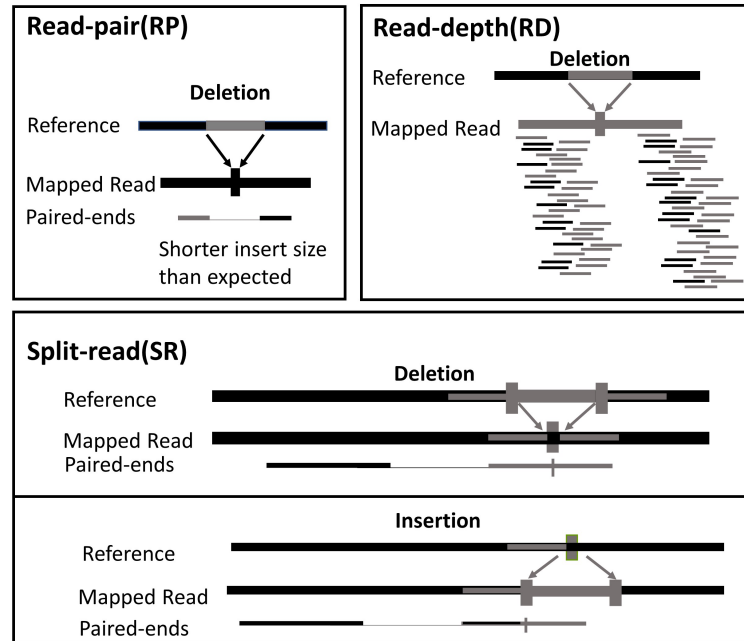


Figure 1.5: Some examples of identification of structural variation events using different whole genome re-sequencing approaches. The figure is adapted from Pirooznia et al.,(2015)

In the paired-end sequencing, DNA fragments are likely to display a specific distribution around the insert size (Korbel et al., 2007). Therefore, read spanning SVs may display a different insert size compared with the genomic average and read pair-based methods use these discordant paired-end reads to identify SVs. However, small sized SVs are difficult to detect using these methods as small disruptions in insert size are difficult to separate from the normal background dispersion in insert size distribution (Medvedev et al., 2009). Further, read pair-based methods are not preferred for detection of SVs in low complexity regions of the genome (Pirooznia et al., 2015). Some methods, in addition to read pair, also consider the split read information to locate precise break-points of SV events (Zhang et al., 2011). Split read methods use reads that remain completely or partially unmapped to the reference genome. Read depth methods exploit the depth of coverage information of genomic alignments to identify deletions or insertions, as there is a direct correlation between the copy number events and depth of coverage (Pirooznia et al., 2015). As opposed to read pair and split read, read depth methods can identify the exact copy number of an event, while the former methods only report the position and the type of event. Moreover, compared to read pair and split read, read depth methods have a higher sensitivity to large CNVs. However, read depth has low efficiency when identifying small CNVs (<1 kbp) (Pirooznia et al., 2015).

There are methods implemented in various tools (Sindi et al., 2012; Layer et al., 2014) that consider a combination of one or more methods described in the previous paragraph. This combined approach often results in a better accuracy of SV identification compared to any single method. In principle, the combination approach-based method combines information from multiple methods, taking advantage of their strength. In doing so, they also overcome the limitation of one method with the unique feature of the another. For instance, combining Read pair, Split read with Read depth has resulted in a high accuracy of identification of small as well as large-sized CNVs (Pirooznia et al., 2015).

1.10 Thesis outline

The overall goal of my research is to investigate the pattern of genetic variation, gene flow and demography in primitive cattle breeds of Europe. By analyzing genotyping data of a large number of cattle breeds, I disentangle the complex relationships between European, African and zebu cattle. Additionally, I also give a broad overview of genetic diversity in some of the least studied cattle breeds of Europe. The practical implications and future direction of the research associated with the results of this thesis are also discussed. In **chapter 2**, I evaluate the relationship between European cattle and an ancient aurochs sample. Also, I report zebu ancestry in several Balkan and Italian cattle breeds. Further, I also assign observed ROH in different European cattle breeds to different demographic changes in breed history. In **chapter 3**, I further explore zebu ancestry in European cattle breeds by including genotype data of several additional Podolian cattle breeds. Furthermore, I also infer African taurine ancestry in several Iberian cattle breeds. I conclude by discussing the origin of Balkan and Italian cattle considering their common non-taurine ancestry. In **chapter 4**, I investigate the breed structure and diversity of Swedish traditional cattle breeds. In particular, I emphasize the low differentiation between different Swedish mountain cattle breeds. In **chapter 5**, I explore the pattern of CNV among different European cattle breeds and discuss these in light of different population genetic forces. I also identify several copy number genes associated with phenotypic diversity. **chapter 6** further discusses the copy number variable genes identified using whole genome sequences of different Eurasian and African cattle breeds. I conclude the thesis with Chapter 7 by discussing my findings, their importance, and applicability in a broader context.

Chapter 2

Genetic origin, admixture and population history of aurochs (*Bos primigenius*) and primitive European cattle

M.R. Upadhyay^{1,2}, W. Chen¹, J.A. Lenstra³, C.R.J. Goderie⁴, D.E. MacHugh^{5,6}, S.D.E. Park⁷, D.A. Magee⁵, D. Matassino⁸, F. Ciani⁸, H.J. Megens¹, J.A.M. van Arendonk¹, M.A.M. Groenen¹, European Cattle Genetic Diversity Consortium⁹ and R.P.M.A. Crooijmans¹

¹Animal Breeding and Genomics Centre, Wageningen University, Wageningen, The Netherlands

²Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden. ³Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands ⁴Taurus Foundation, Nijmegen, The Netherlands ⁵Animal Genomics Laboratory, UCD School of Agriculture and Food Science, University College Dublin, Ireland; ⁶UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland ⁷IdentiGEN Ltd., Unit 2, Trinity Enterprise Centre, Pearse Street, Dublin, Ireland ⁸Consortium for Experimentation, Dissemination and application of Innovative Biotechniques, ConSDABI NFP, I FAO-GS AnGR, Contrada Piano, Italy

Abstract

The domestication of taurine cattle initiated approximately 10,000 years ago in the Near East from a wild aurochs (*Bos primigenius*) population followed by their dispersal through migration of agriculturalists to Europe. Although gene flow from wild aurochs still present at the time of this early dispersion is still debated, some of the extant primitive cattle populations are believed to possess the aurochs like primitive features. In this study, we use genome-wide SNPs to assess relationship, admixture patterns and demographic history of an ancient aurochs sample and European cattle populations, several of which have primitive features and are suitable for extensive management. The Principal component analysis, the model-based clustering, and a distance-based network analysis support previous works suggesting different histories for north-western and southern European cattle. Population admixture analysis indicates a zebu gene flow in the Balkan and Italian Podolic cattle populations. Our analysis supports the previous report of gene flow between British and Irish primitive cattle populations and local aurochs. In addition, we show evidence of aurochs gene flow in the Iberian cattle populations indicating wide geographical distribution of the aurochs. Runs of homozygosity (ROH) reveal that demographic processes like genetic isolation and breed formation have contributed to genomic variations of European cattle populations. The ROH also indicate recent inbreeding in southern European cattle populations. We conclude that in addition to factors such as ancient human migrations, isolation by distance and cross-breeding, gene flow between domestic and wild cattle populations also has shaped genomic composition of European cattle populations.

Key words: Aurochs, primitive cattle, SNP, gene flow, runs of homozygosity

2.1 Introduction

The Domestication of taurine cattle (*Bos taurus*) occurred approximately 10,000 years ago in the Near East (Götherström et al., 2005) and based on mtDNA analysis, it is estimated to have incorporated the maternal lineages from around 80 female aurochs (*Bos primigenius*) in this region (Bollongino et al., 2012). Later on, domesticated cattle became widespread all over Europe through migration of early farmers. As farmers settled in the regions that harboured native European aurochs, sporadic interbreeding might have taken place between domestic cattle and native European aurochs, which persisted in some regions until the Middle Ages. Many previous studies have used mtDNA and Y-chromosome markers to address the hypothesis that local aurochs contributed to the gene pool of modern European domestic cattle. All modern European cattle breeds are characterized by the T mitochondrial haplotype. However, previous studies (Edwards et al., 2007; Scheu et al., 2008) investigating the mitochondrial haplotypes of more than one hundred ancient aurochs samples that were retrieved from multiple sites of northern and central Europe only identified P and E haplotypes, which exclude a significant contribution of local aurochs to the maternal lineages of European domestic cattle. On the other hand, many Italian and one Swiss aurochs not only carried P, but also T mitochondrial haplotypes (Mona et al., 2010; Lari et al., 2011; Schibler et al., 2014), so maternal gene flow from native wild aurochs to local domestic cattle may have taken place there. Contrary to mtDNA, studies involving Y-chromosome haplogroups have remained equivocal as to whether gene flow between local aurochs and domestic cattle has occurred (Götherström et al., 2005; Bollongino et al., 2008). Recently, Park et al., (2015) using genome-wide SNPs identified significant enrichment of British aurochs alleles in north European cattle breeds.

Until approximately 250 years ago cattle productivity was low compared to modern standards. Cattle were mainly used in subsistence farming and extensive management, a situation which still persists in much of Africa and South Asia (Feliuss et al., 2014). Since the Industrial Revolution, an intensification of animal husbandry has led to the development of many specialized breeds with derived traits and uniform appearances. However, several traditional cattle breeds still have retained primitive features of their wild ancestors; such breeds can also be referred to as primitive breeds (Supplementary note 1). Previous studies (Tapio et al., 2006; Medugorac et al., 2009) have recognized primitive cattle breeds as a valuable resource for genetic variation. Most of these cattle breeds are hardier than commercial breeds and endure adverse environmental conditions and extensive management with low quality forage better (Sæther et al., 2006). On the other hand, several of these breeds have declining effective population sizes, which erode their genetic diversity (Tapio et al., 2006).

It has been estimated that by 2030, 3 to 4 % of European farmland will be abandoned and converted to nature reserves. Management of such landscapes by grazing cattle is very easy, cheap and environmentally friendly, and requires primitive cattle breeds that only need a minimum of human intervention (Schaich et al. 2010). Thus, investigating past admixture patterns and recent change in demography of primitive cattle breeds is essential in order to formulate and fulfil the breeding objectives tailored to the purpose of landscape management.

Although economic traits of commercial breeds have been extensively studied (Lu et al. 2014; Raven et al. 2014), the more primitive traits and adaptation to extensive management have not been investigated. Analyses of genome wide single nucleotide polymorphism (SNP) data from aurochs and primitive cattle breeds provide the opportunity to investigate gene flow and admixture processes, which have contributed to the evolution of domestic cattle. Here, we use high density genome wide SNP markers to investigate genetic origins, admixture patterns and demographic histories of 27 primitive cattle populations and one British aurochs sample. These populations originate from all over Europe. The Dutch, and the Swiss commercial dairy breeds, as well as the Nelore zebu (*B. indicus*) are also included for comparing the primitive and highly productive cattle and for assessment of zebu admixture respectively. We also carry out Runs of homozygosity (ROH) based analyses as they provide valuable insights into recent demographic changes and breeding practices in livestock (Purfield et al., 2012; Ferenčaković et al. 2013).

2.2 Material and methods

2.2.1 Sample collection and classification

A total of 144 animals from 38 different cattle populations, consisting mainly of primitive cattle breeds (27 breeds) were sampled (Figure 2.1). Numbers of animals per population varied from 1 to 6 (Table S1). The geographic origin of 135 animals was assigned to one of five regions, largely corresponding Buchanan and Lenstra, (2015): British and Irish (BRI), Dutch-Baltic (NLD, ancestral to large parts of the Lowland Pied and Baltic Red cattle), Balkan and Italian (BAI, representing Podolian and Busha cattle), Iberian (IBR), and Alpine (ALP, combining the Central Brown and Spotted breed cluster). The remaining nine animals did not belong to any of these groups: Jersey (JE), Heck (HE), Nelore zebu (NE), a Maremmana-Pajuna cross (MP) and British aurochs (AU).

2.2.2 Detail of aurochs specimen

The aurochs humerus bone specimen has been retrieved in 1998 from Carsington Pasture cave in Derbyshire, England and was radiocarbon dated to $6,736 \pm 68$ years BP (Edwards et al. 2010; Park et al., 2015). The mitochondrial sequencing revealed P-haplogroup for the aurochs sample (Edward et al., 2010). Park et al (2015) extracted DNA and performed the high throughput sequencing (HTS) using Illumina paired-end DNA sequencing. Further information about the aurochs sample and sample processing has been described in the by Park et al. (2015).

2.2.3 SNP array data

DNA was extracted from hair roots, sperm or blood and genotyped with the Illumina BovineHD Genotyping BeadChip, which contains 777,692 SNPs uniformly spanning the bovine genome. Genotypes were called and processed using the Genome Studio software package (Illumina). Samples with a call rate below 90 % were excluded from the dataset and additional criteria used to filter SNPs from the dataset were as follows: 1) monomorphic SNPs; 2) SNPs having more than 5 % missing genotypes across all the samples; 3) SNPs located on sex chromosomes of the UMD 3.1 assembly (Zimin et al., 2009); and 4) SNPs showing strand discordances. These quality control steps were carried out using PLINK v1.07 (Purcell et al., 2007). The final dataset comprised of 698,452 SNPs for 138 animals.

For the British aurochs sample, 773,659 SNPs, which corresponds to SNP position in Illumina BovineHD Genotyping BeadChip, were extracted from the whole genome sequencing (WGS) reads of DNA. After extracting the 773,659 SNP positions, SNPs were selected for inclusion in the final data set if SNP genotype quality ≥ 40 and SNP position read depth ≥ 5 . After filtering, the final aurochs data set consisted of 562,428 SNPs, of which 544,735 are located on autosomes.

2.2.4 Population genetic structure and admixture

Principal component analysis (PCA) was carried out using the program SMARTPCA 6.0.1 from the EIGENSOFT package (Patterson et al., 2006). The Reynolds' genetic distances between breeds with more than two samples were also calculated using the Gendist module of PHYLIP 3.69 (Felsenstein, 1989), which was followed by construction of a NeighborNet graph using SPLITTREE (Huson, 1998).

Population admixture analysis was carried out after selecting 96,671 SNPs with linkage disequilibrium < 0.25 using sliding windows of 50 SNPs with forward shift of 5 SNPs. ADMIXTURE v1.22 (Alexander et al., 2009) was used to carry out the genetic admixture analysis with cross-validation (CV) for values of K ranging from 1 to 40. In order to test for admixture across primitive cattle populations, we calculated three-population test estimates (f_3 statistics, Reich et al., 2009) and their corresponding normalized value (z-scores) using all 698,452 SNPs in the 'threepop' module implemented in the TREEMIX software package (Pickrell and Pritchard, 2012). Three-population tests considers triplet of the populations (C;A,B), where C is the test population with A and B as reference populations. We performed all possible triplet combinations.

The z-scores were calculated by jack-knifing with blocks of 500 SNPs. If the z-score is significantly negative ($z \leq -3.80$, after Bonferroni correction for multiple testing), test population C must have admixture from both the reference populations A and B (Patterson et al., 2012). For test population "C", only breeds with more than one sample were considered.

To assess the direction of gene flow, we calculated D statistics (Green et al., 2010) as implemented in the ADMIXTOOLS (Patterson et al., 2012) software package. The D statistics method

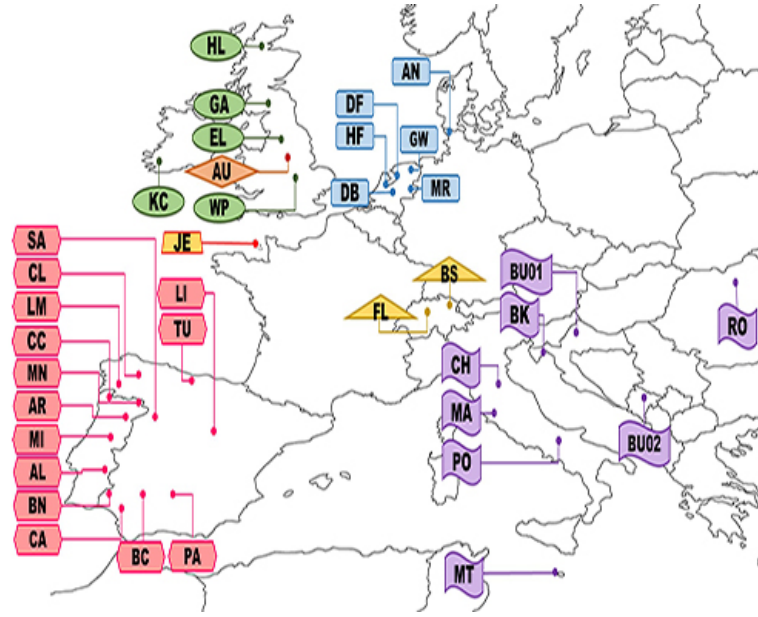


Figure 2.1: Native origin of sampled cattle breeds used in the population genetic analyses. Colours and shape refer to the geographic region of the breeds in the Europe (light green: British and Irish region breeds, light blue: Dutch region breeds, light yellow: Alpine breeds, light pink: Iberian region breed, lavender: Balkan-Italian breeds and dark yellow: Jersey breed). Breed abbreviations: BS-Brown Swiss, FL- Fleckvieh, CH-Chianina, MA-Maremmiana, PO- Podolica, BU- Busha, RO- Romanian Grey, MA- Maltese, BK- Boskarin, NE-Nelore, AN-Angler DB-Dutch Belted, DF- Dutch Friesian, GW- Groningen Whiteheaded, HF- Holstein Friesian, MRY- Meuse Rhine Issel, EL- English Longhorn, GA- Galloway, WP- White Park, HL- Highland, KC- Kerry cattle, HE- Heck, AL- Alentejana, AR- Arouquesa, CC- Cachea, CL- Caldela, MI- Mirandesa, BC- Berrenda en Colorado, BN- Berrenda en Negro, CA- Cardena, LI- Lidia, LM-Limia, MN- Maronesa, PA-Pajuna, SA- Sayaguesa, TU- Tudanca, JE- Jersey.

considers the tree topology $((W, X), Y), Z$ where; Z represents the outgroup, Y the source of admixture and W and X are the test populations. The D statistics method counts the “ABBA” sites where W and Z share the outgroup allele (A) and X and Y share the derived allele (B) as well as the “BABA” sites where W and Y shares the derived allele and X and Z shares the outgroup allele. Admixture between Y and either of the test populations creates a significant difference between the ABBA and BABA counts, with a z-score >3.0 (gene flow between W and Y) or ≤ -3.0 (between X and Y).

For the D-statistics method, we generated a yak (*B.grunniens*) SNP profile to be used as outgroup by splitting the yak scaffolds (Hu et al., 2012) into 600 bp segments followed by local alignment against bovine genome build UMD 3.1 using bwa-mem (Li, 2013) and calling bovine HD array positions using *samtools mpileup* (Li et al., 2009). Two approaches were followed to select reads for *mpileup*. In the first approach, we used all uniquely mapped reads to perform *mpileup* in order to maximize the number of allele positions. This approach generated 483,970 SNPs after filtering SNPs missing in either the aurochs or yak sample or both. In the second

approach, we only used uniquely mapped reads that completely mapped to the bovine genome in order to use exclusively high-confidence allele calling. This approach generated 77,558 SNPs after filtering SNPs missing in either the aurochs or yak sample or both. Aurochs or zebu was considered as a possible source of admixture and as test populations we compared either ALP or BAI with all other remaining populations respectively.

2.2.5 Runs of Homozygosity

ROH were detected as described by Purfield et al. (2012) by using PLINK v1.07 with a sliding window of 50 SNPs, allowing <100 kb between two consecutive homozygous SNPs, less than two missing genotypes and one possible heterozygous SNP and a minimum length of 500 kbp.

2.3 Results

2.3.1 Population genetic structure and admixture

Genetic structure of cattle belonging to various geographical groups was analysed using the PCA and genetic distance-based clustering. The PCA analysis does not take into account breed membership but yields clear structure as samples from the same population clusters together (Figure 2.2). The combination of the first two principal components (PC), namely, PC1 and PC2 separates samples according to their geographic origin. PC1 accounts for 16.5 % of the total variation and separates *B. indicus* (Nelore-NE) and *B. taurus* samples (Figure 2.2a). PC2 separates the cluster of BRI, NLD (north-western) cattle from another cluster of IBR, ALP and BAI (southern-central) cattle. Furthermore, PC2 positions English Longhorn (EL) samples away from other north-western European cattle samples. To explore the genetic structure only among European cattle, the PCA was performed without *B. indicus* (NE) and outlier breeds (EL, MT) as identified in the PCA of all samples (Figure 2.2a & b). Interestingly, the largest PC (PC1) in the PCA of all European samples separates disperse cluster of BAI cattle from the remaining European cattle indicating high divergence of BAI cattle (Figure 2.2c). PC2 separates densely cluster IBR cattle from the overlapping cluster of north-western cattle indicating high genotypic similarities among IBR cattle. The PCA analysis also suggests north-western European cattle ancestry for JE breed. The British aurochs sample is positioned in the centre of the plot near the junction of BAI and IBR samples.

A NeighborNet graph (Figure S1) separates breeds according to geographical origin. As expected, outlier breeds suggested by the PCA, such as EL along with Maltese (MT) and Cachena (CC) have branches originating from the centre of the graph. Short branches in the graph also indicate low divergence among Iberian cattle breeds. The IBS distance-based neighbour-joining tree (Figure S2), with NE samples assigned as the artificial root, separates samples according to their geographic origin in accordance with the PCA. The BAI population appears to be paraphyletic, while IBR, ALP, NLD and BRI formed clear separate clusters.

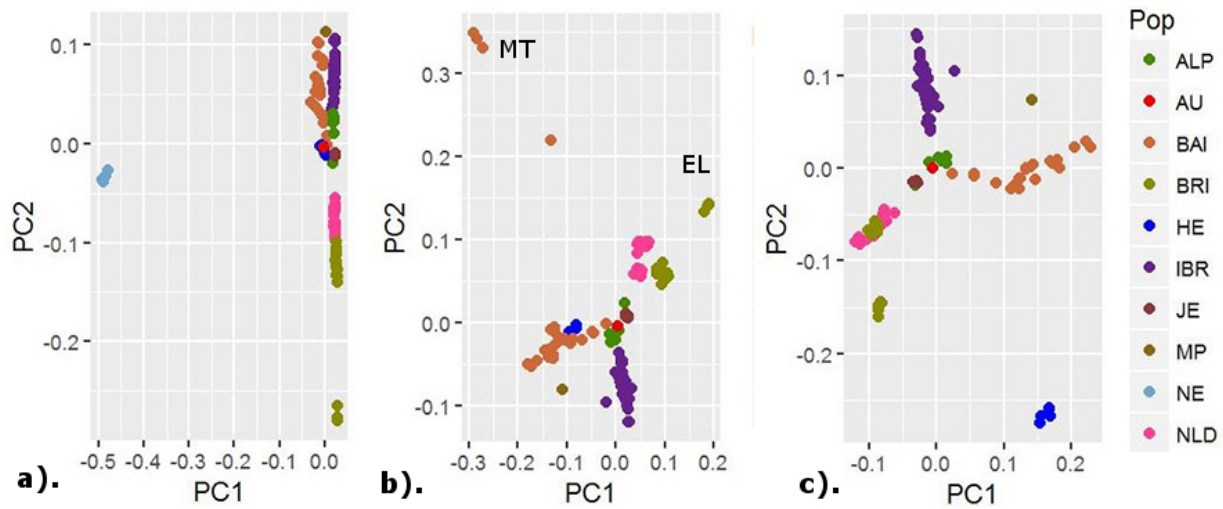


Figure 2.2: Principal component analysis (PCA) plot constructed for PC1 and PC2: (a) based on all available samples; (b) excluding NE samples ;(c) excluding MT, EL and NE samples. Samples are grouped based on their geographical region. Abbreviations: ALP- Alpine, BAI-Balkan and Italy, BRI-British and Irish, HE-Heck, IBR- Iberian, JE- Jersey, MP- Maremmana X Pajuna, NE-Nelore, NLD-Dutch. Breed abbreviations: BS-Brown Swiss, FL- Fleckvieh, CH-Chianina, MA- Maremmana, PO- Podolica, BU- Busha, RO- Romanian Grey, MA- Maltese, BK- Boskarin, NE-Nelore, AN-Angler DB-Dutch Belted, DF- Dutch Friesian, GW- Groningen Whiteheaded, HF- Holstein Friesian, MRY- Meuse Rhine Issel, EL- English Longhorn, GA- Galloway, WP- White Park, HL- Highland, KC- Kerry Cattle, HE- Heck, AL- Alentejana, AR- Arouquesa, CC- Cachea, CL- Caldela, MI- Mirandesa, BC- Berrenda en Colorado, BN- Berrenda en Negro, CA- Cardena, LI- Lidia, LM-Limia, MN- Maronesa, PA-Pajuna, SA- Sayaguesa, TU- Tudanca, JE- Jersey.

In the ADMIXTURE analysis, estimates of cross validation (CV) decreases with increasing number of user-defined ancestral populations (K) from 1 to 3 (Figure S3). Increasing K from 3 to 4 increases the CV only marginally, whereas at $K > 4$, the CV error estimate increases substantially, suggesting $K=3$ as the most likely number of clusters. The population subdivision at $K=2$ reproduces the first PCA coordinate by separating NE samples (*B. indicus*) from all European samples (*B. taurus*) and also indicates an apparent indicine component in BAI and HE cattle (Figure 2.3). The population subdivision at $K=3$ reproduces the second PCA coordinate by separating the IBR cluster from north-western samples. The NLD samples were assigned to a separate cluster at $K=4$ (data not shown). The single aurochs sample displayed a mosaic of genotypes presents in all clusters estimated across a range of K value from 2 to 8 (data not shown).

Significant negative f_3 statistics with NE zebu as one of the reference populations indicates for the BAI breeds (Busha (BU), Chianina (CH) and Maremana (MA)) a zebu ancestry (Table S2). A high number of significant negative f_3 statistics (Table S3) suggests admixture between regional cattle as is observed for all IBR breeds (except Marondesa). The majority of IBR breeds, Chianina (CH), Jersey (JE), White Park (WP), Fleckvieh (FL), Dutch Belted (DB),

Dutch Friesian (DF), MRY and Brown Swiss (BS) exhibits evidence of aurochs ancestry (at least one significant f_3 test with aurochs as one of the reference populations, Table S4). Non-significant f_3 statistics indicate isolated breed structures for majority of the BRI and NLD breeds (Table S4). Contrary to the f_3 statistics, D statistics with yak as outgroup (Z), aurochs as a source of gene flow (Y) and ALP breeds as one of the sister taxa reveals the highest frequencies of the aurochs alleles in the BRI cattle, lower frequencies in the IBR and the NLD cattle and the lowest frequencies in the BAI cattle (Table 1 and Table S5). In agreement with the previous analyses (PCA, ADMIXTURE and F_3), the D-statistics with NE as a source of gene flow (Y) also suggests the highest frequencies of zebu specific alleles in the BAI cattle (Table 2).

2.3.2 Runs of Homozygosity

The number of ROH per animal varied from 49 in a Busa (BU) to 357 in a Groningen White-headed (GW) animal. For the total number of ROH per sample per breed (Table S6), the EL breed displays the highest mean ($\sim 310 \pm 14$), while BU01 cattle displays the lowest mean ($\sim 62 \pm 14$). As expected, the average proportion of the genome coverage by ROH correlates with observed heterozygosity (H_o) (Figure S4). The highest average proportion of the genome covered by ROH is observed in isolated breeds with low mean observed heterozygosity: EL ($\sim 848 \pm 34$ Mbp), MT (825 ± 529 Mbp) and the Iberian Mirandesa (MI, 783 ± 227 Mbp) (Table S6). In two MT animals, the ROH accounted even for a third of their genome. Thus, the ROH coverage in the genome varies considerably within regions with generally high values in a BRI breeds (except WP breed) and a large spread of the value for BAI and IBR populations. The ROH profiles (size vs. number) (Figure 2.4) display differences in pattern across the different cattle populations. On the regional level, the BRI cattle exhibit the highest average cumulative size as well as the highest average number.

Table 2.1: Result of D statistics performed to detect admixture from aurochs (Y) to either W or X. The negative D statistics indicates that gene flow has occurred from Y to X and the positive D statistics indicates that gene flow has occurred from Y to W. The asterisk (*) indicates the significant Z-value. Abbreviations: ALP-Alpine, BRI-British and Irish, NLD-Dutch, JE- Jersey, IBR- Iberian, BAI- Balkan and Italy.

W	X	Y	Z	D-stat	Z-value
ALP	BRI	AU	Yak	-0.0289	-10.46*
ALP	NLD	AU	Yak	-0.0117	-4.55*
ALP	IBR	AU	Yak	-0.0081	-3.48*
ALP	JE	AU	Yak	-0.0168	-4.46*
ALP	BAI	AU	Yak	0.0655	24.14*
IBR	NLD	AU	Yak	-0.0037	-1.81

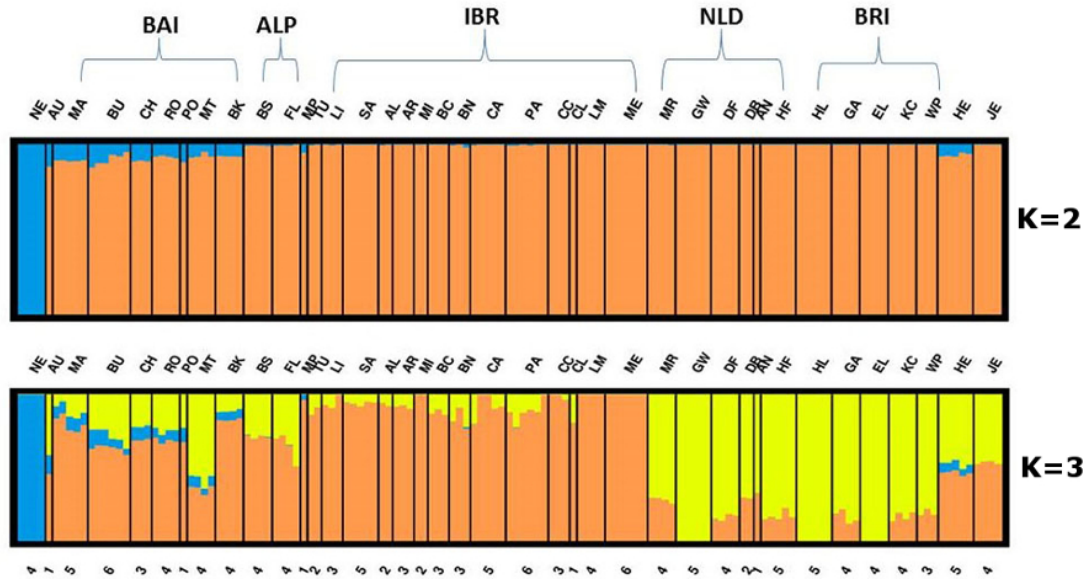


Figure 2.3: ADMIXTURE analysis plot showing model based population assignments for the values of K=2 (upper) and K=3 (lower). Almost all BAI, AU and HE samples display a possible zebu genetic component at K=2 and K=3. The number below each bar represents the number of samples for the respective population. Abbreviations: ALP-Alpine, BAI-Balkan and Italy, BRI-British and Irish, HE-Heck, IBR- Iberian, JE- Jersey, MP-Maremmna X Pajuna, NE-Nelore, NLD-Dutch. Breed abbreviations: BS-Brown Swiss, FL- Fleckvieh, CH- Chianina, MA- Maremmna, PO- Podolica, BU- Busha, RO- Romanian Grey, MA- Maltese, BK- Boskarin, NE-Nelore, AN-Angler DB-Dutch Belted, DF- Dutch Friesian, GW- Groningen Whiteheaded, HF- Holstein Friesian, MRY- Meuse Rhine Issel, EL- English Longhorn, GA- Galloway, WP- White Park, HL- Highland, KC- Kerry Cattle, HE- Heck, AL- Alentejana, AR- Arouquesa, CC- Cachea, CL- Caldela, MI- Mirandesa, BC- Berrenda en Colorado, BN- Berrenda en Negro, CA- Cardena, LI- Lidia, LM-Limia, MN- Maronesa, PA-Pajuna, SA- Sayaguesa, TU- Tudanca, JE- Jersey.

On the other hand, many samples from the BAI and IBR cattle display a cumulative size comparable to cattle sampled from the BRI and NLD populations, but with a small number of ROHs, indicating significant contribution primarily from long ROH. Long ROHs in the genome are a consequence of recent inbreeding, Whereas short ROHs indicate distance ancestral relatedness such as breed founder effect. In order to assess the distribution of ROHs length, ROH were further classified into short (0.5-2Mbp), intermediate (2-8 Mbp) and long (> 8Mbp) (Figure 2.5). We also note that north-western as well as central European breeds compared to southern European breeds display more instances of similar ROH patterns for all samples within breed. For instance, all samples of GW breed display a similar distribution of ROH across different categories of ROH length. In fact, for all GW samples more than 95 % of the total ROH length belongs to the short to intermediate ROH categories.

Table 2.2: The result of D statistics performed to detect admixture from Nellore (NE) (Y) to either W or X. The negative D statistics indicates that gene flow has occurred from Y to X and the positive D statistics indicates that gene flow has occurred from Y to W. The asterisk (*) indicates the significant Z-value. Abbreviations: ALP-Alpine, BRI-British and Irish, NLD- Dutch, IBR- Iberian, BAI- Balkan and Italy.

W	X	Y	Z	D-stat	Z-value
BAI	BRI	NE	Yak	0.0353	16.30*
BAI	NLD	NE	Yak	0.0297	14.30*
BAI	IBR	NE	Yak	0.0284	15.84*
BAI	ALP	NE	Yak	0.0250	11.74*

Individual cattle from the same breed are expected to display similar ROH patterns because of shared population history and similar treatments due to a common herdbook. Comparison of BRI cattle to southern (IBR-BRI) populations shows over representation of all classes of ROH (Figure 2.5).

To assess the within the diversity and effect of selection, exact sharing of ROH haplotypes between animals from the same breed was also assessed (Table S7). The NE and EL samples display highest numbers of shared ROHs between two or more individual cattle. Also, BRI (except WP and KC breeds), HF and GW breeds compare to majority of southern European cattle breeds display higher numbers of shared ROH segments.

2.4 Discussion

The inherent ascertainment bias in SNP arrays (Matukumalli et al., 2009) can result in an overestimation of the homozygosity if applied to populations not included in the design of the SNP array. However, the Illumina BovineHD Genotyping BeadChip was designed from a diverse range of breeds comprising zebu, taurine and crossbred individuals. We therefore, do not expect primitive samples to exhibit a high level of ascertainment bias. Small sample size per breed may distort breed allele frequencies in ADMIXTURE analysis, thus, we only considered the initial K values (from 2 to 4). The validity of D-statistics depends on the number of genetic markers but does not require a large sample size per population (Patterson et al., 2012).

2.4.1 Population origin, admixture and history

The differentiation of indicine and taurine cattle as suggested by PCA and ADMIXTURE is a likely consequence of separate domestication events that occurred within the Neolithic cultures of the Indus Valley and the Fertile Crescent, respectively (Loftus et al. 1994). The PCA and ADMIXTURE analysis also indicated significant genetic differentiation of north-western

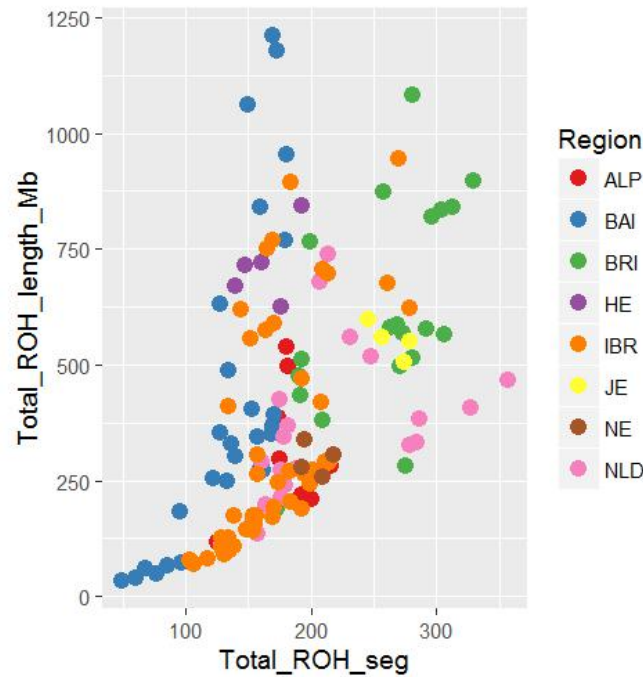


Figure 2.4: The plot displays total number of ROH segments and total ROH size (in Mbp) for all genotyped individuals. Many BAI and IBR samples display a total length comparable to samples from the BRI population. Abbreviations: ALP-Alpine, BAI-Balkan and Italy, BRI-British and Irish, HE-Heck, IBR- Iberian, JE- Jersey, NE-Nelore, NLD-Dutch.

(BRI and NLD), IBR and BAI cattle populations, which is in agreement with previous findings (Cymbron et al. 2005; Decker et al. 2009; Gautier et al. 2010; Edwards et al., 2011). This north-south difference has been hypothesised to reflect migrations of agriculturalists during the Neolithic transition along the Mediterranean coasts and the Danube river, respectively (Cymbron et al. 2005; Feliu et al., 2014).

The divergence of the BAI cattle as suggested by PCA (Figure 2.2c) can be attributed to an indicine genomic component which is identified in the ADMIXTURE (Figure 2.3) and D-statistics (Table 2) analyses. By analysing the genome-wide SNP markers, McTavish et al. (2013) and Decker et al. (2014) also reported an indicine influence on Italian cattle breeds. Using whole genome sequences of ancient human DNA, Jones et al. (2015) and Haak et al. (2015) suggested massive migration of Yamnaya steppe herders as a source of dispersion of Indo-European languages to both northern-central Europe and India. These herders might also have mediated gene flow between Indian zebu and Ukrainian steppe cattle. Contrary to the ADMIXTURE analysis, the three population tests failed to display evidence for an indicine genetic component in MT and BK which can be the result of substantial post-admixture drift in these breeds. Investigating population history of Indian population using genome-wide SNPs, Reich et al. (2009) also reported the failure of three population tests to detect admixture when applied to populations with substantial genetic drift since the admixture.

Given the high genetic diversity in near Eastern cattle population (Utsunomiya et al., 2014), we

expect low ROH abundance in the genome of BAI cattle due to its proximity to the near East (centre of cattle origin) and presence of an indicine component. The result is in concordance with the expectation, as BAI cattle display low number of ROH segments (Figure 2.4). This pattern of diversity is also seen in humans and pigs, where a relationship exists between ROH abundance and distance to source of origin (Kirin et al., 2010; Bosse et al., 2012). Several BAI breeds such as Busha (BU02), Maltese (MT) and Boskarin (BK), however, display high cumulative ROH length manifested as long ROH (Figure 2.5c) indicating inbreeding due to recent reduction of the effective population size.

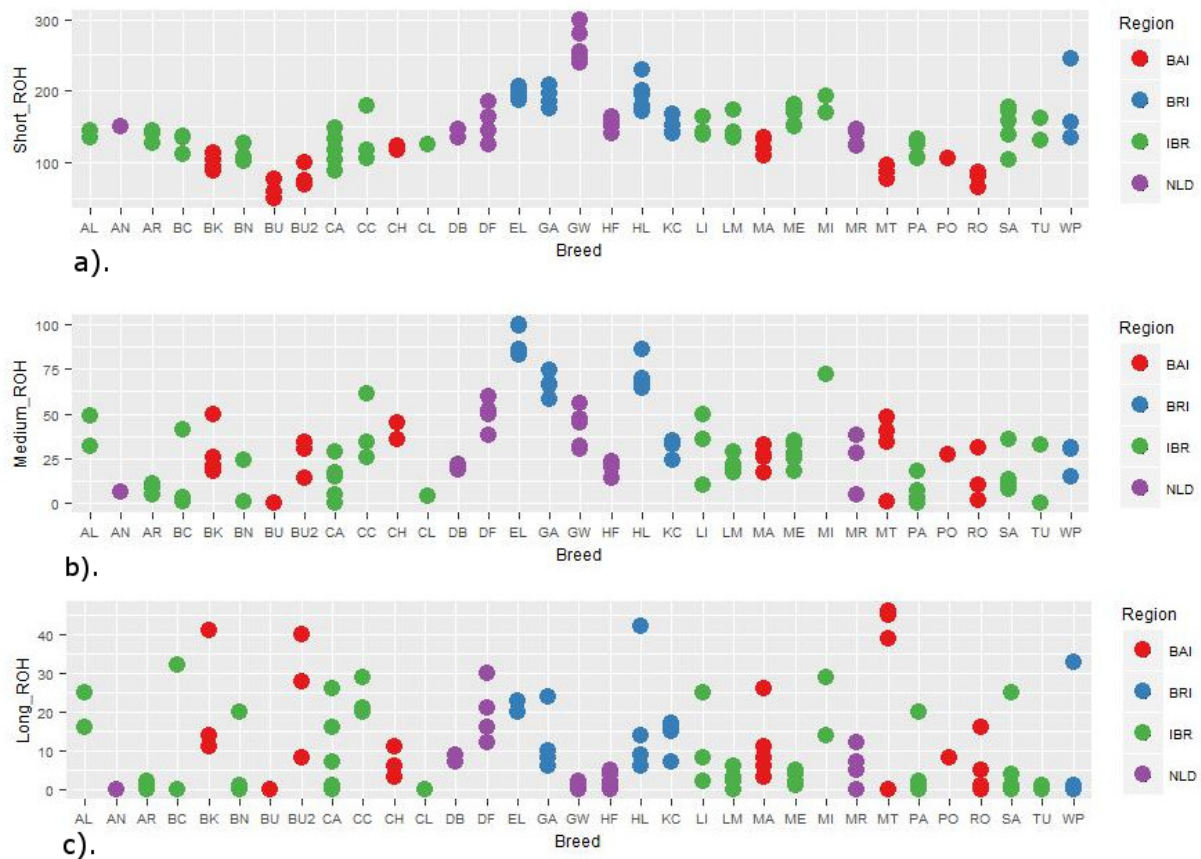


Figure 2.5: The plot displays breed-wise distribution of ROH segments for each individual in three size classes: a) short ROH (0.5-2Mbp); b) medium ROH (2Mbp-8Mbp) and c) long ROH(>8Mbp). Abbreviations: ALP-Alpine, BAI-Balkan and Italy, BRI-British and Irish, IBR- Iberian, NLD-Dutch.

The separate position of the dense IBR cluster in the PCA (Figure 2.2), short branches in the NJ network (Figure S1), and a high number of significant three population tests (Table S3), all suggest a high level of gene flow between IBR cattle breeds which may have resulted in high genetic similarities and low divergence among IBR breeds. These results probably reflect the geographic isolation of the Iberian Peninsula, and the relatively recent (after 1920) introduction of herd books for Iberian cattle breeds (Feliu et al., 2014). Under this scenario, the formal definition of local Iberian breeds may have been an afterthought. Analysing BovineHD SNP

array data, Cañas-Álvarez et al (2015) also reported low divergence among Spanish beef cattle breeds which they attributed to local admixture.

Because of the high admixture among local Iberian cattle identified in the present study and the presence of African cattle ancestry in the genomes of Iberian cattle as reported previously (Decker et al., 2014), we expect low ROH abundance across the genomes of IBR cattle. The result was in concordance with the expectation as IBR cattle display low ROH abundance compared to BRI cattle. Exceptions were the Cachena (CC) and Mirandesa (MI) (Table S6) as well as several individuals of other breeds (Figure 2.5) with high cumulative ROH length as well as high number of ROH. We propose that this reflects a breed management that does not always avoid crossing of related animals. This may be typical of local populations with a small geographical range. Analysing genetic diversity in Iberian cattle breeds using microsatellite markers, Ginja et al. (2010) also reported low genetic diversity in Mirandesa breed.

The similarity of productive NLD and the more primitive BRI breeds as suggested by the PCA (Figure 2.2) and ADMIXTURE analysis (Figure 2.3) may reflect the export of Dutch sires to England since the 16th Century (Feliuss et al., 2014). The three-population tests (Table S3) suggest a well differentiated breed structure for the majority of the BRI and NLD breeds. However, post-admixture genetic drift in the population can mask the signal of the test (Patterson et al., 2012) which is likely in the case of Heck cattle, a complex admixed population derived from several diverse European cattle breeds (Van Vuure, 2005), in particular, the Hungarian grey, which can explain the minor signature of zebu admixture evident in the ADMIXTURE results. The separate position of the Jersey (JE) breed in the PCA plot (Figure 2.2c) indicates its virtually complete isolation since 1763 (Feliuss, 2007).

The fact that cumulative ROH size (Figure 2.5) is dominated by an abundance of short to intermediate ROHs (0.5-8 Mbp) in NLD as well as BRI breeds indicates the pattern of an isolated population based on a source population of substantial size. An abundance of ROH due to genetic isolation has also been reported in other species such as in Japanese Wild boar (Bosse et al., 2012). The low abundance of long ROH in NLD breeds (Figure 2.5), particularly in GW and HF breeds, reflects a breed management that is successful in avoiding crossing of related animals probably by maintaining a sufficient effective population size. Isolation of breeds with a relatively small population size increases the probability of inheritance of identical segments as evident clearly in British and Irish region cattle breeds (English Longhorn (EL), HL (Highland) and Galloway (GA)). On the other hand, breeding systems involving intensive culling, usage of few elite bulls and selection in commercial populations may also contribute to increase the overall homozygosity and produce ROH patterns similar to those seen as a result of a bottleneck (Kim et al., 2013). Thus, it is difficult to disentangle the effect of a bottleneck from selection or vice-versa. Hence, high sharing of ROH haplotypes (Table S7) as detected between the individuals within BRI breeds can also be the result of confounding process like selection or a bottleneck. The predominant short ROH fragments in EL, HL and KC also indicate a low degree of recent consanguinity. The low genetic diversity of the BRI breeds as observed in terms of high cumulative ROH is in line with previous studies using microsatellite, AFLP and SNP markers (Cymbron et al., 2005; Purfield et al., 2012; European Cattle Genetic Diversity Consortium, 2006; Utsunomiya et al., 2014).

2.4.2 Local aurochs' gene flow in the gene pool of domestic cattle

The D-statistics (Table 1 and Table S5) showed the highest frequency of aurochs derived alleles in BRI cattle. Conversely, the lowest number of aurochs-like derived alleles is reported in BAI cattle. One of the most likely explanations for this result is gene flow from British aurochs (or continental European aurochs from the same meta-population) into the ancestors of north-western and southern-central (IBR-ALP) European cattle breeds. Several scenarios have been put forward for post-domestication hybridization between wild aurochs and the ancestors of extant domestic cattle breeds (Götherström et al., 2005; Beja-Pereira et al., 2006; Achilli et al., 2008; Schibler et al., 2014). Using 15,498 SNP markers derived from the same aurochs WGS data used here, Park et al. (2015) also provided evidence for gene flow from aurochs into the ancestors of north-western European cattle breeds. Our study, in addition to north-western European breeds, also include various breeds from the Iberian Peninsula and report a high frequency of aurochs-specific derived alleles in the Iberian cattle compared to central European and Balkan-Italian cattle. Archaeological findings from the Mesolithic period also indicate a higher concentration of wild aurochs in north-western Europe than southern Europe. A large proportion of aurochs' fossils has been recovered from Britain indicating high concentration of wild aurochs in Britain during the Mesolithic period (Wright, 2013). Additionally, Anderung et al. (2005) also identified P mitochondrial haplotypes in a Bronze Age aurochs sample excavated from northern Spain, indicating a wide geographic distribution of aurochs. On the other hand, the low frequency of aurochs-specific derived alleles in the BAI cattle may indicates no or very limited contact between the ancestor of the BAI cattle and the British aurochs. Our results, however, do not rule out the possibility of gene flow from different aurochs populations with T haplotypes into the ancestors of the extant BAI cattle.

Further analysis of ancient DNA samples of aurochs and primitive cattle populations from across Eurasia is needed to thoroughly examine aurochs gene flow into ancestors of extant cattle, and to identify the population genomics processes that accompanied domestication. Our data suggest that this is also relevant for further understanding the genetic variation of the present European cattle breeds.

2.5 Data archiving

Data available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.f2d1q>.

2.6 Conflict of interest

The authors declare no conflict of interest.

2.7 Acknowledgement

MR Upadhyay benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate ‘EGS-ABG’. We would like to thank Bert Dibbits, Animal Breeding and Genomics Centre, Wageningen University for the DNA extraction and Dr KM (Kyle) Schachtschneider, Animal Breeding and Genomics Centre, Wageningen University for his critical remarks on overall format of manuscript. We would also like to thank Goran Andersson, Department of Animal Breeding and Genetics, Swedish Institute of Agricultural Sciences for his comments on the draft and Professor Mark P Brincat, Department of Obstetrics and Gynaecology, Mater Dei University Hospital, Malta, for providing us samples of Maltese breed.

2.8 Consortium members

PA Marsan¹⁰, V Balteanu¹¹, S Dunner¹², JF Garcia¹³, C Ginja¹⁴, J Kantanen^{15,16}

¹⁰Istituto di Zootecnica, Università Cattolica del Sacro Cuore, Piacenza, Italy; ¹¹University of Agricultural Sciences and Veterinary Medicine, Faculty of Animal Science and Biotechnologies, Cluj-Napoca, Romania; ¹²Facultad de Veterinaria, Universidad Complutense de Madrid, Madrid, Spain; ¹³Departamento de Apoio, Produção e Saúde de Animal, Faculdade de Medicina Veterinária de Aracatuba, UNESP - Univ Estadual Paulista, Aracatuba, São Paulo, Brazil; ¹⁴CIBIO-InBIO–Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Campus Agrário de Vairão, Vairão, Portugal ¹⁵ University of Eastern Finland, Kuopio, Finland; ¹⁶University of Eastern Finland, Kuopio, Finland.

2.9 Corrigendum

Correction to: *Heredity* (2017) 118, 169–176; doi:10.1038/hdy.2016.79; published online 28 September 2016 Following the publication of this article the authors have noticed an error in the code used for three-population tests which led to change in the results, in particular, for Iberian cattle (revised Supplementary Materials are provided with this correction). Therefore, the below paragraphs should be amended:

In the RESULTS section, the fourth paragraph should be changed from: Significant negative f_3 statistics with NE zebu as one of the reference populations indicates for the BAI breeds (Busha (BU), Chianina (CH) and Maremana (MA)) a zebu ancestry (Supplementary Table S2). A high number of significant negative f_3 statistics (Supplementary Table S3) suggests admixture between regional cattle as is observed for all IBR breeds (except Maronesa). The majority of IBR breeds, Chianina (CH), Jersey (JE), White Park (WP), Fleckvieh (FL), Dutch Belted (DB), Dutch Friesian (DF), MRY and Brown Swiss (BS) exhibits evidence of aurochs ancestry

(at least one significant f_3 test with aurochs as one of the reference populations, Supplementary Table S4). Non-significant f_3 statistics indicate isolated breed structures for the majority of the BRI and NLD breeds (Supplementary Table S4).

To: Significant negative f_3 statistics with NE zebu as one of the reference populations indicates a zebu ancestry for the BAI breed Busha (BU) (Supplementary Table S2). A high number of significant negative f_3 statistics (Supplementary Table S3) suggests admixture between regional cattle, as is observed for some IBR breeds. Only one IBR breed displayed aurochs ancestry. Non-significant f_3 statistics indicate isolated breed structures for all of the BRI and NLD breeds (Supplementary Table S4). In the DISCUSSION section, result will slightly change from: The three-population tests (Supplementary Table S4) suggest a well differentiated breed structure for the majority of the BRI and NLD breeds. To: The three-population tests (Supplementary Table S4) suggest a well differentiated breed structure for all the BRI and NLD breeds.

It is worth noting that no conclusions have been drawn solely based on the three-population tests. Overall interpretations are also based on PCA, ADMIXTURE, distance based phylogenetic tree, D-statistics and Runs of homozygosity. Hence, these changes do not affect any of the central messages of the paper. All the analysis carried out in the paper remains perfectly reproducible. The authors wish to apologize for any inconvenience caused.

2.10 Supplementary information

Supplementary information can be found at <https://www.nature.com/articles/hdy201679#supplementary-information>

Chapter 3

Deciphering the pattern of genetic admixture and diversity in southern European cattle using Genome-wide SNPs

M.R. Upadhyay^{1,2,*}, C. Bortoluzzi¹, M. Barbato³, P.A. Marsan³, L. Colli³, C. Ginja⁴, T.S. Sonstegard⁵, M. Bosse¹, J.A. Lenstra⁶, M.A.M. Groenen¹ and R.P.M.A. Crooijmans¹

¹ Animal Breeding and Genomics, Wageningen University & Research, Wageningen, The Netherlands. ² Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden. ³ Department of Animal Science Food and Nutrition – DIANA, Nutrigenomics and Proteomics Research Centre – PRONUTRIGEN, and Biodiversity and Ancient DNA Research Centre – BioDNA, Università Cattolica del Sacro Cuore, Piacenza, Italy. ⁴ CIBIO-InBIO—Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Vairao, Portugal. ⁵ Acceligen, Recombinetics, Saint Paul, MN, United States. ⁶ Faculty of Veterinary Medicine, Utrecht University, Utrecht, Netherlands.

The manuscript is accepted for publication in Evolutionary Application.

Abstract

The divergence between indicine cattle (*Bos indicus*) and taurine cattle (*Bos taurus*) is estimated to have occurred $\sim 250,000$ Years ago, but a small number of European cattle breeds still display shared ancestry with indicine cattle. Additionally, following the divergence of African and European taurine, the gene-flow between African taurine and southern European cattle has also been proposed. However, the extent to which non-European cattle ancestry is diffused across southern European cattle has not been investigated thoroughly. Also, in recent times, many local breeds have suffered severe reductions in effective population size. Therefore, in the present study, we investigated the pattern of genetic diversity in various European cattle based on single nucleotide polymorphisms (SNP) identified from whole genome sequencing data. Additionally, we also employed unlinked and phased SNP based approaches on high-density SNP array data to characterize non-European cattle ancestry in several southern European cattle breeds. Using heterozygosity-based parameters, we concluded that, on average, nucleotide diversity is greater in southern European cattle than western European (British and commercial) cattle. However, an abundance of long runs of homozygosity (ROH) and the pattern of Linkage disequilibrium decay suggested recent bottlenecks in Maltese and Romagnola. High nucleotide diversity outside ROH indicated a highly diverse founder population for southern European and African taurine. We also show that Iberian cattle display shared ancestry with African cattle. Furthermore, we show that Podolica is an ancient cross-bred between Indicine zebu and European taurine. Additionally, we also inferred similar ancestry profile of non-European cattle ancestry in different Balkan and Italian cattle breeds which might be an indication of the common origin of indicine ancestry in these breeds. Finally, we discuss several plausible demographic scenarios which might account for the presence of non-European cattle ancestry in these cattle breeds.

Keywords: genetic diversity, SNPs, haplotype, admixture, cattle, indicine ancestry, southern European, African taurine

3.1 Introduction

Modern cattle originated from at least two different species of wild aurochs: *Bos primigenius primigenius* (European aurochs) and *Bos primigenius namadicus* (Indian aurochs). Analyses based on mitochondrial DNA (mtDNA) have estimated the divergence date between these two species from ~117,000 to ~332,400 years before present (YBP) (Achilli et al., 2008; Bradley, Machugh, Cunningham, Loftus, 1996; Loftus, Machugh, Bradley, Sharp, 1994). Subsequently, the near Eastern population of *Bos. p. primigenius* was domesticated ~10,000 YBP giving rise to domesticated taurine population (Troy et al., 2001), while *Bos. p. namadicus* was domesticated ~8,000 YBP, somewhere in the Indus valley giving rise to the modern zebu (cattle with hump aka indicine) population (Chen et al., 2010). The occurrence of a third domestication event involving African aurochs in north-eastern Africa is still debated (Grigson, 1991; Loftus et al., 1994; Pitt et al., 2018). However, recent studies based on genome-wide SNPs refuted this hypothesis of third domestication centre and instead, proposed the gene flow from African aurochs resulting in high divergence of African taurine (Pitt et al., 2018; Decker et al., 2014). This hypothesis of African aurochs introgression, however, is yet to be tested because DNA from an ancient specimen of African cattle is not available. Conversely, analyses of genome-wide SNPs from ancient European aurochs have provided novel insights into the post-domestication contribution of ancestral aurochs to the gene pool of modern European cattle. For instance, studies have suggested that British cattle followed by Iberian and Dutch cattle may carry an abundance of aurochs specific alleles compared to either African or near Eastern taurine breeds (Park et al., 2015; Upadhyay et al., 2016). More aurochs samples from different parts of Europe and at different time points, however, are needed for further validation of these results.

Based on the archaeological evidences (Martins et al., 2015; Zilhao, 1993) and genomic analysis of DNA sequences of ancient farmers (Hofmanová et al., 2016), at least two migration routes have been suggested to explain the expansion of Neolithic farmers along with their domesticated animals after early domestication of European aurochs in the Near East. Following these early migrations, various instances of gene flow involving non-European cattle into the gene pool of southern European cattle have been suggested. For instance, mtDNA, microsatellite markers and genome-wide SNP based studies have shown the presence of African taurine ancestry in Iberian breeds (Beja-Pereira et al., 2003; Cymbron et al., 2005, Ginja et al., 2010a, 2010b; Decker et al., 2014). On the other hand, several Italian breeds also carry complex non-taurine ancestry. For instance, analysing microsatellite markers, Cymbron et al., (2005) reported the highest frequency of indicine population-associated alleles in Greek and Italian cattle breeds, while Decker et al., (2014) reported the dual ancestry—African taurine as well as indicine—in three Italian cattle breeds namely, Chianina, Romagnola and Marchigiana. However, the origin of such dual ancestry remains unclear. Moreover, it is also not known whether other Balkan and Italian (BAI) cattle breeds such as Busha and Maremmana, also carry the similar dual ancestry. Further, the genetic admixture pattern in southern European cattle has mostly been investigated using either mitochondrial DNA (Di Lorenzo et al., 2018; Pellicchia et al., 2007) or microsatellite markers (D’Andrea et al., 2011). Conversely, genome-wide high-density SNP markers have scarcely been used for detailed characterization of the non-European ancestry in

southern European cattle.

The European indigenous native cattle breeds are valuable genetic resources as they are well adapted to local environments. For instance, Maremmana became well adapted to the hot and humid environment of Tuscan Maremma plain, once wetlands where malaria was endemic. Further, it has been postulated that some local breeds like Busha might have conserved an abundance of rare alleles because of their large effective population sizes (Medugorac et al., 2009). Indeed, conservation analyses have prioritized some local European cattle for conservation, namely from Iberia (Canon et al., 2001; Catarina Ginja et al., 2013). Additionally, in some cases, the certified products obtained from local breeds provide an additional value that distinguishes them from non-native breeds (Di Trana et al., 2015). Furthermore, local breeds are also attached to several traditions of cultural heritage. The rapid decline of local breeds, however, remains a major concern. It has been estimated that of the 640 global cattle breeds with known “risk status”, 171 breeds can be classified under “at risk” category while 184 breeds are already extinct (FAO, 2015). In our previous study, we also reported the recent reduction in effective population size for several southern Europe breeds such as Mirandesa and Maltese (Upadhyay et al., 2016). Therefore, a comprehensive understanding of the status of current genetic diversity and demographic processes driving these changes will have a large impact on their ongoing conservation efforts.

In this study, genome-wide SNPs data were generated using two different approaches: BovineHD SNP array genotyping and whole genome sequencing (WGS). At first, we used the genotypes obtained from WGS of European cattle to assess genetic heterozygosity and change in recent demography, followed by identification and assessment of non-European ancestry in southern European cattle (Iberian, and Balkan and Italian cattle) using BovineHD SNP array data.

3.2 Material and methods

3.2.1 Whole genome sequencing data, alignment, and variant calling

Blood, hair roots or semen samples were collected from twelve individuals from Balkan and Italian cattle breeds (one Boskarin, two Busha, one Chianina, four Maltese, one Italian Podolica, and three Maremmana), and seven individuals from Iberian cattle breeds (one Limia, one Maronesa, two Pajuna, two Sayaguesa and one Tudanca). DNA was extracted either using QIAamp DNA blood spin kit (Qiagen Sciences) or [®] Blood & Tissue Kit (Qiagen Sciences). DNA quantification and qualification were carried out using Qubit 2.0 fluorometer (Invitrogen). Library construction was carried out with 0.5-3 μ g of genomic DNA following the Illumina library prepping protocols (Illumina Inc.).

All the 19 individuals were paired-end re-sequenced with the Illumina sequencing technology (Illumina Inc.). We also obtained additional 18 WGS data (15 raw sequenced data and 3 WGS alignment) of several commercial and traditional cattle from previous studies (Bickhart et al., 2016; Daetwyler et al., 2014; Kim et al., 2017; Murgiano et al., 2014). All the detailed sample

information is given in Table S1. To perform the quality-based trimming on each fastq file, sickle (Joshi and Fass et al., 2011) was run with the default settings excepts for the length threshold of 50 bp. Following this trimming, BWA-mem (Heng Li, 2013) algorithm was used to align the quality-trimmed fastq files against the bovine reference genome build UMD 3.1. After the alignment, duplicate reads were removed from the bam files using “Samtools rmdup” (H. Li et al., 2009). Finally, “RealignTargetCreator” and “IndelRealigner” arguments as implemented in Genome analysis toolkit 3.1 (GATK) (H. Li et al., 2009) were used to perform local read realignments.

Multi-sample variant calling was carried out using freebayes (Garrison Marth, 2012) with the default settings except for the parameters minimum base quality (`-min-base-quality 20`) and haplotype length (`-haplotype-length 0`). Under the default parameters, freebayes only considers the alternate allele as a SNP if it is covered by at least two reads or present in at least 0.2 fractions of the total reads aligned at a position in at least one sample. Following this early round of SNP calling, vcftools (Danecek et al., 2011) was used to discard variants using these criteria: 1).indel positions, 2). SNP variants with more than two alleles. 3). SNP variants with genotyping quality less than thirty, 3). SNP with a minimum depth less than four and a maximum depth greater than thirty (avoid erroneous genotypes due to CNV). The concordance between genotypes as called from the whole genome sequencing dataset and BovineHD array dataset was examined using seven samples for which both genotype sets were available.

3.2.2 Estimation of heterozygosity using whole genome sequences

The genetic heterozygosity was estimated in each individual whole genome sequence using mlRho (Haubold et al., 2010). The method implemented in mlRho co-estimates the population mutation rate (θ) and sequencing error rate (ϵ) using Maximum likelihood approach. If the value of θ is small, the estimate of theta approximately reflects heterozygosity under an infinite allele model. For this analysis, we used UMD 3.1 aligned data with mapping quality>15, base quality >25 and sites with the depth between 4 and 30.

3.2.3 Assessment of recent demographic change using Runs of homozygosity (ROH) analysis

Runs of homozygosity (ROHs) were extracted from whole-genome sequence data following the procedure implemented as in Bosse et al., (2012), using sliding windows approach of 10 kbps. We defined an ROH as a genomic region of at least 20 kbps where the number of heterozygous SNPs per bin (or SNP count) was less than 0.25 times the whole-genome nucleotide diversity, and at least 20 consecutive bins showed a total SNP average lower than the total genomic average. Local assembly or alignment errors were minimized by relaxing the threshold for individual bins within a candidate homozygous stretch, allowing the number of SNPs per bin to be maximum twice the genomic average and the average SNP count within the candidate ROH to not exceed 2/3 the genomic average. Both whole-genome nucleotide diversity and nucleotide diversity within a

candidate ROH were estimated from well-covered sites only, which were defined by a depth of coverage between 4X and 30X.

3.2.4 BovineHD genotyping array data and filtering

BovineHD SNP genotyping array data of various European, African and Indian cattle as published in the previous studies (Bahbahani et al., 2017; Upadhyay et al., 2016; Barbato et al., unpublished) were obtained and merged for the top alleles using PLINK (Purcell et al., 2007). For the present study, several additional animals of various Italian and Iberian cattle breeds were genotyped using BovineHD SNP array. The merged BovineHD genotyping array data was filtered using PLINK (Purcell et al., 2007) to keep the animals with more than 90% of genotypes called ($-mind\ 0.1$), SNPs that were present across at least 95% of the samples ($-geno\ 0.05$), and SNP with minor allele frequency $\geq 1\%$ ($-maf\ 0.01$). The complete total BovineHD genotyping dataset consists of 670k SNPs and 358 samples (Table S2).

3.2.5 Population admixture using the unlinked SNPs

Genetic admixture patterns of southern European cattle breeds in relation to non-European taurine and zebu were characterised using: 1). model based clustering as implemented in ADMIXTURE (Alexander, Novembre, Lange, 2009), 2). treemix analysis (Pickrell Pritchard, 2012), 3). three population test (Reich, Thangaraj, Patterson, Price, Singh, 2009). The ADMIXTURE analysis was carried out with 1,000 bootstrap replicates for population cluster (K) values from 2 to 6. Prior to the ADMIXTURE analysis, a linkage disequilibrium (LD) pruning was performed, to reduce the overall pairwise LD to <0.10 . We used the python package pong to generate the figures based on ADMIXTURE result (Behr, Liu, Liu-Fang, Nakka, Ramachandran, 2016). Treemix analysis (Pickrell Pritchard, 2012) was carried out to investigate the migration events of zebu and African cattle during the domestication history of European taurine. In brief, we followed the procedure of Decker et al., (2014): first we generated maximum likelihood based phylogenetic tree of all cattle populations, and iteratively, we added one migration edge to the previously generated graph with “m” migration edge. We rooted the graphs with yak (as an outgroup), used blocks of 1,000 SNPs, and applied the “-se” option to estimate standard errors of migration proportions. For this analysis, the scaffolds of yak genome assembly (Hu et al., 2012; Qiu et al., 2012) was aligned to the bovine UMD 3.1 assembly and processed as described in Upadhyay et al., (2016). Moreover, the SNP genotyping data of the British aurochs (Park et al., 2015) were also used in this treemix analysis. The treemix analysis was run five separate times, each time using different seeds, to assess the consistency of migration edges across different runs. Additionally, the f3 tests, which considers the correlation of allele frequencies across the genome-wide markers, was also carried out to provide support to admixture analysis. The f3 test with Z-score less than -3.0 was considered as significant. The three population tests were carried out using ADMIXtools (Patterson et al., 2012).

3.2.6 Population admixture using the phased SNPs

Phasing

We phased genotypes of each autosomal chromosome separately using Beagle 4.1 (Browning Browning, 2007) by setting all the parameters as default. The recombination map of cattle was used from the previous study (Ma et al., 2015). This recombination map comprises of 59,309 SNPs markers for 29 autosomes with an average distance of 0.043 cM in males and 0.039 cM in females. Our data, however, comprised consisted of ~670K SNPs; and hence, for an SNP with unknown recombination rate in our dataset, we assign it the value of nearest SNP with a known location in recombination map, thus, keeping the overall genetic distance between markers same as that of the consecutive markers in original recombination map (Ma et al., 2015).

3.2.7 ChromoPainterv2 to infer the coancestry matrix

To infer population admixture using the phased data, Li and Stephens (2003) algorithm as implemented in ChromoPainterv2 (Lawson, Hellenthal, Myers, Falush, 2012) was used. The underlying algorithm takes into consideration LD and underlying recombination process along the markers and reconstructs each haplotype as a recipient of a series of genetic chunks from all the other “donor” haplotypes. ChromoPainterv2 first calculates nuisance parameters, n (like effective population size) and M (population mutation rate), to implement them in a probabilistic model that calculates copying probability of genetic chunks in a recipient haplotype given the other donor haplotypes. We used SNPs located on chromosomes 1, 2, 7, and 12 to infer the “ n ” and “ M ” using 10 iterations of Expectation Maximization (EM) algorithms. The obtained values of “ n ” and “ M ” were: 307 and 0.0012 respectively. Finally, these inferred values were fixed in the algorithm to obtain the ChromoPainter coancestry matrix (count matrix as well as length matrix) that measures the haplotype sharing among the samples across all the chromosomes.

3.2.8 FineSTRUCTURE algorithm to cluster the samples

Once we obtained the count of shared haplotypes between different samples, we used the FineSTRUCTURE to cluster the samples into genetically homogeneous groups. Following Leslie et al., (2015), FineSTRUCTURE was run for 2 million Markov-Chain-Monte-Carlo (MCMC) iterations with the initial 1 million iterations discarded as “burn-in” and following these “burn-in”, sampling from the posterior distribution was carried out at every 10,000 iterations. Later, as recommended in Lawson et al., (2012), we ran 10,000 “hill climbing” iterations on the MCMC iteration with the highest posterior probability to get the final cluster assignment.

3.2.9 GLOBETROTTER to estimate the admixture proportion

The multiple linear regression model as implemented in the GLOBETROTTER algorithm was used to assess the ancestral makeup of Balkan and Italian cattle breeds in terms of ancestry contribution from various taurine and zebu clusters. In brief, we model the genome of each BAI cattle breed as a linear mixture of the taurine and zebu donor populations using the method described in Leslie et al., (2015). To estimate the standard error in these ancestry proportions, we applied delete-one chromosome jack-knife approach (Busing et al., 1999) as described in Montinaro et al. (2015).

As the estimates of admixture dates and ancestry proportions depend on the true underlying recombination rate between any two consecutive SNPs, we interpolated true average recombination rate based on the cattle recombination map constructed by Ma et al., (2015).

3.2.10 Linkage disequilibrium decay and estimation of effective population size

To assess the change in effective population size and the pattern of LD decay, the SNeP v1.11 (Barbato, Orozco-terWengel, Tapio, Bruford, 2015) software was used. The LD decay analysis was run with the following parameters: -mindist 5000, -maxdist 2000000, -numBINS 50, -maf 0.05.

3.3 Results

3.3.1 Profile of heterozygosity and Runs of homozygosity

Individual genomes of 19 southern-Eastern European cattle were sequenced and genotyped along with 18 additional whole genome sequences downloaded from NCBI SRA (Detail information in Table S1). These additional sequences were comprised of individuals from Indian zebu (two Gir), African zebu (two Kenana), African taurine (four N'Dama), British (two Belted Galloway, two Dexter, one Hereford), Italian (two Romagnola), and commercial (one HF, Jersey, and Simmental each) cattle breeds. This sampling scheme allowed to compare the genetic diversity and recent demographic changes between cattle from a wide range of geographical regions. The coverage of re-sequenced genomes across samples in the final bam files varied from $\sim 2.6 \times$ to $\sim 16 \times$ (Table S1). The alignment was performed against UMD3.1 using bwa-mem, and the alignment rate was greater than 98% for all the samples. The total number of SNPs identified after performing quality filtering in bam file and post-SNP calling was greater than 30 million. The genotyping concordance (Table S1) between SNPs genotyped using the BovineHD SNP array and SNPs genotyped using whole genome sequencing was $\sim 94\%$ for all seven samples, probably indicating the low proportion of false positive SNPs in the dataset.

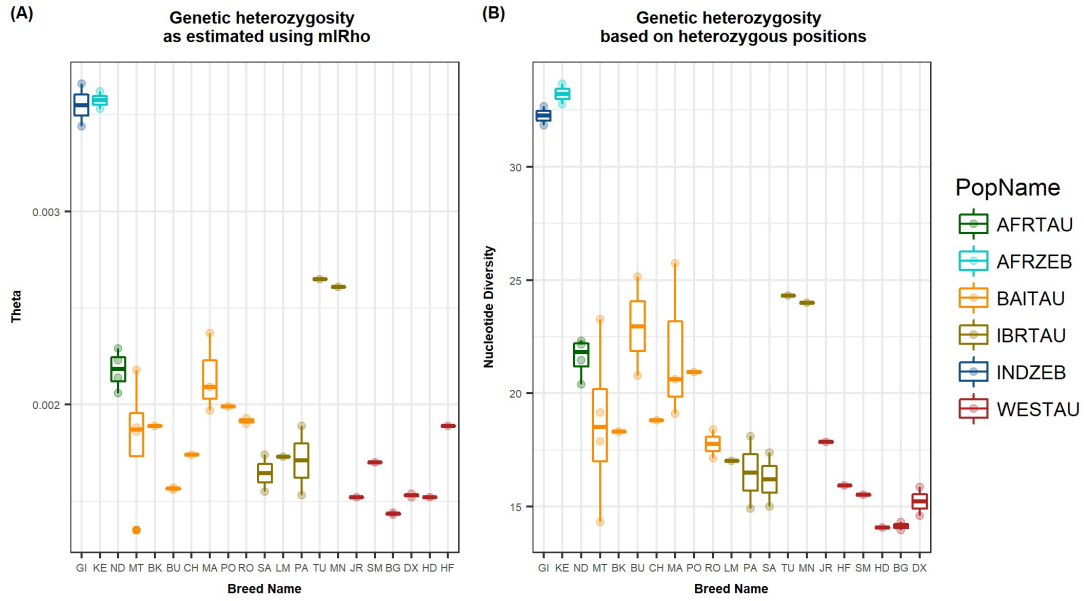


Figure 3.1: Boxplots showing average population scaled mutation rate (A) and average heterozygosity calculated in 10 Kbp window using whole genome sequencing data. Abbreviations (Population Ids): AFRTAU- African Taurine, AFRZEB- African Zebu, BAITAU- Balkan and Italian Taurine, IBRTAU- Iberian Taurine, INDZEB- Indian Zebu, WESTAU- western European Taurine. Refer to Table S1 for the breed abbreviations.

We used two different approaches to estimate heterozygosity from whole genome sequences of cattle. Both the approaches - population mutation rate (θ) as well as nucleotide diversity calculated as an average number of heterozygous sites in a 10 kbp window- indicated relatively low heterozygosity in western European cattle, while several southern European individuals displayed the levels of heterozygosity comparable to African taurine (Figure 3.1). All Maremmiana (MA) individuals consistently displayed a high value of heterozygosity in both the approaches (Figure 3.1A & 3.1B). The African zebu cattle displayed the highest values of heterozygosity, probably due to the admixed nature of their genome (Bahbahni et al., 2017).

The number and total length of ROHs in the genome did not vary sharply among cattle from different regions (Figure 3.2B). In concordance with the previously described heterozygosity estimates, several southern European individuals displayed the number and cumulative length of ROHs comparable to that of N'Dama cattle. On an average, Maltese as well as N'Dama individuals, despite the low number of ROHs in their genome, have a substantial part of their genome under ROHs indicating a significant contribution from long ROHs. This is an indication of a recent reduction in the effective population size of Maltese cattle. The commercial and British individuals displayed the highest number of ROHs as well as a high cumulative length of ROH (Figure 3.2B), probably a result of intense selection. We also estimated the nucleotide diversity outside the ROH regions (π_{out}), and the result (Figure 3.2A) consistently indicated greater values in southern European individuals compared to the west European cattle. Also, N'Dama individuals have the highest π_{out} values among all the taurine cattle. These results could indicate that southern European and African taurine cattle are descendant of diverse ancestral

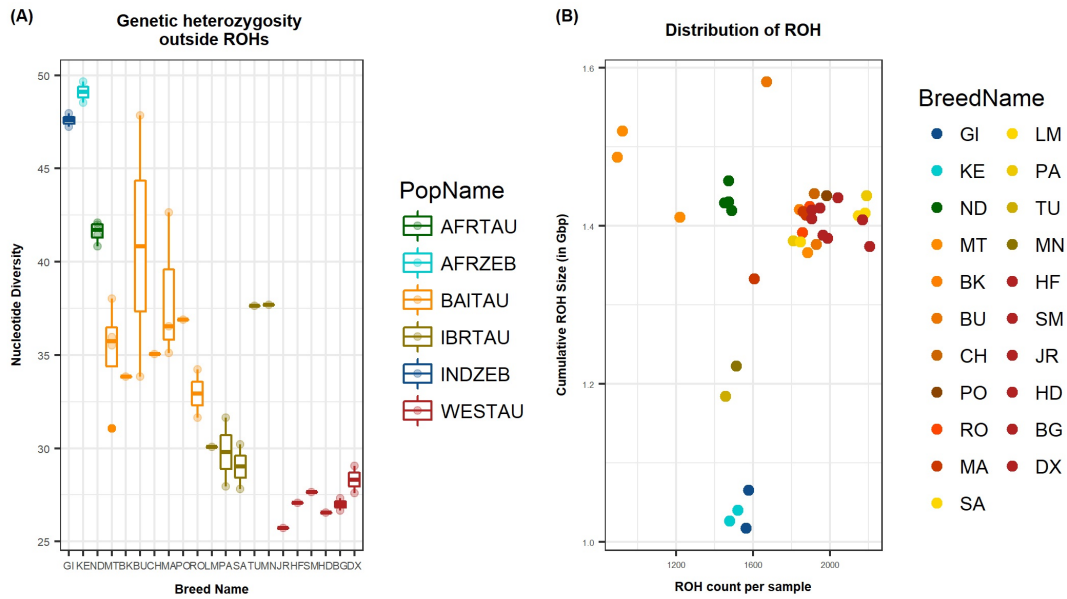


Figure 3.2: Boxplots showing average heterozygosity calculated in 10 Kbp window outside ROH (A) and distribution of ROH (B) using whole genome sequencing data. Abbreviations (Population ids): AFRTAU- African Taurine, AFRZEB- African Zebu, BAITAU- Balkan and Italian Taurine, IBRTAU- Iberian Taurine, INDZEB- Indian Zebu, WESTAU- western European Taurine. Refer to Table S1 for the breed abbreviations.

population. We note that the ROH profile (Figure 3.2B) of Busha, which is characterized by the low number and low cumulative length - might also be a consequence of their low genome coverage ($\sim 2-4 \times$).

3.3.2 Investigation of genetic admixture using unlinked SNPs

To investigate the presence of indicine and African cattle ancestry among southern European cattle, admixture analysis was initially run on the entire dataset of 358 individuals for values of K between 2 and 11. For values of K from 2 to 6, all Iberian samples display minor component of African cattle ancestry in their genome (Figure 3.3A).

Among Balkan and Italian (BAI) breeds, Busha (BU), Maremmmana (MA), Chianina (CH), and Marchigiana (MCGFH) display African as well as indicine cattle ancestry in their genome. On the other hand, at the value of K=5 and K=6, Romagnola (ROMFH) displays only single ancestry which is also one of the ancestral components in almost all European taurine samples. We hypothesized that this unique cluster of Romagnola is the result of a significant bottleneck due to a recent reduction in the effective population size. To test this hypothesis, we performed linkage disequilibrium decay analysis to estimate the change in effective population size for all the breeds with sample size more than >13 (Figure 3.3B). The analysis indicated greater historical effective population size, but slow LD decay in Italian cattle breeds compared to the commercial cattle breed HF. Such slow LD decay indicates the presence of long haplotypes which are usually

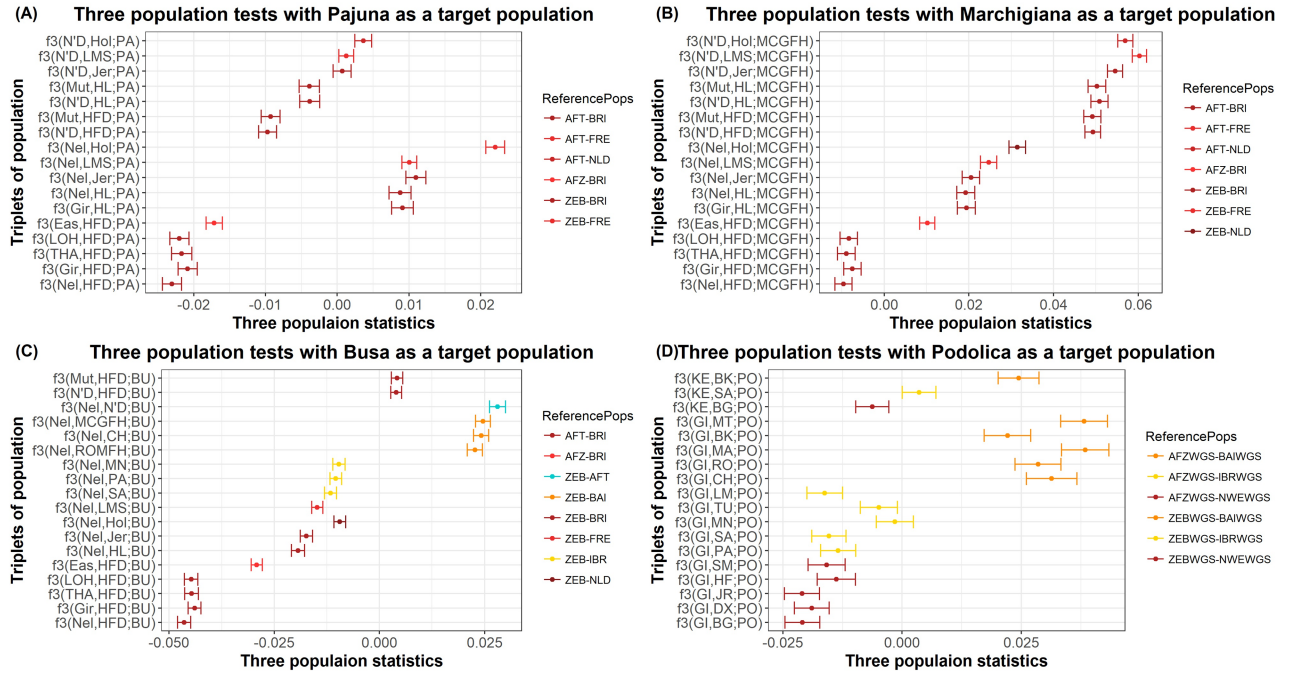


Figure 3.4: Three population tests with Pajuna (A), Marchigiana (B), Busha (C), and Podolica (D) as target populations. In case of Podolica, f3 tests were performed on whole genome sequencing data while for the remaining population f3 tests were performed on SNP array data. The dot shows f3 statistics value, while horizontal bar shows plus or minus standard error. Refer to Table S2 for the breed abbreviations.

performed three-population tests using two sets of genotypes obtained from WGS data: 1). Genotypes identified from aligning the cattle sequences against the taurine reference (UMD 3.1) and 2). Genotypes identified from aligning the cattle sequences against indicine reference (*Bos indicus*.1.0). Interestingly, across both genotype sets, only a Podolica sample generated significant negative z-value for f3 tests with British, Dutch or Iberian cattle as one and Gir as another reference population (Figure 3.4D).

To investigate the cattle phylogeny and relationship across Indicine, African and taurine lineages, the phylogenetic tree was constructed by applying the maximum likelihood (ML) algorithm as implemented in the software treemix. The ML based phylogenetic tree (Figure 3.5) topology perfectly captures known relationships among taurine and zebu cattle populations. Using Yak as an outgroup, the tree displays the first split between taurine and zebu cattle which is followed by a split between domestic taurine and wild British aurochs. African taurine appears to be the most divergent among all domestic taurine, while Balkan and Italian breeds appear to be paraphyletic. All Iberian breeds display short branch lengths and form a single clade.

To investigate the migration events involving indicine and African cattle in southern European cattle, we run five independent treemix analysis, each initialized with different random seeds. Across all the runs, the graph model without any migration edges explained about 98.73% of the

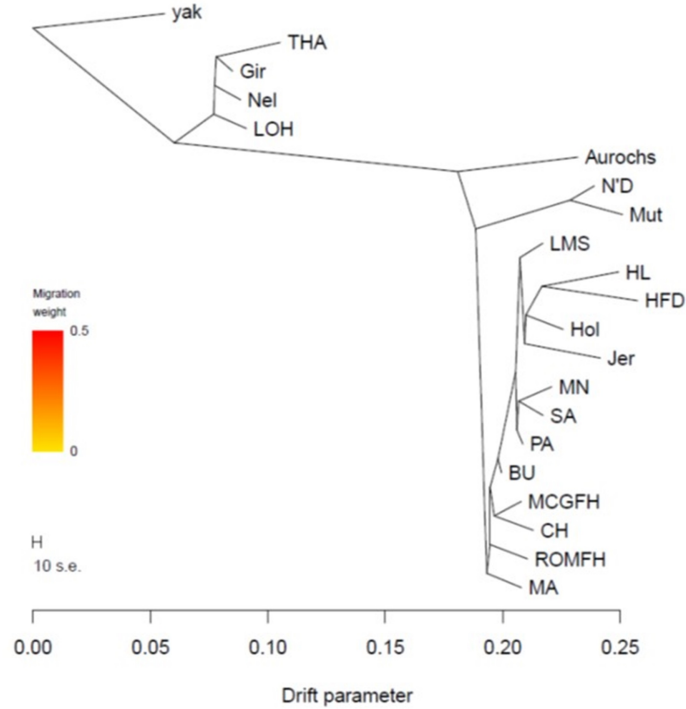


Figure 3.5: Maximum Likelihood based phylogenetic tree. Scale bar shows 10 times the average standard error of the estimated entries in the sample covariance matrix. Refer to Table S2 for the breed abbreviations.

total variance (Figure S1) in relatedness between populations meaning that adding migration events would improve the fit within a graph. Across all the treemix runs (Figure S2), we either observed migration edges between African taurine and Iberian breeds, or the same clade for African taurine and Iberian samples. Interestingly, all BAI breeds displayed inconsistent migration edges across all the different treemix runs (Figure S2). Decker et al., (2014) also reported inconsistent placement of Italian breeds across the different treemix runs, probably as the result of an underlying complex phylogeny.

3.3.3 Investigation of admixture pattern using phased SNPs

The pattern of haplotype sharing and in-depth analysis of the complex admixture pattern of southern European cattle were investigated by ChromoPainter and fineSTRUCTURE. The pattern that emerged from ChromoPainter co-ancestry matrix (Figure 3.6 and Figure S4) as well as the clustering of samples based on the fineSTRUCTURE analysis not only reinforced the finding of the ADMIXTURE and treemix analysis but also provided additional insight into the relationships between zebu and taurine cattle. Based on the fineSTRUCTURE-inferred tree, we

classified our cattle dataset of 358 animals into the following major clusters: 1). indicine zebu, 2). African cattle, 3). southern European cattle, 4). western European cattle.

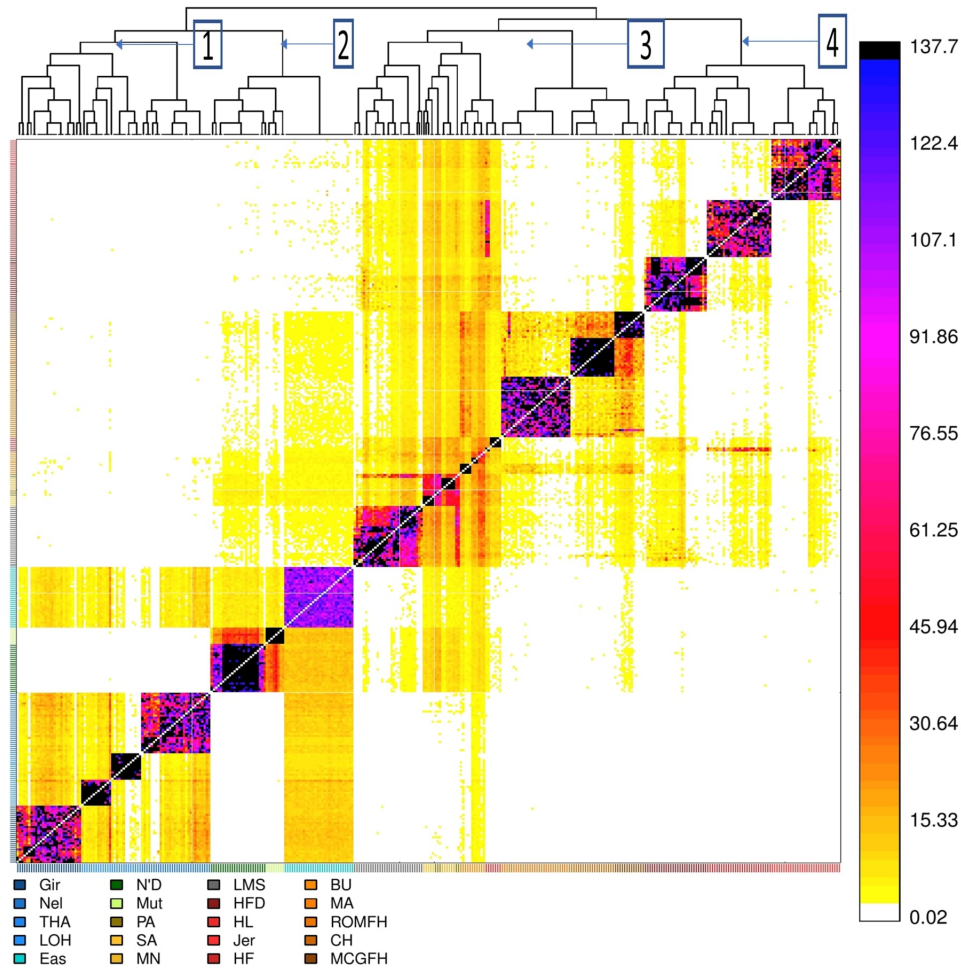


Figure 3.6: fineStructure inferred phylogenetic tree. Note that breeds are coloured according to their geographic origin. The intensity of colour indicates shared haplotypic segments (values in terms of centiMorgan). The number beside the clusters indicate: 1. Zebu cattle, 2. African cattle (N'Dama and East African zebu), 3. southern European cattle (Iberian and Italian), 4). west European cattle (Commerical and British cattle). Note that African cattle clusters with European when performed the same analysis but with smaller number of zebu samples (Figure S3). Refer to Table S2 for the breed abbreviations.

In indicine zebu cluster, we observe that Nellore (NEL) forms a separate sub-cluster from Gir, Tharparkar (THA) and Lohani (LOH) which could be attributed to the fact that Nellore is derived from south-eastern Indian zebu, while the remaining are from north-western Indian sub-continent. In African cattle cluster, we observed that N'Dama (N'D) and East African Zebu (EAZ) display a significant haplotype sharing which is consistent with the hybrid history of EAZ. It is worthwhile to note that EAZ also displays significant haplotype sharing with southern European cattle. Among the southern European cattle, Limousin (LMS) forms a cluster

with Iberian cattle (Pajuna (PA), Sayaguesa (SA), Maronesa (MN)) and they both display a significant haplotype sharing with African taurine and EAZ. The haplotype sharing pattern of Maremmana (MA) and Busha (BU) represents the mixture of a genetic component of all the cattle present in the dataset which can be attributed to their small sample size. Nevertheless, ADMIXTURE analysis also represented the Maremmana (MA) and Busha (BU) ancestry as a mosaic of all genetic ancestry of all the cattle in the dataset. However, none of the Italian cattle breeds display significant length of haplotype sharing with either African taurine or zebu except Marchigiana (MCGFH), when compared to Romagnola (ROMFH) and Chianina (CH), display evidences of slightly greater haplotype sharing with non-European cattle which could be due to differences in genetic drift after splitting from the common ancestor.

To account for the drift effect, we re-run the ChromoPainter analysis with Italian breeds only allowed to receive the ancestry from the non-Italian cattle donor groups (four clusters) assigned based on the fineStructure-inferred tree. Also, we did not consider East African zebu (EAZ) as donor population because it is the cross-bred of African taurine and indicine zebu. With this analysis, we specifically wanted to compare the ancestry profile of different BAI breeds. In particular, we note that if introgression events involving non-European cattle ancestry are ancient and split between Italian cattle is relatively recent then all Italian cattle would display similar ancestry profile. The ancestry profile of Italian cattle displays similar zebu and African taurine ancestry (Figure 3.7). The zebu and African taurine ancestry (Table S3) in these cattle breeds varied between approximately 9-12.5% (except BU02) and 8-10% respectively.

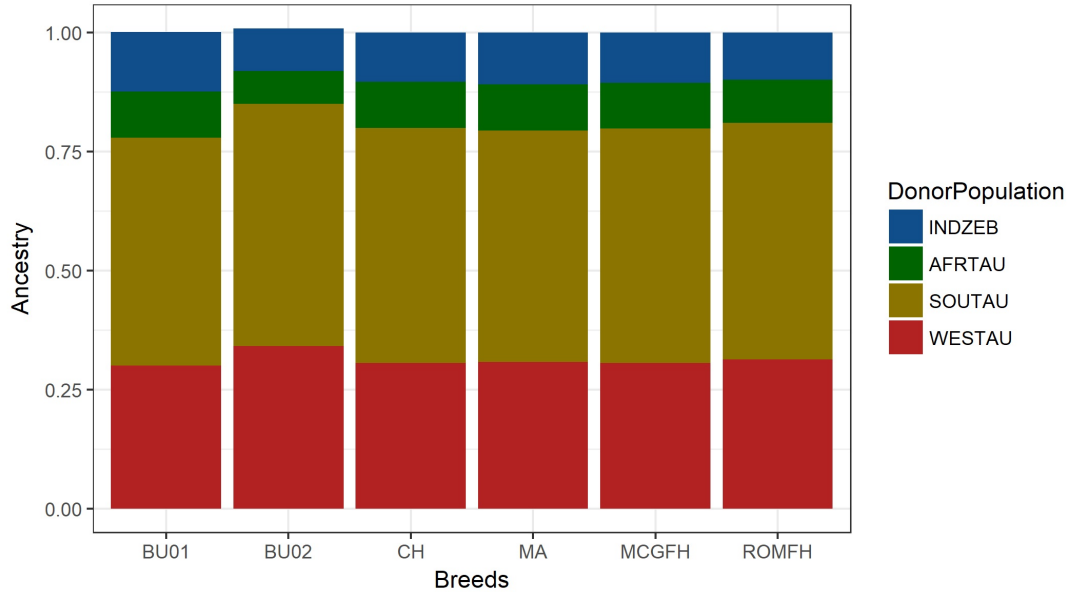


Figure 3.7: Ancestry proportion of different donor population in genome of Italian and Balkan cattle breeds. Abbreviations: INDZEB-Indian Zebu, AFRTAU- African taurine, SOUTAU-southern European taurine (without Italian cattle breeds), WESTAU: western European taurine. Note that Busha individuals come from two different sub-populations and hence, treated separately in this analysis. Refer to Table S2 for the breed abbreviations.

3.4 Discussion

3.4.1 Genetic diversity and change in recent demography

The average autosomal heterozygosity estimated using mlRho did not sharply contrast the cattle from different regions except that, on average, zebu and African cattle showed greater heterozygosity compared to the European cattle (Figure 3.1). However, differences in the estimates between different breeds were clearly visible. For instance, all Maremmana (MA), Romagnola (RO) and Podolica (PO) individuals showed greater heterozygosity compared to all western European cattle. In fact, all the estimates for Maremmana (MA) were either equal to or more than the previously reported estimates for European commercial European cattle that ranged from 1.27×10^{-3} to 1.97×10^{-3} (Gautier et al., 2016). As Intensive artificial selection or/and genetic isolation usually led to a reduction in effective population size and associated genetic diversity, the low values for the estimates were expected for western European cattle. On the other hand, high genetic diversity in Maremmana (MA), Podolica (PO) and Romagnola (RO) can be attributed to their relatively shorter history of artificial selection and greater historical effective population size compared to western European (WESTAU) cattle (Figure 3.1 and Figure 3.2). However, we also, note that LD decay analysis using SNP array markers indicates a recent effective population size in Romagnola (ROMFH) that is comparable to that of the commercial cattle (Figure 3.2). This result can be attributed to the strong decline in the effective population size of Romagnola.

The fact that Maltese (MT) displayed large between the individual variations in theta as well as in ROH profile indicates that genetic sub-structure exists in a population (Figure 3.2). This genetic sub-structure may have been formed as a result of the varying degree of Chianina admixture in a population (Lancioni et al., 2016). Also, Maltese (MT) individuals display a large proportion of genome under long ROHs indicating a strong recent bottleneck event (Figure 3.2B). This result is in good agreement with our previous study, where we observed a similar long ROH pattern using BovineHD SNP array (Upadhyay et al., 2016). The nucleotide-diversity outside ROH is a good indicator of ancestral haplotype diversity as it reflects the haplotype variations that remained weakly affected by recent consanguinity (Bosse et al., 2012). Our analysis of nucleotide-diversity outside ROH (Figure 3.2A) consistently resulted in greater values for Italian individuals compared to the rest of European samples which indicates that either the ancestral founder population for these breeds was large or/and the ancestral population received gene-flow from some genetically distinct population. Both these scenarios are likely to explain these results as Italy is much closer to the centre of domestication compared to western Europe and hence, it could be hypothesized that serial founder effect was severe in western Europe (Scheu et al., 2015) compared to Italy, while the ADMIXTURE (Figure 3.3A) and ancestry profile using ChromoPainter analysis (Figure 3.7) in this study also pointed toward presence of complex non-European cattle ancestry in Italian cattle. Similarly, the highest π_{out} values (Figure 3.2A) observed for African taurine also indicate the possibility that the ancestral population of N'Dama (ND) was much diverse compared to European taurine. Hence, the hypothesis of introgression from diverse cattle population in the genepool of African taurine cannot be ruled

out.

3.4.2 Characterizing non-European cattle ancestry

Our results from linked and unlinked SNPs based approaches, to detect admixture pattern, not only confirmed the previous reports but also extended the support for the complex origin of southern European cattle breeds. For instance, Decker et al., (2014) analysing 50k SNP markers reported the presence of complex ancestry-African taurine and zebu like-for central Italian cattle breeds (Chianina, Romagnola and Marchigiana) and they also proposed either African or Near-eastern origin for this complex ancestry. Here, we show (Figure 3.3A and Figure 3.7) that at least two other cattle breeds, namely, Maremmana (MA) and Busha (BU), also carry similar non-European ancestry as previously identified in Central Italian cattle. The ancestry profile of Busha, however, was slightly different from that of the Central Italian cattle. We note that It is also likely that Busha and central Italian cattle received a similar contribution from the same sources and subsequently evolve independently. The results of treemix analysis also indicate this possibility as in three out of five treemix runs, we observed Busha (BU) and Maremmana (MA) receiving migration edges from the same branch in phylogenetic networks. Nevertheless, our results (Figure 3.7) point toward the common origin of non-European cattle ancestry BAI cattle breeds.

At least two hypotheses can explain the origin of the complex ancestry in BAI cattle breeds: 1) the donor populations that contributed the African and indicine ancestry in BAI cattle were different (multiple introgression events), 2) or the single donor population that carried both the ancestries contributed in the gene pool of BAI cattle breeds (single event of introgression from a single donor population). We note that it is not surprising for BAI cattle to display shared ancestry with African taurine as they both likely have the same centre of domestication. However, shared ancestry between BAI and zebu due to divergence is less likely to occur, since both the lineages are estimated to have diverged about 250,000 YBP. To test the two hypotheses related to introgression, we applied the algorithm implemented in a tool, GLOBETROTTER (Hellenthal et al., 2014). In the absence of true admixing population, GLOBETROTTER can identify the most suitable proxy for the true donor (aka “surrogates”) among all the populations in the dataset. The algorithm, however, did not identify clear admixture signal involving any surrogate population in our dataset indicating that either the admixture event is very old, or the admixture events were recurring and involved multiple donor populations. Indeed, BAI cattle breeds are inhabiting in their respective regions since ancient times. Therefore, it is likely that these admixture events are too old to date using the GLOBETROTTER approach as the simulated data have shown that GLOBETROTTER can only estimate the admixture events accurately if they have occurred in about last 170 generations (Hellenthal et al., 2014). In fact, analysing microsatellite markers from a thousand-year-old bone sample, Gargini et al., (2015) reported that, compared to Iberian and western European cattle, Chianina and Romagnola were genetically closer to the thousand-year-old ancient bovine sample. Perhaps, genotyping ancient cattle bone samples from Italy will shed more light on the age and origin of the complex ancestry present in modern central Italian cattle.

Applying three population tests on SNPs identified from whole genome sequencing data, we provide clear evidence of the cross-bred (taurine and zebu) ancestry in an Italian Podolica sample. However, we note that this zebu ancestry might have not necessarily come from the Indian zebu itself; instead, it could have come from some zebu related population not sampled in our dataset. Because Podolica is believed to have originated in the Podolian region of Eastern Europe, it can be hypothesized that this Zebu ancestry is either of the Eastern European or western-Asian origins. These results confirmed the previous hypothesis of cross-bred origin for Podolica based on the similarity of the α -globin variant between Podolica and Zebu (Pieragostini, Scaloni, Rullo, Di Luccia, 2000).

Our admixture (Figure 3.3A) and ChromoPainter (Figure 3.6) results indicated shared ancestry between Iberian cattle (PA, SA and MN), Limousin (LMS) and African taurine (ND and MUT) which is in concordance with previous reports (Beja-Pereira et al., 2003; Cymbron et al., 2005, Ginja et al., 2010a, 2010b; Decker et al., 2014). Moreover, we also observed shared ancestry between east African zebu (EAZ) and southern European cattle (Figure 3.3A and Figure 3.6), however, as EAZ itself is a cross-bred of African taurine and zebu, it is difficult to interpret this shared ancestry. While the clustering of LMS with Iberian cattle probably reflects the use of this commercial breed to upgrade local Iberian cattle, namely Pajuna (Martínez et al., 2011), previous studies have also reported the presence of EAZ specific microsatellite alleles in Iberian breeds such as Mertolenga and Pajuna (Beja-Pereira et al., 2003; Martín-Burriel et al., 2011). The fineStructure-inferred tree (Figure 3.6) also clusters Marchigiana (MCGFH) with Chianina (CH) which is consistent with the known history of Marchigiana as a cross-bred with a high fraction of Chianina related ancestry.

The three population tests involving British cattle as one of the reference population resulted in relatively greater significant z-values for Busha and Podolica (Figure 3.4A and Figure 3.4D). This could mean that British cattle is the most suitable surrogate for European taurine in our dataset. We also note that the British cattle displayed the least average heterozygosity estimates (Figure 3.1 and Figure 3.2). Taken together both these results, it is safe to propose that British cattle is the least admixed population among all European cattle populations in the dataset and thus, they might have preserved a large number of European taurine-specific-variants. Interestingly, this hypothesis may also partially explain the previous reports of British cattle sharing a high frequency of derived alleles with ancient aurochs sample, while Italian cattle reportedly displayed the least frequency of aurochs specific alleles (Park et al., 2015; Upadhyay et al., 2016).

Taken together, our results in this study provided a holistic view on non-European cattle ancestry in southern European cattle. The fine-scale dissection of the demographic history and evolutionary events that have shaped the genome of modern southern European cattle, however, still demand improved methods and data.

Chapter 4

Genomic relatedness and diversity of Swedish native cattle breeds

M.R. Upadhyay^{1,2}, S. Eriksson^{1*}, S. Mikko¹, E. Strandberg¹, M.A.M. Groenen², R.P.M.A. Crooijmans², G. Andersson¹ and A.M. Johansson¹

¹Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden. ²Animal Breeding and Genomics, Wageningen University & Research, Wageningen, The Netherlands.

The draft has been submitted.

Abstract

Background

Native cattle breeds are important genetic resources given their adaptation to their local environments. However, the widespread use of commercial cattle breeds has resulted in a marked reduction in population size of several native cattle breeds worldwide. Therefore, conservation management of native cattle breeds requires urgent attention to avoid their extinction. To this end, we performed genotyping of genome-wide 150K SNPs in nine Swedish native cattle breeds to investigate the level of genetic diversity and relatedness between these breeds.

Results

We used various linked and unlinked SNP-based approaches on this data to connect demographic history with genetic diversity and population structure of these Swedish cattle breeds. Our results suggest isolation and low genetic diversity in Väneko and Ringamålako, two breeds originating from southern Sweden. Conversely, we inferred a large founder population and relatively high genetic diversity in Swedish Mountain cattle (Fjällko). Based on the shared ancestry and phylogenetic trees, we identified two major clusters in Swedish native cattle. In the first cluster of mountain cattle breeds, we found low differentiation among Swedish Mountain cattle, Unique Swedish Mountain cattle called Fjällnära cattle, and Bohus Polled cattle (Bohuskulla). The second cluster consists of breeds from southern Sweden: Väneko, Ringamålako and the Swedish Red cattle. Interestingly, we also identified sub-structures in the Fjällnära cattle breed and indicated differences in breeding practices across the farms keeping this breed.

Conclusions

This study represents the first comprehensive genome-wide analysis of genetic relatedness and diversity of Swedish native cattle breeds. We show that different demographic pattern such as genetic isolation and cross-breeding have shaped the genomic diversity of Swedish native cattle breeds. We also show that Swedish mountain native cattle breeds have retained their authentic distinct gene pool without a significant contribution from any of the other European cattle breeds they were compared within this study.

4.1 Background

Based on the types of livestock management, European cattle breeds can be broadly categorized as either commercial or traditional/native. Commercial cattle breeds are mainly used in intensive animal farming, aimed at maximizing the overall production and economic profit, for which a few popular breeds have been disseminated throughout the world (1,2). In contrast, traditional cattle breeds have a long history of adaptation to their respective environments (Medugorac et al., 2009). While commercial cattle breeds far exceed the traditional cattle breeds in terms of milk and meat production, traditional cattle breeds have a substantial cultural value, are often adapted to the environment and the climate conditions where they are held, and sometimes display a few superior production or functional traits compared with commercial cattle. For instance, Swedish Mountain cattle (Fjällko) produces milk with a superior protein composition for cheese making compared to commercial cattle such as Swedish Red cows (Poulsen et al., 2017). Also, in a study comparing the grazing pattern of Holstein Friesian with the Swedish Mountain cattle in a grass-dominated pasture area, the latter was found to travel over a larger area and to show preferences towards diverse vegetation types (Lien et al., 1999). Therefore, traditional/native cattle breeds can be valuable for grassland management with diverse vegetation.

In a recent FAO report (FAO, 2015) many cattle breeds were classified as being “at risk”. It is important to consider that this classification was based only on a fraction of all native breeds worldwide. The diversity status of about 50% of all cattle breeds globally, is currently unknown and therefore proper conservation strategies cannot be designed for these breeds. Advancement in affordable high throughput genotyping techniques has made it possible to genotype a large number of molecular markers, i.e. single nucleotide polymorphisms (SNPs), at a reasonable cost. Therefore, it is now feasible to infer population history and diversity status of breeds for which only scant recorded information is available. In fact, many recent studies have used a relatively high number of SNPs to explore genetic diversity, demographic history, and relatedness between different traditional Eurasian cattle breeds (Browett et al., 2018; Mastrangelo et al., 2018; Sermyagin et al., 2018; Upadhyay et al., 2016; Yurchenko et al., 2018). For instance, by using 30,000 SNPs, showed patterns of gene-flow among different Italian native cattle breeds and also reported recent inbreeding in several of them.

Swedish native cattle breeds display a large phenotypic diversity in various phenotypic and production related traits (Swedish Board of Agriculture, 2011b, 2011a). For instance, while Swedish Mountain cattle display white coat color with black or brown spots, most Swedish breeds such as Swedish Red, Swedish Red Polled (Rödkulla), and Ringamåla cattle (Ringamålako), display a solid/spotted red coat color. A large fraction of individuals display polledness in Swedish Mountain cattle, Bohus Polled (Bohuskulla) and Swedish Red Polled cattle breeds (more detailed information in Table S1). Additionally, many of these breeds have inhabited the local landscape for many generations and may harbor unique gene variants providing adaptation to the local climate. Thus, these traditional cattle breeds are important resources for future breeding programs aimed at novel/alternate breeding goals.

While many studies have investigated the pattern of genetic diversity and structure in traditional European cattle breeds (Browett et al., 2018; Mastrangelo et al., 2018; Upadhyay et al., 2016), limited knowledge is available concerning the genetic structure and diversity of Swedish native cattle breeds. In fact, to the best of our knowledge, only a few studies (Kantanen et al., 2000, 2009; Korkman, 1988; Li and Kantanen, 2010; Tapio et al., 2007) have investigated the genetic relatedness and diversity of Swedish cattle breeds. However, these studies included only a small number of Swedish cattle breeds and used either mitochondrial or microsatellite markers. Herein, we genotyped about $\sim 140,000$ SNPs in 96 samples, whereof 94 of sufficient quality, that represent all the nine native cattle breeds from different parts of Sweden. We carried out standard population genetic analyses that use either independent SNPs or haplotypes with the aim to explore genetic diversity, demography and relatedness among all native Swedish cattle breeds defined by the Swedish Board of Agriculture.

4.2 Methods

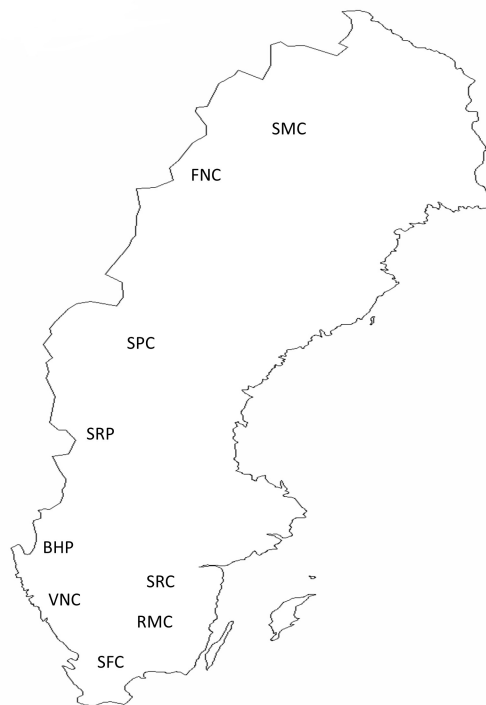


Figure 4.1: A map of Sweden indicating approximate geographic origin of the analysed Swedish native cattle breeds (some of the breeds origin from a wider geographic area). Breed abbreviations: BHP-Bohus Polled, FNC- Fjällnära, RMC-Ringamålako, SFC-Swedish Holstein Friesian, SMC- Swedish Mountain, SPC- Swedish Polled, SRC-Swedish Red, SRP-Swedish Red Polled, VNC- Väneko (map outline from D-maps.com-https://d-maps.com/carte.php?num_car=23103lang=en)

4.2.1 Sample collection and DNA extraction

Table 4.1: Summary statistics related to genetic diversity in studied Swedish cattle breeds. NC refers to the summary statistics that are not calculated for breeds with sample size less than five. MAF_{av} refers to the average minor allele frequencies, H_e refers to the average expected heterozygosity, and P_m refers to the proportion of polymorphic markers

Breed	Breed abbreviation	Sample size	MAF_{av}	H_e	P_m
Swedish Mountain	SMC	23	0.281	0.341	0.941
Bohus Polled	BHP	6	0.280	0.302	0.817
Fjällnära	FNC	16	0.289	0.342	0.924
Swedish Polled	SPC	3	NC	NC	NC
Swedish Red Polled	SRP	15	0.297	0.349	0.936
Ringamålake	RMC	13	0.291	0.320	0.891
Väneko	VNC	9	0.292	0.290	0.799
Swedish Holstein Friesian	SFC	5	0.325	0.337	0.880
Swedish Red	SRC	4	NC	NC	NC

Samples were included from 96 individuals representing nine Swedish cattle breeds (Figure 4.1 and Table 1). Old frozen blood samples and DNA samples from 35 and 59 individuals respectively, were used for genotyping. These biological materials were used from the frozen collection kept in the Department of Animal Breeding and Genetics of the Swedish University of Agricultural Sciences. Additional, nasal swabs from 2 individuals were collected in 2017. Information about original farm and pedigree was used when available, to select animals that represent different locations and to avoid including very close relatives such as full-sibs or parent and offspring. For DNA extraction from blood samples, we used either a standard phenol-chloroform method (18) or the BioRobot[®] EZ21TM Robotic Liquid Handler (Qiagen, Hilden, Germany) following the manufacturer's instructions. DNA quantification and qualification were carried out using Quant-iTTM PicoGreenTM dsDNA Assay Kit (ThermoFisher Scientific, Waltham, MA, USA).

4.2.2 SNP genotyping and filtering

DNA samples from the 96 individuals were genotyped using GeneSeek[®] Genomic Profiler High-Density Bovine 150K (GGP HD150K) array which contains about 140,000 SNPs with an average spacing of about 19 kb. Genotypes were called using GenomeStudio[®] software (Illumina, San

Diego, CA, USA) and two samples with genotyping rate less than 95% were discarded from the dataset. A bash script was used to convert Illumina genotypes report file to plink file —“ped” and “map”—format. To keep only the high-quality SNPs in the dataset, we excluded SNPs that were assigned to sex chromosomes and unassembled contigs, and SNPs with missing genotypes in more than 5% individuals. No filtering for low minor allele frequency and deviations from Hardy Weinberg equilibrium was performed to keep rare variants and SNPs potentially under selection in a specific the breed(s) in the dataset. All the quality filtering steps were carried out using PLINK 1.9 (Purcell et al., 2007).

4.2.3 Genetic diversity and a recent change in demography

To assess genetic diversity in each breed, we estimated average minor allele frequency (MAF_{av}), the average expected heterozygosity (H_e), the proportion of polymorphic loci (P_m), and the mean inbreeding coefficient (F). We used “-family” command in PLINK to generate all these estimates for each locus within each breed separately and mean values of these estimates were calculated separately for each breed.

Runs of homozygosity (ROH) are long stretches of identical by descent (IBD) homozygous genotypes that can provide valuable insight into recent and past demography of a population (Bosse et al., 2012). Therefore, we estimated ROH in PLINK using default settings except that we allowed only two heterozygous genotypes and one missing genotype per window of 50 SNPs. Additionally, the pattern of linkage disequilibrium (LD) decay between pairwise-SNPs was generated for the breeds with sample size more than 10 to supplement the demographic pattern inferred using ROH. For this purpose, we used SNeP (Barbato et al., 2015) to calculate pairwise r^2 values between pairs of SNPs located within the distance of 2 Mbp.

4.2.4 Assessment of genetic structure and relatedness among Swedish native cattle breeds

To assess population structure among the nine Swedish native cattle breeds, we first used three approaches that consider individual SNPs: 1) Principal component analysis (PCA), 2) Admixture analysis, 3) F_{st} and maximum likelihood-based phylogenetic tree. To perform PCA, the “bed” format of PLINK was converted into “gds” format using “gdsfmat” package before applying the “snpgdsPCA” function of SNPrelate package (Zheng et al., 2012, 2017). ADMIXTURE (Alexander et al., 2009) was carried out for population cluster analysis, with K-values ranging from 2 to 9. Prior to the ADMIXTURE analysis, LD pruning was performed using the “-indep” function in PLINK to reduce the overall pairwise LD to <0.15 . The output from ADMIXTURE was visualized using the python package PONG (Behr et al., 2016). Pairwise F_{st} distance was calculated using “hierfstat” R package (Goudet, 2005). Neighbor-joining (NJ) tree was constructed from the F_{st} distance matrix using “phangorn” R package (Schliep, 2011). We also calculated phylogenetic tree using maximum likelihood (ML) algorithm as implemented in “Treemix” software (Pickrell and Pritchard, 2012).

Finally, the genetic structure/clustering pattern was assessed using the algorithm implemented in CHROMOPAINTER and fineSTRUCTURE (Lawson et al., 2012). Because these algorithms take phased data as an input, Beagle 4.1 (Browning and Browning, 2007) was used to phase genotypes of each chromosome separately. The underlying algorithm in CHROMOPAINTER considers the pattern of LD and underlying recombination process along the markers to reconstruct each haplotype as a recipient of a series of genetic chunks from all the other “donor” haplotypes. As recommended by (Lawson et al., 2012), we first used “fs” pipeline to calculate nuisance parameters: n (similar to effective population size) and M (population mutation rate). The inferred values of these parameters were fixed in CHROMOPAINTER algorithm to obtain the ChromoPainter coancestry matrix (count matrix as well as length matrix) that measures the haplotype sharing among samples across the genome. Using the count matrix (number of chunks), which represents the number of haplotypic chunks copied among individuals, we ran a fineSTRUCTURE analysis to cluster the samples into genetically homogeneous groups. Following (Leslie et al., 2015), fineSTRUCTURE was run for 1 million Markov-Chain-Monte-Carlo (MCMC) iterations after discarding the first one million iterations as “burn in”; sampling was performed at every 10,000 iterations. Finally, 10,000 “hill climbing” iterations on the MCMC iteration with the highest posterior probability to get the cluster assignment.

4.2.5 Genetic relationship of Swedish cattle breeds with other European cattle

To investigate the relationships between Swedish native cattle breeds and other European cattle breeds, bovine SNP genotyping array data of various European cattle breeds as published in previous studies (Bahbahani et al., 2017; Upadhyay et al., 2016) were obtained and merged for the top alleles using PLINK (Purcell et al., 2007). Two datasets were created merging the data from these previous studies (Table S2). The first datasets consisted of 118,224 SNPs in 118 individuals which represented various North-western European cattle. The second dataset included additional samples from an indicine cattle breed—Gir—which was used as an outgroup. The genetic relationship analyses were performed using PCA and ML-based phylogenetic tree as described earlier in this section.

4.3 Results

4.3.1 Genetic diversity and demographic inference

After applying filtering criteria, our final dataset consisted of 114,000 SNPs and 94 individuals. The summary statistics related to genetic diversity are reported in Table 1. It is noteworthy that a majority of the Swedish native cattle breeds displayed a high proportion of polymorphic loci (P_m), ranging from 0.79 in Väneko to 0.94 in the Swedish Mountain cattle. This indicates that SNPs selected for GGD HD150K are highly informative for Swedish cattle breeds. The MAF_{av} is highly consistent among the Swedish cattle breeds with values ranging from 0.28 in Bohus

Polled to 0.33 in Swedish Holstein Friesian. Similarly, the expected heterozygosity (H_e) is also observed to be relatively high with values ranging from 0.29 in Väneko to 0.34 in Fjällnära.

Demographic inferences in Swedish cattle breeds were carried out using analysis based on ROH and LD decay. The Swedish cattle breeds display significant variability in ROH profile (Figure 4.2A). The Swedish Mountain cattle and Ringamålako display less within breed variability compared to Fjällnära and Bohus Polled. The Swedish Mountain cattle, Swedish Red and Swedish Holstein Friesian display relatively lower ROH counts and cumulative ROH size, indicating large/diverse ancestral populations. However, unlike the Swedish Mountain cattle, some Swedish Red Polled individuals display cumulative ROH size comparable to other individuals but with lower ROH counts, indicating several instances of mating between closely related individuals. Conversely, Väneko and Ringamålako display relatively large ROH counts and cumulative ROH size, indicating genetic isolation and a relatively small founder population. As expected, the inference drawn based on the pattern of pairwise LD decay (Figure 4.2B) is consistent with that of the ROH profiles. For instance, small haplotype diversity can be inferred for the Ringamålako breed as it displayed the largest r^2 values at all the pairwise SNP distances and overall slow LD decay. On the other hand, the LD pattern in Swedish Mountain cattle displays lowest r^2 value at all pairwise SNP distances and rapid LD decay which is indicative of large haplotype diversity.

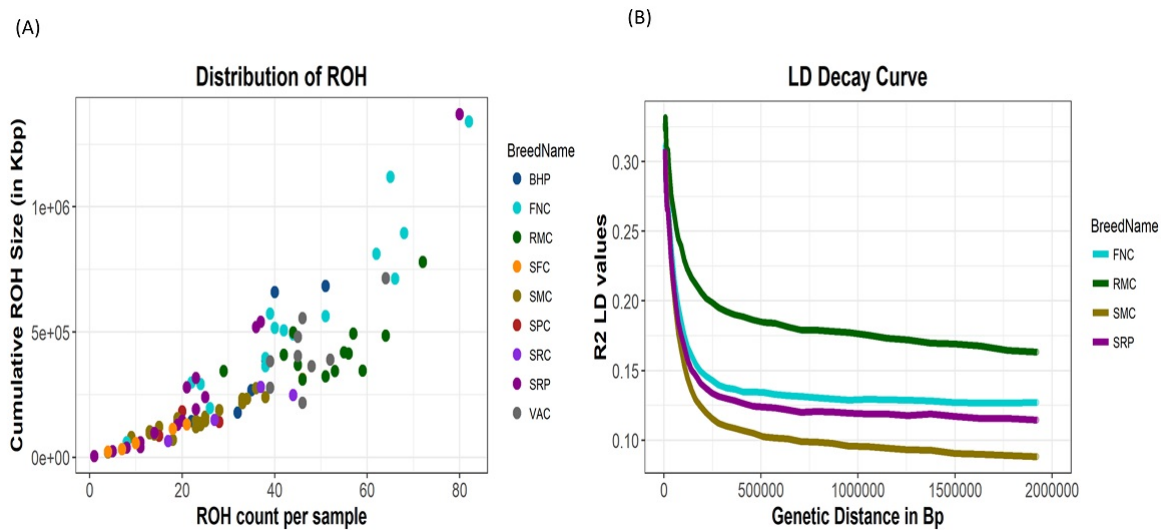


Figure 4.2: (A). ROH profile of Swedish cattle breeds, where each circle represents an individual. (B). Linkage disequilibrium decay in four Swedish cattle breeds with sample size greater than 10. Breed abbreviations: BHP-Bohus Polled, FNC- Fjällnära, RMC-Ringamålako, SFC- Swedish Holstein Friesian, SMC- Swedish Mountain, SPC- Swedish Polled, SRC- Swedish Red, SRP- Swedish Red Polled, VAC- Väneko

4.3.2 Genetic structure and relationships among Swedish native cattle breeds

To assess genetic relationships among Swedish native cattle breeds, we first performed a principal component analysis (PCA). The first principal component (EV1), which explains 6.39% of the total variance, separated individuals of Swedish Mountain cattle, Fjällnära and Bohus Polled from that of Swedish Red, Väneko and Ringamålako. The second principal component (EV2), which explains 3.65% of the total variance, separated Väneko individuals from the cluster of Swedish Red and Ringamålako (Figure 4.3A). We also observed a low genetic differentiation within the cluster of Swedish mountain breeds (including Bohus Polled). Interestingly, individuals of Swedish Holstein Friesian and Swedish Red Polled occupy the central position on the plot. Moreover, individuals of Swedish Polled display a large variation; two of its individuals form a cluster with Swedish mountain breeds, while the remaining occupy a position on the plot which appears to be a little further than the cluster of Swedish mountain cattle breeds.

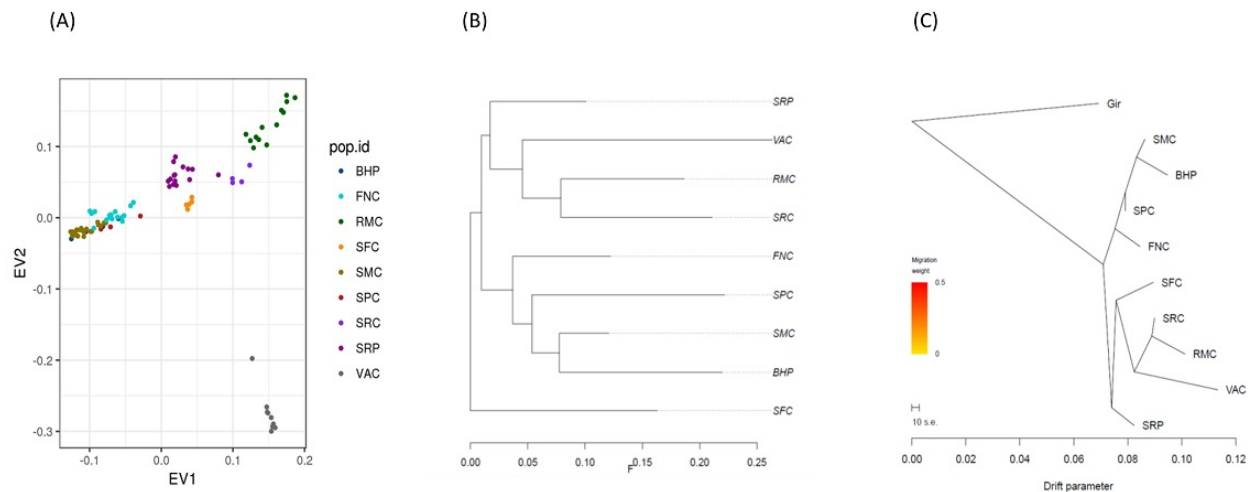


Figure 4.3: Genetic relatedness among Swedish cattle breeds using, (A). Principal component analysis, (B). F_{st} -based phylogenetic tree, and (C). Maximum likelihood based phylogenetic tree. Breed abbreviations: BHP-Bohus Polled, FNC- Fjällnära, RMC-Ringamålako, SFC- Swedish Holstein Friesian, SMC- Swedish Mountain, SPC- Swedish Polled, SRC-Swedish Red, SRP-Swedish Red Polled, VAC- Väneko.

The NJ-tree constructed based on F_{st} (Figure 4.3B) and the phylogenetic tree constructed using maximum likelihood (ML) approach (Figure 4.3C) reinforce the patterns identified using PCA (Figure 4.3A). Broadly, these phylogenetic trees divide Swedish native cattle breeds into two clusters: one cluster of the Swedish Mountain cattle, Fjällnära and Bohus Polled, and another cluster of southern horned Swedish cattle breeds (Swedish Red, Väneko and Ringamålako). These trees also suggest close relationships between Bohus Polled and the Swedish Mountain

cattle, and between Ringamålako and Swedish Red. Moreover, a long branch for Väneko also indicates considerable genetic drift/divergence. Swedish Red Polled are not close to any of the other Swedish breeds.

The clustering pattern based on shared ancestry inferred using ADMIXTURE analysis also indicates a first split (at $K=2$) between Swedish mountain breeds (including Bohus Polled) and southern Swedish breeds (Figure 4.4). Interestingly, starting from K value of 4, we observe sub-structures within the Fjällnära breed.

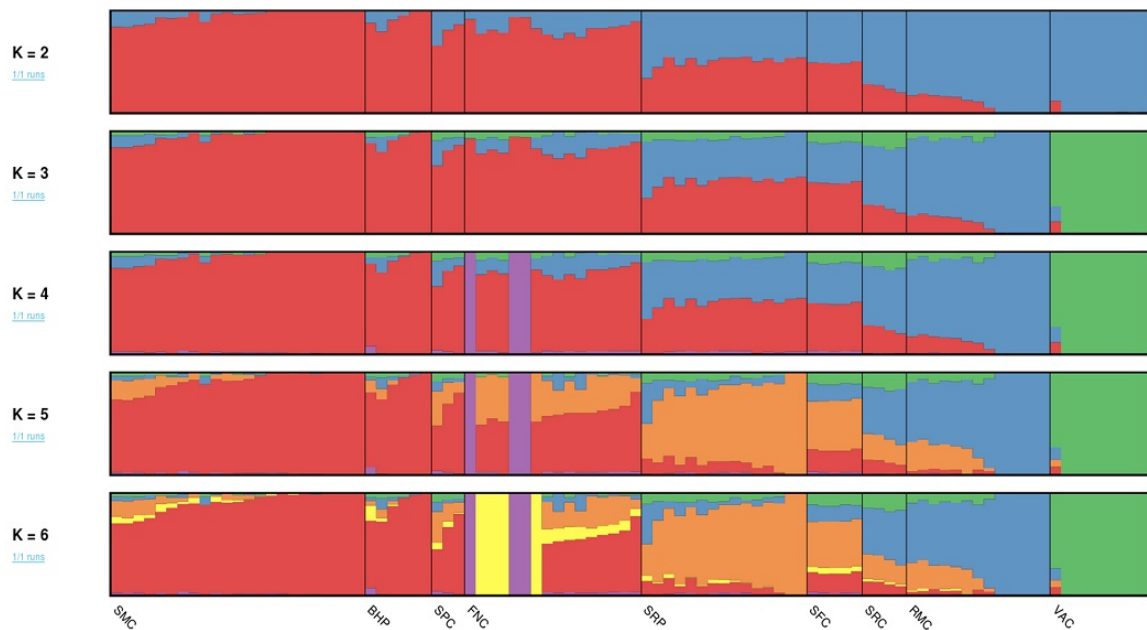


Figure 4.4: Model-based clustering of Swedish cattle breeds based on the estimated membership fraction of individuals. Breed abbreviations: BHP-Bohus Polled, FNC- Fjällnära, RMC-Ringamålako, SFC- Swedish Holstein Friesian, SMC- Swedish Mountain, SPC- Swedish Polled, SRC-Swedish Red, SRP-Swedish Red Polled, VAC- Väneko.

The pattern of haplotype sharing (Figure 4.5) as investigated using ChromoPainter and fineStructure, refines the clustering pattern of Swedish cattle breeds. In particular, these results facilitated characterizing the low differentiation among Swedish mountain cattle breeds shown in Figure 4.2A in more detail. Within the cluster of Swedish mountain breeds, except Bohus Polled, individuals of the same breed do not cluster together. For instance, we observe two clusters of the Swedish Mountain cattle; one of which forms the cluster with individuals of Bohus Polled, while the other forms the cluster with one of the sub-clusters of Fjällnära individuals. Further investigation of individuals assigned to either of the Fjällnära sub-clusters reveals genetic differentiation due to the farms from which the sampled individuals had ancestry. Though, overall, Swedish Polled form a cluster with the Swedish Mountain cattle in PCA and phylogenetic trees (Figure 4.3 and Figure 4.4). The fineStructure based analysis (Figure 4.5)

clusters two of the three Swedish Polled with the individuals of Swedish Holstein Friesian and Swedish Red Polled. Nevertheless, the clustering of southern Swedish cattle breeds is in perfect agreement with the pattern obtained using PCA and phylogenetic trees.

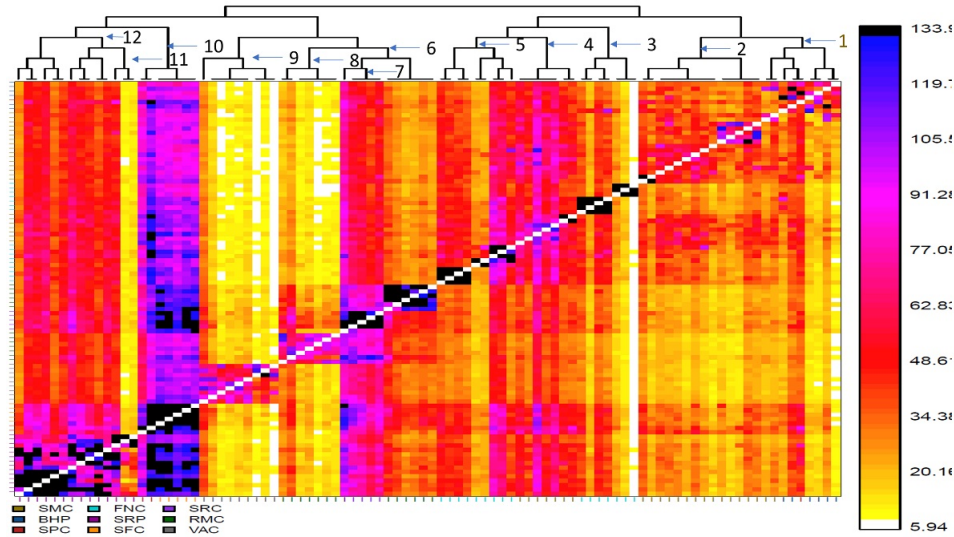


Figure 4.5: fineSTRUCTURE based phylogenetic tree. The intensity of colour indicates the number of shared haplotypic chunks. The number besides the clusters indicates: 1. Bohus Poll, 2. Swedish mountain cattle (SMC) sub-group, 3. Fjällnära (FNC) sub-group, 4. Swedish mountain cattle (SMC) sub-group, 5. Fjällnära (FNC) sub-group, 6. Ringamålako (RMC) sub-group, 7. Swedish Red cattle (SRC), 8. Ringamålako (RMC) sub-group, 9. Väneko (VNC), 10. Swedish Friesian cattle (SFC), 11. Swedish Polled (SPC), 12. Swedish Red Polled (SRP).

The relationships between Swedish cattle breeds and other European cattle breeds were investigated using PCA and phylogenetic tree. In PCA, the first two principal components, which explain altogether only $\sim 8\%$ of the total variance, separate breeds based on their geographical origin (Figure S1). Among Swedish cattle, individuals of Swedish Holstein Friesian form a cluster with Dutch cattle breeds, as expected. This relationship is also observed in phylogenetic trees (Figure S2). Interestingly, we did not identify any historical relatedness between the Swedish mountain cattle breeds and the European cattle breeds studied here, probably indicating that their contribution has been insignificant.

4.4 Discussion

Rare native cattle breeds are vulnerable and important genetic resources as they represent a unique gene pool formed as a result of long-term adaptation to the local environment. Because the majority of these breeds have not been intensively selected for production traits, they may harbor diverse alleles compared to their commercial counterparts. For instance, analyzing a

large number of genome-wide SNPs, (Herrero-Medrano et al., 2014) identified about 100 non-synonymous polymorphisms nearly fixed in commercial breeds whereas local cattle breeds displayed relatively high frequency of alternative alleles at those positions. Therefore, conservation of local cattle breeds facilitates preservation of diverse gene pools. However, the development of optimal conservation strategies requires sufficient knowledge of genetic diversity and population structure of a breed. To achieve this objective for native Swedish cattle breeds, we performed genotyping of about 140,000 SNPs in nine Swedish cattle breeds and analyzed these data using standard population genetic techniques. Our results suggest varying degrees of genetic diversity and historical relatedness among Swedish cattle breeds.

4.4.1 Genetic diversity and demography

The relatively high average minor allele frequency (MAF_{av}) and proportion of polymorphic markers (P_m) across Swedish cattle breeds (Table 4.1) indicate that the SNPs selected on the GGP HD150K array are polymorphic in these breeds. The 150K array SNPs have also been found to be highly polymorphic in Russian cattle breeds (Yurchenko et al., 2018). However, the possibility of ascertainment bias cannot be ruled out as the largest value of MAF_{av} is observed for Swedish Holstein Friesian which has a history of recent cross-breeding with Dutch dairy cattle breeds. The overall summary statistics (MAF_{av} , P_m , and H_e) indicates low genetic diversity in Väneko, Bohus Polled and Ringamålako. Inferences drawn based on ROH pattern and analysis based on LD-decay (Figure 4.2) provide further insight into demographic changes leading to low genetic diversity in these breeds. Ringamålako displays an overall high ROH count and cumulative length as well as largest r^2 values across the entire pairwise distances up to 2 Mb. These patterns indicate high autozygosity in the genome of Ringamålako individuals, which can be attributed to small founder populations leading to ancestral relatedness and/or lack of gene flow from distantly related populations due to isolation. These inferences are consistent with the known recorded history. Ringamålako has been kept isolated from other Swedish dairy breeds for a long time and the size of the populations is very small (Table S1) (Swedish Board of Agriculture, 2011b, 2011a). Interestingly, the relationship between ROH count and cumulative length (Figure 4.2B) in Ringamålako and Väneko falls within the trend observed in individuals of other breeds, indicating a low frequency of long ROH. The long ROH are often the consequence of consanguineous mating in which recombination has not had enough time to break these haplotypes. Therefore, low frequency of long ROH is an indication of a breeding strategy which has managed to avoid breeding between closely related individuals.

The overall summary statistics (MAF_{av} , P_m , and H_e in Table 4.1) indicate relatively high genetic diversity in the Swedish Mountain cattle and Fjällnära. Furthermore, the Swedish Mountain cattle display a low ROH count and cumulative length as well as rapid LD-decay (Figure 4.2), which is an indication of high haplotype diversity in this population. Before the beginning of the 20th century, the Swedish Mountain cattle had a large population size and displayed a large phenotypic diversity (Swedish Board of Agriculture, 2011b, 2011a). However, the effective population size of the breed continued to decline until the end of the 20th century because with the arrival of intensive farming other breeds were preferred. Our results suggest that despite that

reduction in population size, the Swedish Mountain cattle harbors significant genetic diversity, probably as a result of using distantly related purebred individuals in breeding programs. On the other hand, the large variation in the ROH profile in Fjällnära cattle may be related to the sub-structures in the population that we identified using ADMIXTURE and fineStructure analyses (Figures 4 and 5). Moreover, unlike the Swedish Mountain cattle, several Fjällnära individuals display an abundance of ROH counts and a large portion of the genome under ROH indicating several instances of mating between closely related individuals. Swedish Red Polled seems to have maintained high haplotype diversity as it shows relatively rapid LD-decay (Figure 4.2B). In fact, like Swedish Mountain cattle, a large historical effective population size is also recorded for Swedish Red Polled.

4.4.2 Genetic structure of Swedish native cattle breeds

The clustering pattern found for Swedish cattle breeds (Figures 3-5) is in concordance with the known history for the respective breeds. Based on their shared ancestry, we identified two major clusters within the studied Swedish cattle breeds. Low differentiation among Swedish Mountain breeds including Bohus Polled in the first cluster was indicated by results of PCA (Figure 4.3), ADMIXTURE (Figure 4.4) and fineStructure (Figure 4.5) analyses. Bohus Polled originates from the south-western part of Sweden (Figure 4.1) but is phenotypically similar to Swedish Mountain cattle and the use of semen from Swedish Mountain Cattle bulls in this breed has been documented. Earlier studies, analyzing microsatellite markers in northern European cattle, also assigned Bohus Polled cattle to the cluster of the Swedish Mountain and Fjällnära cattle breeds (Kantanen et al., 2009; Tapio et al., 2007). Similarly, Fjällnära cattle was recognized as a subpopulation of the Swedish Mountain cattle breed that was less selected for milk production. Interestingly, we also identified sub-structures within Fjällnära using ADMIXTURE and fineStructure based analyses (Figures 4 and 5). These sub-structures/clusters correspond to the farms from which these sampled individuals have their ancestry. Some individuals of the Swedish Mountain cattle clustered with sub-structures of Fjällnära (Figure 4.5), indicating several instances of gene-flow between the two populations.

The Swedish Polled cattle breed was formed in 1938 by merging the breeds Swedish Red Polled and Swedish Mountain cattle (Swedish Board of Agriculture, 2011a). Some crossbreeding between the two original breeds was practiced, but also pure-bred groups of the original breeds remained. Swedish Red Polled and Swedish Mountain cattle became a separate breed again in 1984 and in 1995, respectively; however, Swedish Polled cattle persists as a breed. Attempts have also been made to improve the production of Swedish Polled cattle by crossbreeding with commercial cattle breeds. This could explain the somewhat unclear clustering pattern seen for Swedish Polled cattle when comparing results from different analyses, but the results should be interpreted with caution as only three individuals of this breed were included in our study.

The second cluster (Figures 3-5) consist of Ringamålako, Swedish Red and Väneko. The ringamålako breed originates from southern Sweden (Figure 4.1), displays resemblance with Swedish Red cattle, and the two breeds have been suggested to share some ancestry (Swedish Board of Agriculture, 2011a). Also, fineStructure-inferred clustering pattern (Figure 4.5) sug-

gested a low differentiation between Swedish Red and Ringamålako. Interestingly, based on ADMIXTURE and fineStructure analyses, we also observed two sub-structures in Ringamålako which corresponded to the different farms from which these samples had their origin. Based on PCA (Figure 4.3) and ADMIXTURE analysis (Figure 4.4), it can be concluded that Väneko is the most diverged breed in this cluster, which might be explained by drift due to genetic isolation.

Swedish Red Polled occupies an intermediate/central position on the PCA plot (Figure 4.3A). It also displays an ambiguous clustering pattern (Figure 4.3B and 3C). Interestingly, fineStructure-based clustering (Figure 4.5) indicates a close relationship between Swedish Red Polled and Swedish Holstein Friesian.

4.5 Conclusions

To conclude, herein we provide the first detailed analyses of genetic relatedness and diversity of all Swedish native cattle breeds under the auspice of the Swedish Board of Agriculture. The information generated from these analyses will aid in the conservation management of these breeds. Our results demonstrate that some Swedish cattle breeds (such as Väneko and Ringamålako) seem to have accumulated a relatively large genetic drift and therefore, require special attention for conservation. Moreover, we also show that Swedish Mountain cattle breeds (including Fjällnära) are unique in that they have maintained authentic local ancestry. Future studies should aim at genotyping a larger number of individuals using high-density genome-wide SNP array or whole-genome sequencing approaches allowing the possibility to identify genetic factors providing adaptive potential.

4.6 Declarations

4.6.1 Ethics approval and consent to participate

All the old blood samples and frozen DNA samples were already available in our laboratory; in fact, some of these samples had already been used for microsatellite analysis in previous studies, and therefore, no ethical approval was needed to use those samples. However, cattle owners' consent was taken before collecting the nasal swabs from their animals. Moreover, they were also informed about planned research on their livestock.

4.6.2 Consent for publication

Not applicable

4.6.3 Availability of data and material

The genotyping data will be deposited to DRYAD public data repository upon acceptance of the manuscript.

4.6.4 Competing interests

The authors declare that they have no competing interests.

4.6.5 Funding

This study was generously support by the funds of Swedish Research Foundation (FORMAS). MU benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate “EGS-ABG”.

4.6.6 Author’s contribution

MRU, AMJ, SE and GA, conceived, designed, planned and directed the study. MRU, AMJ, SE, SM performed the experimental work. MRU analyzed the data. ES, RPC, MAG, GA provided significant inputs in interpretation of the data and in manuscript preparation. MRU, AMJ prepared figures, compiled breed history and wrote the manuscript. AMJ, SE and GA supervised the work. All authors read and approved the manuscript.

4.6.7 Competing financial interests

The authors declare that the research was conducted without any financial interests.

4.6.8 Acknowledgment

We gratefully acknowledge contributions of samples and information about breeds and animals from the different breed organizations and individual animal owners. We also acknowledge Kaj Sandberg and Birgitta Danell for coordinating and building a sample collection of some of the old samples in earlier projects. Other samples were sent to the animal genetics laboratory for paternity testing.

We want to thank Susanne Gustafsson, Åsa Ohlsson and Gabriela Bottani for help with the DNA extraction and measurement of DNA concentration.

Chapter 5

Distribution and functionality of copy number variation across European cattle populations

M.R. Upadhyay^{1,2*}, V.H. da Silva^{1,2}, H.J. Megens¹, M.H.P.W. Visker¹, P. A. Marsan³, V.A. Bâlteanu⁴, S. Dunner⁵, J.F. Garcia⁶, C. Ginja⁷, J.K. antanen^{8,9}, M.A.M. Groenen¹, and R.P.M.A. Crooijmans¹.

¹Animal Breeding and Genomics, Wageningen University & Research, Wageningen, The Netherlands. ²Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden. ³Institute of Zootechnics and Nutrigenomics and Proteomics Research Center, Università Cattolica del S. Cuore, Piacenza, Italy. ⁴Institute of Life Sciences, Faculty of Animal Science and Biotechnologies, University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, Romania. ⁵Animal Production Department, Veterinary Faculty, Universidad Complutense de Madrid, Madrid, Spain. ⁶Departamento de Apoio, Produção e Saúde Animal, Faculdade de Medicina Veterinária de Araçatuba, UNESP – Univ Estadual Paulista, Araçatuba, São Paulo, Brazil. ⁷CIBIO-InBIO—Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Campus Agrário de Vairão, Vairão, Portugal. ⁸Green Technology, Natural Resources Institute Finland, Jokioinen, Finland. ⁹Department of Environmental and Biological Sciences, University of Eastern Finland, Kuopio, Finland.

Abstract

Copy number variation (CNV), which is characterized by large-scale losses or gains of DNA fragments, contributes significantly to genetic and phenotypic variation. Assessing CNV across different European cattle populations might reveal genetic changes responsible for phenotypic differences, which have accumulated throughout the domestication history of cattle as consequences of evolutionary forces that act upon them. To explore pattern of CNVs across European cattle, we genotyped 149 individuals, that represent different European regions, using the Illumina BovineHD Genotyping array. A total of 9,944 autosomal CNVs were identified in 149 samples using a Hidden Markov Model (HMM) as employed in PennCNV. Animals originating from several breeds of British Isles, and Balkan and Italian regions, on average, displayed higher abundance of CNV counts than Dutch or Alpine animals. A total of 923 CNV regions (CNVRs) were identified by aggregating CNVs overlapping in at least two animals. The hierarchical clustering of CNVRs indicated low differentiation and sharing of high-frequency CNVRs between European cattle populations. Various CNVRs identified in the present study overlapped with olfactory receptor genes and genes related to immune system. In addition, we also detected a CNV overlapping the *KIT* gene in English longhorn cattle which has previously been associated with colour-sidedness. To conclude, we provide a comprehensive overview of CNV distribution in genome of European cattle. Our results indicate an important role of purifying selection and genomic drift in shaping CNV diversity that exists between different European cattle populations.

Key words: Copy number variations, European cattle, High density SNP array, population differentiation, purifying selection, drift, *KIT* gene

5.1 Introduction

Copy number variation (CNV), which is defined as large-scale losses and gains of DNA fragments, forms one of the major classes of genetic variation (Zhang et al., 2009). In the terms of total bases involved, CNV affects a larger fraction of the genome compared to Single Nucleotide Polymorphisms (SNP) (Redon et al., 2006). In addition, estimates from several autosomal dominant diseases have indicated a higher de novo locus-specific mutation rate for CNV than for SNP (Lupski, 2007). Copy number losses and gains of genetic sequences roughly make up 5-10% of the human genome, and many of the regions affected are associated with phenotypic variations including susceptibility to specific diseases (Redon et al., 2006; Stankiewicz and Lupski 2010; Zarrei et al. 2015). For example, the presence of lower *CCL3L1* gene copy number compared to the population average is associated with contracting HIV and developing AIDS (Gonzalez et al., 2005). Differences in copy number of genomic segments can result in changes in gene expression and phenotypic variation through gene disruption and altering gene dosage. For example, duplication of the *APQ7* gene in human has been linked with the emergence of traits related to endurance running as a consequence of increased expression (Dumas et al., 2007; Lupski, 2007b).

A number of studies have been undertaken in various domestic animals to characterize CNV and their effects on phenotypes (Chen et al., 2012; Liu et al., 2013; Bickhart et al., 2016; Zhu et al., 2016). For example, duplication of a set of fibroblast growth factor (*FGF*) genes and the *ORAOV1* gene in Rhodesian and Thai Ridgeback dogs causes a characteristic dorsal hair ridge (Salmon Hillbertz et al., 2007). A partial or complete duplication of the *KIT* gene causes different patterns of white coat coloration in pigs and in some of the cattle breeds such as White park, Galloway and Belgian blue (Pielberg et al. 2002; Durkin et al., 2012; Brenig et al., 2013). Similarly, white coat colour in sheep has been associated with a duplication of the *ASIP* gene (Norris and Whan, 2008). Because CNV is known to be common in regions of the genome that regulate important physiological functions, they have also been studied for association with economically important traits in domesticated animals, such as milk production and fertility. For example, several CNVs associated with milk production traits in Holstein Friesian (HF) cattle have been identified (Xu et al., 2014).

The aurochs (*B. primigenius primigenius*) is the ancestor of European cattle. Although the wild ancestor no longer exists, many extant European cattle breeds still possess primitive, aurochs-like features. These breeds are often referred to as primitive cattle breeds (Upadhyay et al., 2016). By contrast, commercial cattle breeds, including Holstein-Friesian (HF), Brown Swiss (BS) and Jersey, display derived phenotypic traits such as polledness and early maturity. It is likely that some of these differences in traits between primitive and modern cattle may result from CNV. Systematically assessing CNV between commercial and primitive cattle breeds might, therefore, reveal the genetic changes responsible for phenotypic differences, which have accumulated throughout the domestication history of cattle as result of natural and artificial selection. Moreover, contrasting cattle populations from different European regions may provide insight into the role of CNVs in population differentiation. Also, as previous studies (Park et al., 2015; Upadhyay et al., 2016) have indicated a higher frequency of aurochs specific alleles in

North-western European cattle breeds compared to Balkan and Italian cattle breeds probably as a result of secondary contact between local aurochs and ancestor of domestic cattle, it is possible that comparing CNV patterns between cattle of different regions may reveal unique CNV that might have been introgressed into the ancestors of present domestic European cattle during this secondary contact. Hence, the objectives of the current study are to provide comprehensive overview of CNV distribution in various European cattle populations and to carry out comparative evaluation of CNV patterns between cattle populations originating from different regions of Europe and thus, to address the role of CNVs in population differentiation.

5.2 Materials and Method

5.2.1 Sample collection and genotyping

A total of 196 animals from 38 different cattle breeds, consisting mainly of primitive cattle breeds (27 breeds), were sampled. Numbers of animals per breed varied from 1 to 6 (Table S1), except for Holstein-Friesian (HF), for which 55 animals were used to identify CNVs. The geographic origin of 189 animals was assigned to one of five regions or breed groups: British and Irish (BRI), Dutch (NLD, ancestral to large parts of the Lowland Pied and Baltic Red cattle), Balkan and Italian (BAI, representing Podolian and Buscha cattle), Iberian (IBR), and Alpine (ALP, combining the Central Brown and Spotted breed cluster). Because five Heck (HE) cattle did not belong to any of these groups and geographic origin of two samples were not assigned confidently they were not categorized in any of these five groups.

DNA was extracted from hair roots, blood or semen. Hair and blood samples were collected by a veterinarian in accordance with EU legislation. Semen samples were obtained from commercial AI services. The research did not involve experimentation on animals requiring approval of Animal Experiments Committee (DEC), Netherlands. The samples were genotyped with the Illumina BovineHD Genotyping BeadChip, which contains 777,692 SNPs uniformly spanning the bovine genome. All the SNP were clustered and analysed using the Illumina BEADSTUDIO software (2.0).

5.2.2 Identification of CNVs

We used the PennCNV software to identify CNVs (Wang et al., 2007). A Hidden Markov Model (HMM) algorithm as employed in PennCNV incorporates multiple parameters, such as total signal intensity (LRR), allelic intensity ratio (BAF) values of each marker for each individual, and the population frequency of B allele (PFB) of SNPs. Both the LRR and BAF values of each marker for each sample were generated from the Illumina Genome Studio software package using the default clustering file (Illumina Inc USA). The PFB was calculated as the average BAF for each marker in this population. We only used autosomal markers for detection of CNVs. The chromosomal positions of the SNPs were derived from the bovine UMD3.1

genome sequence assembly. To reduce the number of false positives, the LRR of each SNP was adjusted for the GC content of 1 Mb window surrounding the SNP using the ‘-gcmodel’ option as employed in PennCNV. The PennCNV algorithm (with options: -test) was applied to all 29 autosomes (with option: -lastchr 29) to detect cattle CNV. All samples with standard deviation of LRR(>0.30, standard deviation of BAF>0.001 and wave factor>0.05 were discarded for downstream analysis. Finally, 47 low quality samples were discarded from further analysis. Of the remaining samples (Table S1), a CNV was included in the downstream analysis if it spanned minimum of 3 SNPs (default). Furthermore, a CNV region (CNVR) was defined as a union of overlapping CNVs detected in two different samples (Redon et al., 2006). The identification of CNVRs was performed using a custom python script.

5.2.3 Comparison of cumulative CNV counts and CNV size

To compare differences in CNV counts and CNV size between the five major breed groups, we first removed the samples showing outlier values (mean ± 3 standard deviations) for either CNV counts, or CNV size, or both, if more than 5 samples represented a breed. In the second step, we used the Kruskal-Wallis test to assess overall differentiation of cumulative CNV counts, and One Way Anova to assess overall differentiation of cumulative CNV size among the five breed groups of cattle. If the overall P-value was significant ($P < 0.05$), we performed a post-hoc Mann-Whitney test for cumulative CNV counts and a T-test for cumulative CNV size to assess pair-wise population differences followed by Bonferroni correction for multiple testing.

5.2.4 qPCR validation

CNVs were validated by Real-time qPCR using the 7500 Fast and RT (Real-time) PCR system (Applied Biosystems). Primers were designed using the Primer3 webtool (<http://frodo.wi.mit.edu/primer3/>). Information about Primers and samples used in qPCR are given in Supplementary Table S2. All PCR primers were designed from UMD 3.1 reference genome based on the first 1000bp regions of CNVs. Before PCR, the quality and quantity of every DNA sample was measured with Qubit® 2.0 Fluorometer. PCR amplifications were performed in a total volume of 12.5 μ L containing reagents described in Table S2. The BTF3 gene was chosen as internal standard in all qPCR experiments.

5.2.5 Hierarchical clustering of CNVR data

To cluster samples according to their CNV similarities, we made a vector of “0”s and “1”s for each individual based on presence or absence of a specific CNVR in that particular individual. A hierarchical clustering was performed using the DendroUPGMA (<http://genomes.urv.cat/UPGMA/>). We used Jaccard index as a distance measure and the unweighted pair-group method with average mean (UPGMA) as the clustering method.

5.2.6 Identification of Segmental duplications

We applied Whole Genome Assembly Comparison (WGAC) to identify intrachromosomal segmental duplications as described previously (Khaja et al., 2006). In the first step, we retrieved the Repeat-masked UMD 3.1 cattle genome assembly from the UCSC database (<http://hgdownload.soe.ucsc.edu/goldenPath/bosTau8/bigZips/>). Subsequently, we used MegaBlast to perform sequence similarity searches within the assembly. Finally, we retrieved all paralogous sequences which displayed > 90

5.2.7 Assessment of population differentiation using V_{st}

To detect population differentiated CNVs between cattle populations of different regions, V_{st} was calculated. We calculated pair-wise V_{st} as defined previously (Redon et al., 2006) by using the equation: $(V_s - V_T) / V_T$, where V_T is the total variance in mean of LRR of a probe across all individuals of two populations and V_s is the average variance of a probe in samples within each breed. We used a window-based approach to identify groups of minimally 3 SNP probes, each showing significant V_{st} ($V_{st} > 0.35$) with the window shift of a single SNP probe. Finally, we only referred to a CNV as population differentiated if it contained the group of SNPs identified as having significant V_{st} in this window based approach.

5.2.8 Gene contents and functional annotation

The unique cattle gene list based on UMD 3.1 was retrieved from Ensembl biomart (Cow release 84). The PANTHER classification system was used to assess the probability of overrepresented genes in CNVRs within biological process, cellular component, and molecular function using Bonferroni correction for multiple comparisons (Mi et al., 2005).

5.3 Results

5.3.1 Overview of copy number variation (CNV) across all groups

A total of 9,944 autosomal CNVs were identified in 149 European cattle samples (Table S3). Out of 9,944 CNVs, 1,941 were identified as singletons (Table S4), while the remaining CNVs had minimally a base overlap with at least one CNV identified in another sample. The average number of CNV identified per sample was 67. For the different breed groups, BAI, BRI, IBR, NLD, ALP, the average number of CNV per sample was 80, 79, 70, 58 and 55 respectively. We found overall significant differences ($P < 0.05$, Kruskal-Wallis test) as well as pair-wise population differences ($P < 0.05$, post hoc Mann-Whitney test) in the average number of identified CNV per individual in each of the five major breed groups (Figure 5.1a).

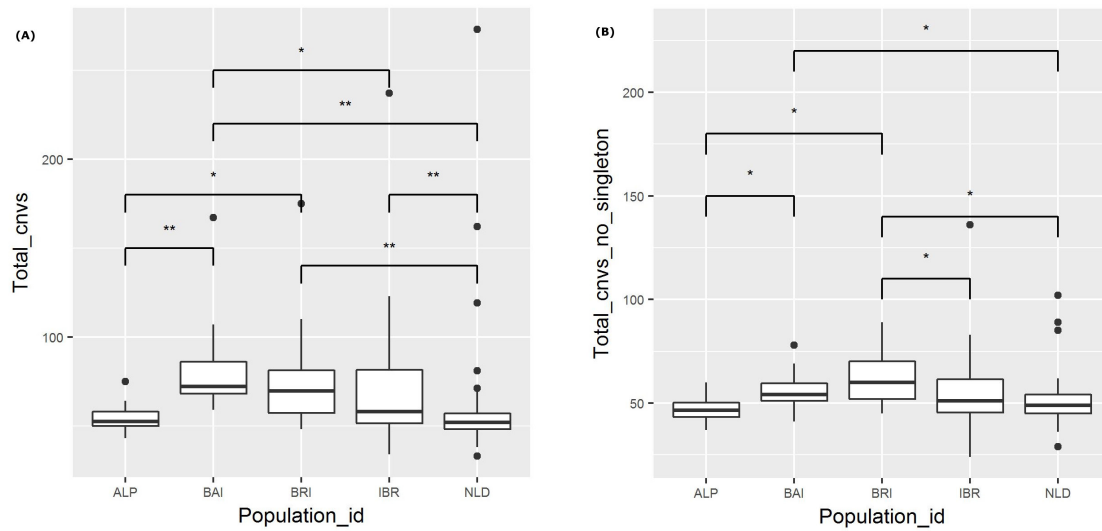


Figure 5.1: Number of detected CNVs per sample. Samples are categorized based on their origins. The median of the population is indicated by central line in a box, while the black dots represents outliers. Connecting bar above box plots between pair of populations displays significant P-values. Please note that “*” indicates $P < 0.05$ and “**” indicates $P < 0.01$. A). Total CNV counts with singleton. B). Total CNV abundance excluding singleton.

Excluding singletons from the comparison also resulted in significant differences in average number of CNVs between the major breed groups (Figure 5.1b). On the other hand, despite overall significant differences ($P < 0.05$, One way ANOVA) in average cumulative length of CNVs per individual in each of the five major breed groups, pair-wise post-hoc T-test did not result in a significant difference between any of these groups.

5.3.2 CNV validation using qPCR

To validate CNVs identified in the present study, we performed quantitative PCR (qPCR) assays for 18 CNV loci in 33 animals. These loci were chosen to represent different copy number states (loss, gain and complex) and different frequency ranges (from singleton to multiple individuals). Of the 18 CNVs tested, 13 could be confirmed by qPCR. The copy number estimated by qPCR for all confirmed CNVs, agreed with the state estimated by the PennCNV analysis. Of the 5 non-validated CNVs, 3 were identified only in one animal, while the remaining 2 were identified in at least two animals. Only 1 of the 5 non-validated CNVs did not amplify because of the poor DNA quality of samples, while nonamplification of the remaining CNVs indicated normal copy instead of hemizygous deletion as identified by PennCNV for all locus. These results also indicate high likelihood of singleton CNVs, i.e. CNVs that occurs only once in the dataset, being false positive (Table S2).

5.3.3 Overview of CNVRs

A total of 923 CNVRs were identified by aggregating overlapping CNVs with overlap identified in at least two animals (Figure 5.2 & Table S5). These CNVRs cover 61.06 Mb of the cattle UMD 3.1 genome assembly, which corresponds to 2.5% of the 29 bovine autosomes. However, the distribution of CNVRs across all autosomes varies considerably, with the highest number (55) on chromosome 1 and the lowest (10) on chromosome 27. The estimated length of CNVRs varies from 1.53 Kb to 3.51 Mb, with an average of 66.15 Kb (Figure 5.3). The ratio of total estimated CNVR length per chromosome to the length of that chromosome varies from 8.20% for chromosome 12 to 0.040% for chromosome 26. Chromosome 23 displays the highest density of CNVRs with an average distance of 1.46Mb between CNVRs, while chromosome 20 exhibits the lowest density of CNVRs with an average distance of 5.24 Mb between CNVRs. These 923 CNVRs are comprised of 587 losses, 179 gains and 157 complex (both loss and gain) events. Furthermore, the frequency of these CNVRs in the populations under study ranged from 1.34% (present in 2 of the 149 animals) to 97.31% (present in 145 of the 149 animals). The 157 complex CNVRs, on average, displays much higher frequencies (13%) than the average frequency of only losses (3.6%) or only gains (3.4%) CNVR events. One complex CNVR (id: CNVR527) displays the highest frequency (97.31%) and is located on chromosome 12 between 73.2Mbp and 76.7Mbp. Out of all 923 CNVRs, 198 CNVRs (21%) have a frequency of more than 5% in the studied cattle populations (Table S5). Of these 198 CNVRs, 49 were identified in at least one individual of each of the five breed groups (Table S5). The BRI and NLD populations revealed the highest (26) and the lowest (8) number of CNVRs per sample respectively.

5.3.4 Hierarchical clustering based on CNVRs

To assess population differentiation, hierarchical clustering was carried out after converting CNVRs into binary data (See methods). Samples belonging to low diversity breeds, as identified in our previous SNP based studies (Upadhyay et al., 2016), such as English longhorn (EL) and Boskarin (BK), cluster together (Figure S1). However, samples belonging to IBR breeds show a lower degree of differentiation compared to samples belonging to the other regions. Also, the breeds from the same geographic region do not cluster together. For example, Brown swiss (an Alpine breed) samples form a clade with the Iberian samples belonging to Pajuna (PA), indicating sharing of high frequency CNVRs.

5.3.5 Association between high frequency CNVRs and Intra-Chromosomal Segmental duplications (SD)

The large proportion (>21% of total) of CNVRs displaying a frequency above 5% indicates that either; (1) they fall into regions of CNV hot spots in the cattle genome, (2) these are likely to be of more ancient origin when compared to the low frequency CNVRs or (3) they are under strong positive selection. CNV hot spot regions in the genome usually overlap with genomic segmental duplications (SDs) (Sharp et al., 2005). Thus, to assess the correlation between SDs

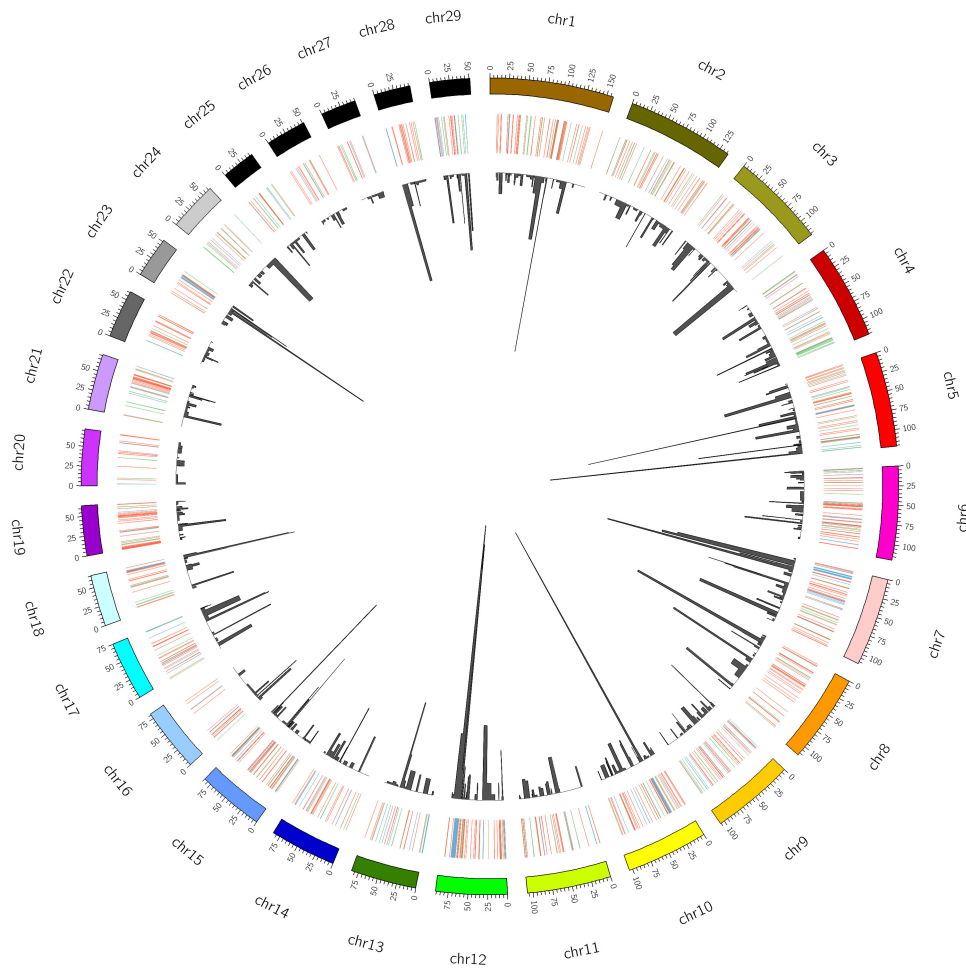


Figure 5.2: Distribution, status and frequency of CNVR in the bovine genome (UMD 3.1). The status of CNVRs are shown in outer circle in Red (loss), Green (gain) and Blue (both), while the inner circle is indicative of the frequency.

and high frequent CNVRs, we identified SDs from the UMD 3.1 bovine genome assembly. In the analysis, we only considered intra-chromosomal SDs longer than 5 Kb, as these SDs are more likely to cause misalignment and aberrant recombination (Stankiewicz and Lupski, 2002) than the small size SDs (<5Kb). Of the 21 CNVRs that overlapped with large SDs, 14 were present at a frequency of more than 5% across all cattle populations (Table S6).

5.3.6 Assessment of population differentiated CNVs based on Vst

To identify CNVs contributing to population differentiation, we calculated the pairwise Vst between every possible combination of the five major breed groups, and between, HF and the four major breed groups. In the latter Vst analysis comparing HF samples with other breed groups we did not include NLD, as large proportion of its sample size consisted of HF samples. The value of Vst varies between 0 (no population differentiation) and 1 (complete population differentiation), with high Vst values suggesting a difference between populations in the fre-

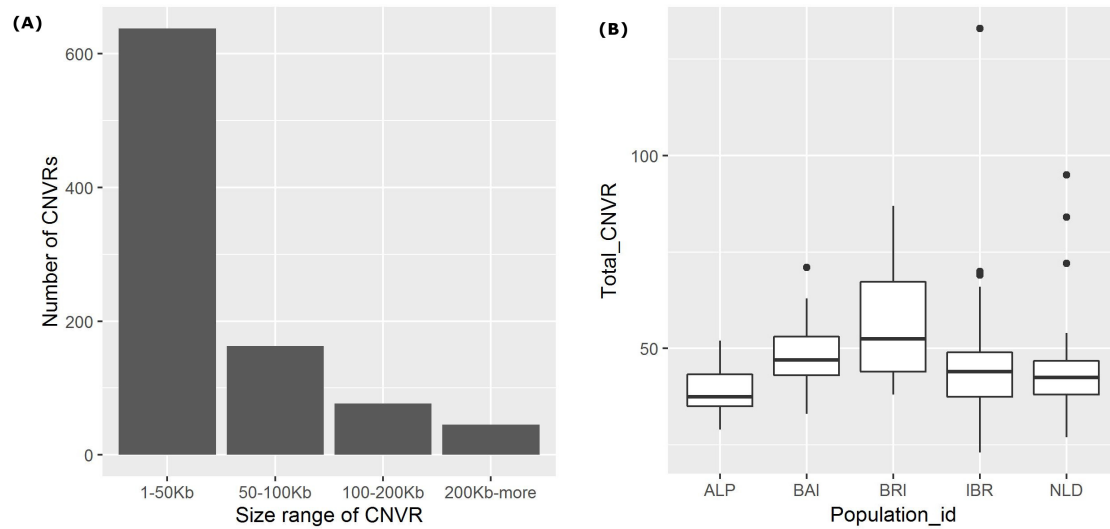


Figure 5.3: A). Distribution of CNVR size in the genome. B). Distribution of CNVR per sample categorized based on its origin. The median of the population is indicated by central line in a box, while the black dots represents outliers.

quencies of copy number states of underlying sequences. Interestingly, we observed a very few breed group differentiated CNVs for all combinations except for HF vs BRI, where we observed quite a few breed group differentiated CNVs. We also considered the effect of mis-assembly on lineage differentiated CNVs as it has been reported previously that incorrect placement of the sequence from the sex chromosome on autosomes may distort the dosage ratio between male and female which can lead to false positive lineage population differentiated CNVs (Zhou et al., 2016). Clearly, except for three CNVRs, all CNVRs identified as breed group differentiated (in HF vs BRI) displayed difference in LRR values between bull and cow samples (Figure S2, Table S7). Thus, the high V_{st} observed between BRI and HF for CNVs involving miss-assembled SNPs can be attributed to the highest proportion of female samples in the BRI group, while all HF samples were male.

5.3.7 Gene content of CNVRs

Of the 923 CNVRs identified in the present study, 415 (45%) span (with at least one bp overlap) 916 unique cattle genes (Table S8). Interestingly, genes from certain immune response related gene families such as the *TRAV* like gene family, and *IGL* were covered by complex CNVRs indicating differences in copy number between different cattle breeds. The most frequent CNVR (id:CNVr527) covered four related genes, which are located in tandem and display similarity to *ABCC4* in human. We also found genes related to economically important traits of livestock covered by CNVRs such as *MTHFSD*, and *GTF2I*. In addition, CNVRs also covered some essential genes such as *MSH4* and *ATF2* (Table S9). The gene ontology (GO) analysis for 916 ensemble unique genes using the Panther classification system (REF) showed that terms related to immunity and olfactory activity were overrepresented in the CNVRs that were identified in

the present study (Table S10).

Interestingly, 5 cattle samples displayed a CNVR (id: CNVr2375) covering the *KIT* gene. However, occasionally, CNVs contributing to a CNVR may or may not fall in the gene covered by that CNVR. Hence, we investigated the CNVs within CNVr 2375. Detailed analysis of these CNVs revealed that of the 5 samples, 3 English Longhorn (EL) samples had a CNV gain covering the *KIT* gene, while the remaining samples had CNVs outside the gene. In addition, the estimated size and position of this CNV within CNVr 2375 (Figure 5.4A) displayed similarity with the genomic segment involved in a serial translocation from chromosome 6 to chromosome 29 that was previously reported in Belgian Blue, Galloway and White Park cattle breeds (Durkin et al., 2012; Brenig et al., 2013). Thus, to test the presence of that same serial translocation overlapping the *KIT* gene, we PCR amplified the known fusion points of the translocation (for more details refer to Durkin et al., 2012). The amplification of fusion points (-D,E-A and C-) confirm the presence of the Belgian Blue type allele (Cs29) in 2 English Longhorn (EL) samples (Figure 5.4b), whereas the remaining sample did not amplify due to poor DNA quality. The White Park cattle samples, which were discarded from the analysis due to high standard deviation in LRR, also reveal the presence of the Cs29 allele. These results show a high prevalence of the Cs29 allele in BRI cattle breeds.

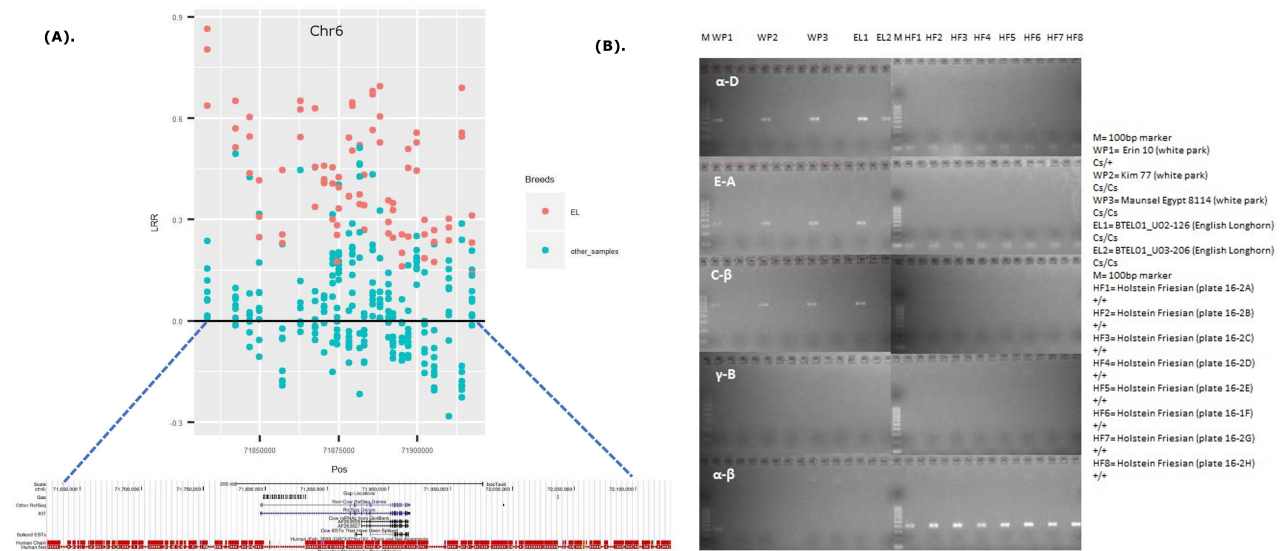


Figure 5.4: CNVr2375 completely overlaps the *KIT* gene in English longhorn samples. (A). Regional SNP plot of CNVr2375. The mean LRR for each marker in EL samples are displayed in Red, while that of the remaining samples are displayed in light green. (B). Result of PCR performed to validate the presence of Cs29 allele in English Longhorn and White Park cattle, where -D, E-A, and C- refers to the fusion points of Cs29 allele (for more details refer to Durkin et al., 2012), while - refers to the wild type (+) allele. Abbreviation: WP- White Park cattle, EL-English Longhorn cattle, M-100 bp ladder Marker, HF-Holstein Friesian.

5.4 Discussion

5.4.1 Difference in abundance of CNV counts between different cattle populations

We found significant differences in CNV counts between different cattle populations. The BAI and the BRI breed groups displayed relatively high number of CNVs per individual compared to the breed groups from other regions. The mean CNV cumulative length per individual between different cattle groups, however, was not significantly different. Such difference in CNV abundance between different cattle populations have already been reported previously. For instance, high CNV abundance has been reported in indicine and African taurine cattle breeds than in European taurine, which has been attributed to their breed divergence and population history (Liu et al., 2011). Similar observations were also reported in other species as well. For instance, Pezer et al. (2015) reported difference in total CNV abundance and total genic deletions between several natural populations of the house mouse, which they attributed to difference in effective population size between mouse populations. It is evident from these studies that population history such as change in past effective population size, gene flow, and selection process may contribute to differential CNV abundance between different populations. Thus, we hypothesize that persistence of small effective population size over many generations in BRI breeds such as English longhorn (EL) and Highland (HL) may have resulted in relaxation on purifying selection against slightly deleterious CNVs and which in turn, may have resulted in accumulation of large number of CNVs and genic deletion events. To test this hypothesis, we calculated the number of genic deletion CNVs, percentage of genic CNVs as well as cumulative genic CNV length in breeds with more than or equal to 3 samples. (Figure 5.5, Figure S3).

Indeed, we observed a higher number of genic deletion as well as higher cumulative length of genome under deletion in BRI breeds (English Longhorn (EL) and Highland (HL) (Highland)) compared to breeds from other regions. This observation supports the hypothesis that genetically isolated small populations may accumulate abundance of CNVs, in particular, deletion CNVs. However, we note that, since some SNP arrays display bias towards detection of deletions (Pinto et al., 2011) and the present study suffers from low sample size per breed, large samples from multiple breeds are needed to be sampled to explore this hypothesis further.

On the other hand, within population variability in number of genic and total number of CNVs in several BAI and IBR breeds may be attributed to high admixture pattern of their genomes (Decker et al., 2014; Upadhyay et al., 2016) or higher historical effective population size compared to NLD, ALP or BRI breed-groups.

5.4.2 Comparison of Identified CNVRs with previous studies

To characterize the CNVRs identified in the present study in more detail, we compared them to the CNVRs identified in eighteen previous studies using various methods such as whole genome sequencing(Stothard et al. 2011; Boussaha et al., 2015; Bickhart et al. 2016; Ben Sassi et

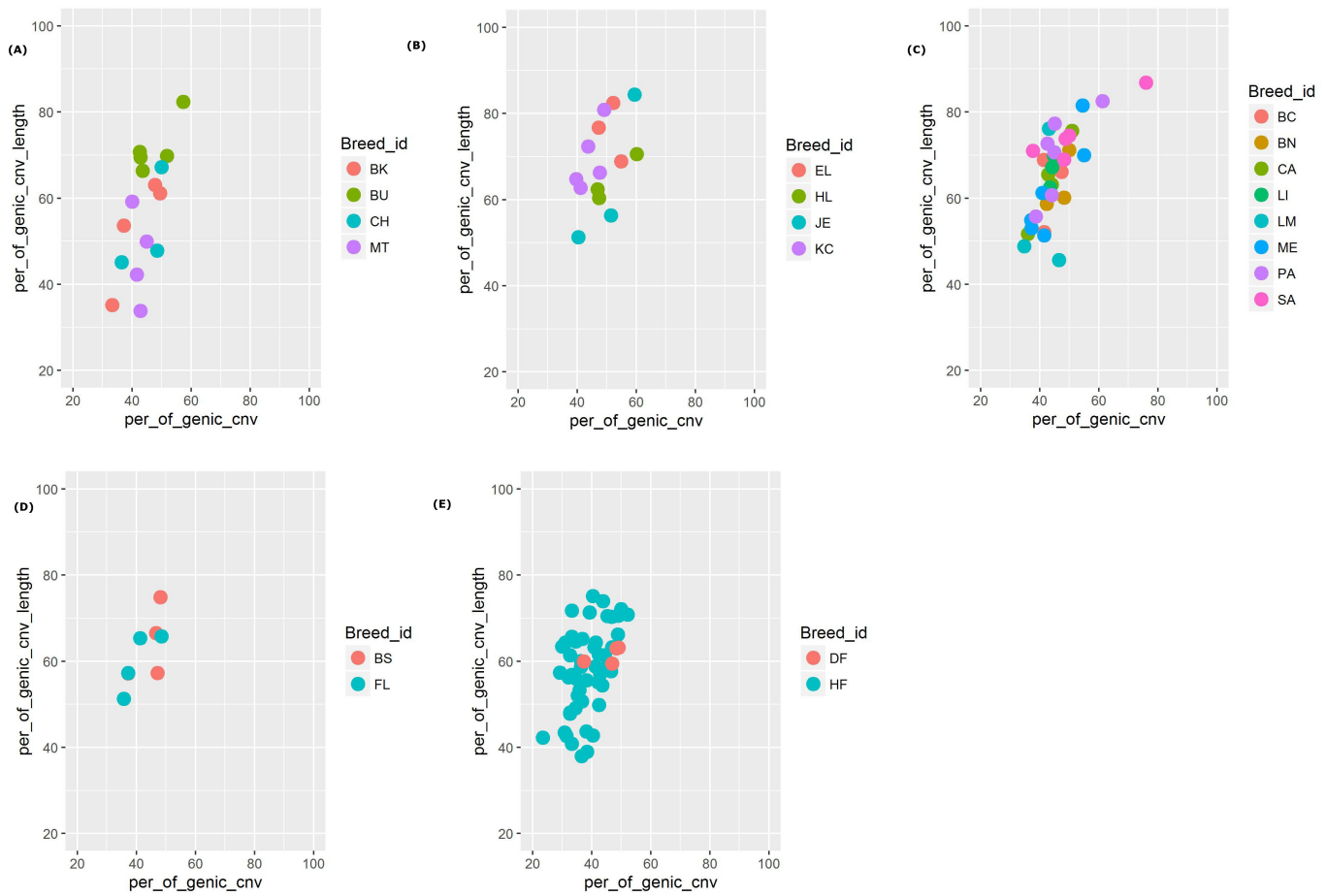


Figure 5.5: Percentage of genic CNV counts (on X axis) and cumulative genic CNV length (on Y-axis) per individual split based on breed and its geographical region. (A) Balkan and Italian cattle breeds (BAI), (B) British and Irish cattle breeds (BAI), (C) Iberian cattle breeds (IBR), (D) Alpine cattle breeds (ALP), and (E) Dutch (NLD) cattle breeds.

al. 2016), comparative genomic hybridization (Fadista et al. 2010; Kijas et al. 2011; Liu et al. 2010; Zhang et al. 2014, 2015), Illumina BovineHD BeadChip Arrays (Hou et al. 2012; Jiang et al. 2013; Zhang et al. 2014b; Sasaki et al. 2016), and Illumina Bovine50K SNP array (Bae et al., 2010; Hou et al., 2011; Jiang et al., 2012). The comparisons revealed that 737 (80% of the total) number of CNVRs detected in the present study overlapped completely or partially (by at least a single bp overlap) with CNVRs from these previous studies detected in literature (Table 1). The inconsistency of overlaps with different studies can be attributed to differences in size and structure of populations under investigation, platforms and algorithms of CNV calling, definitions of CNV and CNVR between various studies. As expected, high overlaps are reported in studies that have investigated CNVs in diverse cattle breeds using the Illumina BovineHD BeadChip array (Table 1).

The identification of 186 Novel CNVRs suggests that a substantial number of CNVs in cattle

genomes are yet to be identified (Table S11). Of the 186 CNVRs (20% of the total) identified as novel in this study, 49 are breed specific CNVRs for various breeds. Of the 49 private CNVRs, 27 are observed only in HF. We note that HF might have shown the highest percentage of breed specific CNVRs due to the larger sample size investigated in our study.

Table 5.1: Comparison between CNVRs identified in the present study with previous studies in terms of count and length. Please note that the term “overlap” refers to the number or percentage of CNVR of previous study that display overlap with the present study.

Methods	Study	Total CNVR segment	Total cumulative CNVR length	Over lapped CNVR	Over lapped CNVR (%)	Over lapped CNVR cumulative length (%)
50K	Bae_et_al_2010	368	63,107,899	39	10.60	3.36
	Hou_et_al_2011	743	15,8021,612	188	25.30	15.24
	jiang_et_al_2012	101	23781,338	19	18.81	3.63
	wang_et_al_2015	389	70748371	60	15.42	3.40
770K	Hou_et_al_2012	3438	146902512	658	19.14	25.97
	Jiang_et_al_2013	357	34407298	119	33.33	47.32
	Sasaki_et_al_2016	861	43654930	347	40.30	48.30
	Xu_et_al_2016	257	12443243	95	36.96	80.11
	Zhang_et_al_2015	365	13128818	116	31.78	53.60
	Fadista_et_al_2010	254	15760830	22	8.66	7.94
CGH	Kijas_et_al_2010	27	6085066	6	22.22	9.98
	Liu_et_al_2010	200	36171861	91	45.50	41.32
	Zang_et_al_2014	353	42915082	41	11.61	9.14
	Zhang_et_al_2015	339	36596362	35	10.32	9.23
	Bickhart_et_al_2012	1265	55590961	87	6.88	7.90
WGS	Boussaha_et_al_2015	4199	1012466378	1314	31.30	41.83
	Sassi_et_al_2016	823	45420220	89	10.81	26.47
	Stothard_et_al_2011	790	3287618	48	6.08	4.93

Abbreviations: 50K-Illumina BovineSNP50 BeadChip, 770K-Illumina BovineHD BeadChip, CGH- Comparative Genomic Hybridization, WGS- Whole Genome Sequencing.

However, despite the small sample size, all BRI breeds displayed at least one breed specific CNVR. Interestingly, the English Longhorn (EL) displayed quite a few breed specific private CNVRs (6) followed by Heck (HE), which displayed 4 breed specific CNVRs.

5.4.3 Sharing of highly frequent CNVRs and low differentiation between European cattle populations

Hierarchical clustering of CNVRs revealed that animals which belonged to breeds with a relatively low diversity such as English longhorn (EL), Boskarin (BK), Dutch Friesian (DF), Maltese (MT) and Heck (HE) formed a clear cluster. This type of a clustering pattern suggests that low diversity led to an increase in shared CNVRs between animals. For example, the Boskarin (BK) breed displayed more than 40% of the total CNVRs as shared between two or more samples (data not shown). However, unlike SNP based clustering of these same samples in our previous study (Upadhyay et al., 2016), samples from the same region did not cluster together (Figure S1). This indicates a relatively low level of differentiation or sharing of high frequent CNVRs among different European cattle breeds. This discordance in inference of European cattle population structure indicates that either our analysis suffered from low sample size per breed or most CNVs are transient enough to not have followed the same pattern of demographic events in the history of cattle domestication that typical neutral genetic variants have experienced. In addition, it can be speculated that de-novo CNVs, CNV hot-spot regions in the genome and false CNV calls due to variation in genotyping intensities can also affect inference of population stratification.

To investigate the highly frequent CNVRs in more detail, we performed their association with SDs. It has already been shown that SDs provide substrate for non-allelic homologous recombination (NAHR), which in turn, produces novel chromosomal rearrangements and copy number changes. Therefore, CNVRs that overlap with SDs typically display high frequencies as compared to the CNVRs that do not overlap SDs. Accordingly, we found an enrichment of highly frequent CNVRs in cattle SDs as previously observed in human, mice, and apes (Gazave et al., 2011; Pezer et al., 2015; Redon et al., 2006).

Recently, Sudmant and colleagues (2015), while analysing CNV patterns across different human populations and a Denisovan sample, identified large duplications that introgressed from the extinct Denisovan lineage exclusively into Oceanic population, and also were present at high frequencies. Since north-western European cattle breeds harbour high frequency of aurochs' specific alleles, probably as a result of secondary aurochs introgression (Park et al, 2015; Upadhyay et al., 2016), we investigated whether animals from these regions carry any unique CNVs. However, Vst based analysis did not identify any region specific CNV that might have introgressed from aurochs during secondary contact, i.e. CNV present only in animals of certain regions as a result of aurochs introgression. In the future, the availability of high-coverage sequence from archaic aurochs samples might aid researchers in identification of ancient CNVs in the genome of European cattle. Additionally, Vst based analysis also did not identify any breed-group differential CNVs, when contrasting HF (commercial breed) against IBR or BAI animals. This observation is consistent with a recent study on bovine population structure, where authors reported only few lineage-specific CNVs in breeds from the same continent i.e. Holstein and Angus cattle breeds (Xu et al., 2016). However, quite a few breed-group differentiated CNVs between HF and BRI were identified, except a few, all of which turned out to be false positive CNVs.

5.4.4 Copy number variable genes

Cattle genomes display enrichment of CNVs in genes related to immune response and environmental interaction such as sensory perceptions of smell and chemical stimuli. Many of these immune related genes appeared to be copy number variable between different cattle breeds. These variations may explain inter-population differences in immunological response to different clinical conditions. For example, *BoLA-DRB3* locus which partially lies within a high frequency complex CNVR (id: CNVr1586, Table S8) has been associated with differential response to various clinical conditions such as Mastitis and Bovine leukaemia virus infection in various cattle breeds (Rupp et al., 2007; Yoshida et al., 2012). Another interesting example is the *CIITA* gene, which lies within a high frequency gain CNVR identified in multiple breeds (id: CNVr1680, Table S8) and which was found to be duplicated in Angus cattle that showed nematode resistance (Liu et al., 2011). On the other hand, enrichment of CNVs in genes related to sensory perception of smell is either indicative of physiological requirements of domestic animals as described previously in case of pig (Paudel et al., 2013) or can, alternatively, be the result of drift due to random duplication and deletion of olfactory genes (OR) (Nozawa and Nei, 2007). Similar overrepresentation of CNVs in immune related genes and OR genes was reported previously in various species of domestic animals (Paudel et al., 2013; Liu et al., 2013; Da Silva et al., 2016). The most frequent CNVR displayed variable copy numbers and covered genes similar to *ABCC4* in human. The *ABCC4* genes encode a protein related to ATP-binding cassette (ABC) transporters which transport various molecules across extra- and intra-cellular membranes. Previous studies have reported association between *ABCC4* genes and phenotypic traits related to disease resistance and feed efficiency in cattle (Liu et al., 2011; Chen et al., 2012). Interestingly, Lee et al., (2013), using whole genome sequencing data of Hanwoo cattle, reported a very high number of non-synonymous SNPs, splice-site variants, and coding indels in *ABCC4* gene. These observations imply that either *ABCC4* gene has evolved into multiple copies for environmental adaptation, or, alternatively, mis-assembly at chromosome 12 led to distortion of signal intensity ratio resulting in detection of false CNVRs. The CNVR1206 that covered the *MTHFSD* gene, has been associated with milk protein yield in Spanish HF cattle (Sassi et al., 2016), while the *GTF2I* gene that lies within CNVR1703, has been identified as a candidate gene related to traits associated with feed conversion efficiency in chicken (Reyer et al., 2015). Novel CNVRs, i.e. CNVRs that were identified only in the present study, spanned important genes such as *MSH4* and *ATF2* etc. The *MSH4* gene encodes a protein essential for reciprocal recombination and proper segregation of homologous chromosomes at meiosis. Additionally, deficiency of *MSH4* gene, which is covered by CNVr1975 (Table S8) and identified only in 2 animals, has been associated with impaired gamete formations in laboratory mice (Kneitz et al., 2000). Recently, Ma et al., (2015) and Kadri et al., (2016) also have associated the *MSH4* gene with recombination rate in cattle. Also, deficiency of the *ATF2* gene, which is partially covered by CNVr1217 and identified only in 4 animals, led to early postnatal lethality in laboratory mice (Bhoumik and Ronai, 2008). Since both these CNVRs (id: CNVr1975 and CNVr1217) have been identified in low frequency and in heterozygous deletion state, the hypothesis of purifying selection against such deleterious CNVs cannot be ruled out.

Recently, Durkin and co-workers (Durkin et al., 2012) have shown that a duplication of a *KIT*

gene segment from chromosome 6 and its aberrant insertion on chromosome 29 led to the “colour-sided” white coat colour phenotype in Belgian blue cattle. Additionally, Brenig and co-workers (Brenig et al., 2013) identified the Belgian-blue type allele (Cs29) in White Park and Galloway cattle. Interestingly, they also suggested a dose dependent effect of Cs29 in which heterozygous (Cs29/wild allele) animals exhibited variable degrees of pigmented spots on white body trunk and homozygous (Cs29/Cs29) animals produced no pigmentation on the body trunk. In our study, we show the presence of the Belgian Blue type allele in English longhorn cattle, most likely introduced in the English longhorn (EL) due to cross breeding with other cattle breeds such as Galloway and White park that also carry the same allele. In addition, as both English longhorn (EL) animals were homozygous for Cs29 and as this breed harbours low diversity, we hypothesize that frequency of Cs29 allele is high in this breed.

In summary, we utilized signal intensity data from Illumina BovineHD genotyping array to identify copy number variation in cattle populations sampled from different regions of Europe. The comparative evaluation indicated a higher abundance of CNV counts in British and Balkan-Italian cattle breeds, probably because of high historical effective population size or relaxation on purifying selection of slightly deleterious CNVs. Also, clustering based on copy number variation regions displayed low population differentiation indicating the effect of transient CNVs or CNV hot-spot region. Functional analysis revealed enrichment of CNVR in genes related to immunological responses and environmental interaction such as sensory perceptions of smell and chemical stimuli. In addition, we also detected a CNV overlapping the *KIT* gene in English longhorn cattle which has been identified previously in Belgian blue, white park and Galloway cattle and associated with colour-sidedness.

5.5 Conflict of interest

The authors declare no conflict of interest.

5.6 Author contributions

MU, RC, and MG conceived the study, its design and coordination. MV, PM, VB, SD, JG, CG, and JK identified pure bred animals and carried out sampling. MU and VS identified CNV and performed filtration for subsequent analysis. MU performed all the downstream analysis. MU drafted the manuscript, MG, RC, HM, PM, VB, SD, JG, CG, VS, and JK provided critical remarks on the content and all authors read and approved the manuscript.

5.7 Funding

MU benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate “EGS-ABG”.

5.8 Acknowledgements

The authors would like to thank Bert Dibbits, Wageningen University & Research, Animal Breeding and Genomics, for DNA extraction and qPCR validation.

5.9 Supplementary information

The supplementary information can be found at: <https://www.frontiersin.org/article/10.3389/fgene.2017.00108/full> #supplementary-material

Chapter 6

Comparative evaluation of structural variations in taurine and indicine cattle using individual whole genome sequences

M.R. Upadhyay^{1,2}, M.F.L. Derks¹, G. Andersson², M.A.M. Groenen¹ and R.P.M.A. Crooijmans¹

¹ Animal Breeding and Genomics, Wageningen University & Research, Wageningen, The Netherlands. ² Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden.

The draft has been submitted.

Abstract

Background

Structural variations (SVs) are large size genomic regions which include inversion, balanced translocation, insertions or deletions. Although SVs are relatively less frequent than SNPs, some have a significant contribution to individual fitness. In this study, we used 29 whole genome sequences representing various European taurine, African taurine and zebu cattle to identify SVs in cattle genomes. Additionally, SVs were also identified from whole genome sequence (WGS) of ancient aurochs to further facilitate their comparison with modern domesticated cattle.

Results

Applying multiple approaches of SV identification and stringent filtering criteria, we identified 11,185 autosomal SVs in the domestic cattle studied. The average number of SVs identified per sample in domestic cattle was 1820 and an ancient aurochs WGS harbored 749 high-quality SVs. Functional annotation revealed that these SVs encompassed various genes related to metabolism, coat color and meat quality. Additionally, several novel lineage-specific copy number variants (CNVs) were identified. Moreover, many duplication events identified in an ancient aurochs sample overlapped with those identified in modern domestic cattle, indicating that these events might already have been segregating in the ancestral aurochs population. Furthermore, some of these shared duplications were identified in genes encoding olfactory receptors and proteins with functions related to immunity, probably indicating their evolutionary importance.

Conclusions

We identified various protein-coding genes and regulatory elements encompassing SVs which represents valuable resources for future studies aimed at finding the association between physiological processes and SVs. Additionally, our results also highlight the important contribution of SVs in relation to cattle diversity and evolution.

6.1 Introduction

Sequence diversity consists of various types of genomic variations such as single nucleotide polymorphisms (SNPs), structural variations (SVs), small insertions and deletions (INDELs). SVs encompass large size (usually > 1 kbp) genomic features such as duplications, deletions, inversions and translocations were observed (Layer et al., 2014). Although SVs occur at relatively low frequencies, they may have a high impact on gene structure and function. SVs can influence mRNA and protein expression levels, if overlapping with regulatory regions or coding sequences of genes (Aldred et al., 2005; Perry, 2009), thereby contributing to the fitness of an individual. E.g., in humans, different expression levels of the *AMY1* gene between populations has a direct relationship with the *AMY1* copy number, as each duplicated copy also carries the regulatory regions of the gene (Bank et al., 1992; Perry et al., 2007). Conversely, deletions can also affect the phenotype by disrupting the regulatory elements of the gene as reported for goat polled intersex syndrome (Schibler et al., 2000).

Being widespread in mammalian genomes, SV contributes significantly to diversity. For instance, Asian pig carry increased copy numbers of the *UGT2B10* gene compared to European pigs; this increase in copy numbers was suggested to be associated with detoxification (Paudel et al., 2013). Also, in cattle, genes encoding proteins related to parasitic disease resistance, such as *CATHL4* and *ULBP17*, display a difference in copy number between indicine and taurine cattle (Bickhart et al., 2012). Moreover, beef cattle also have been shown to carry high copy numbers of genes related to lipid metabolism and transportation (Bickhart et al., 2012). Similar to pig and sheep, SV in the genes controlling coat color, such as *KIT*, were also reported in cattle (Durkin et al., 2012; Upadhyay et al., 2017). In fact, Durkin et al., (2012) reported two alleles involving SV affecting the *KIT* gene resulting in differential patterns of coat color in Brown Swiss and Belgian Blue cattle.

Different approaches have been used to identify SVs in the bovine genome including comparative genomic hybridization (CGH), SNP microarray and whole genome resequencing (Gao et al., 2017; Kijas et al., 2011; Liu et al., 2010; Stothard et al., 2011; Upadhyay et al., 2017; Zhan et al., 2011). While CGH and SNP microarray-based methods suffer from low accuracy in estimation of proper break-points and length of SV events, advances in next-generation sequencing technology and computational algorithms allow the identification of common and rare SVs with relative accurate estimation of the break-points (Bickhart et al., 2012; Mesbah-Uddin et al., 2018).

SV identification methods based on whole genome resequencing data can make use of the following four categories: Read-pair (RP), Split-reads (SR), Read depth (RD) and assembly-based methods (Pirooznia et al., 2015). Each of these methods has its strengths and limitations. For instance, while RP-based methods are relatively accurate in identifying medium to large-sized SVs, small sized SVs are difficult to ascertain as small disruptions in insert size are difficult to separate from the normal background dispersion in insert size distribution (Medvedev et al., 2009). Therefore, a variety of tools has been developed that combine multiple methods like e.g. LUMPY (Layer et al., 2014) which combines RP and SR, and GenomeSTRiP (Handsaker et al., 2015) which combines RP and RD approaches.

In this study, we used LUMPY (Layer et al., 2014) to identify SVs from whole genome re-sequencing data of 29 cattle representing African, European and Zebu cattle. In European cattle, we investigated commercial cattle that represent individuals selected intensively for traits related to fertility and milk production, and southern European cattle which have short selection history with emphasis on traits related to meat production. Therefore, our goal was to identify SVs that might be associated with phenotypic variations observed across these diverse cattle populations. Additionally, we also identified SVs from whole genome sequencing data of an ancient aurochs sample (Park et al., 2015) to further facilitate variant comparison with modern cattle breeds.

6.2 Material and Methods

6.2.1 whole genome sequencing data

The samples used for CNV identification are described in Table S1. DNA extraction was performed from either blood, hair roots or semen samples which were collected from 19 individuals of Balkan-Italian and Iberian cattle breeds. Library construction for the whole genome sequencing (WGS) was carried out with 0.5-3 μ g of genomic DNA following the Illumina library preparation protocols (Illumina Inc.). Following library construction, all 19 DNA samples were paired-end re-sequenced with 150 bp read length using the Illumina sequencing technology (Illumina Inc.). An additional 10 WGS data (Table S1) of individuals representing several commercial and traditional cattle breeds were downloaded from the NCBI short read archive; all of these sequences were included in the data published previously (Bickhart et al., 2016a; Daetwyler et al., 2014; Kim et al., 2017; Murgiano et al., 2014).

6.2.2 Raw fastq file processing and alignment

To perform the quality-based trimming of each fastq file, the program Sickle (Joshi and Fass et al., 2011) was run with default settings (except for the length threshold of 50 bp). BWA-mem (Li, 2013) algorithm was used to align the sequences against the bovine reference genome build UMD 3.1. Following the alignment, samtools rmdup was used to remove PCR duplicate reads from the bam files (Li et al., 2009). Finally, local read re-alignment was performed using “RealignTargetCreator” and “IndelRealigner” with arguments as implemented in Genome analysis toolkit 3.1 (GATK) (Li et al., 2009).

6.2.3 Detection of structural variation using *smoove*

We used the *smoove* pipeline (<https://github.com/brentp/smoove>), which wraps various existing software and adds a number of internal read-filtering, to reliably call and genotype SVs (Figure 6.1). First, we extracted discordant read pairs and split read mapping to separate

sam files using SAMBLASTER (Faust and Hall, 2014); later, we used samtools (Li et al., 2009) to sort and convert these sam files into bam files. Following this extraction, we used *smoove* to call SVs. The approach implemented in *smoove* filters out spuriously aligned reads from bam files containing discordant read pairs and split read mapping based on multiple criteria from bam files containing discordant read pairs and split read mapping (described at: <https://brentp.github.io/post/smoove/>) before using LUMPY (Layer et al., 2014) to call SV. The algorithm implemented in LUMPY uses a number of alignment signals such as discordant paired-end and split read mapping to determine break-points and SV type. Following this single sample SV calling, we used SVTyper (Chiang et al., 2015), which is internally implemented in *smoove*, to extract, merge and genotype all the sites from all the groups.

6.2.4 Post-filtration and SV processing

The LUMPY SV calling approach does not consider read depth evidence exclusively, therefore, we developed an in-house python script which used pysam and pyvcf module to add “CV” (normalized coverage) tag in the vcf file containing genotyped SVs. The in-house python script estimates “CV” by dividing the average read-depth across the genomic regions containing SVs by the average read-depth across the entire genome. To reduce the discovery of false positive SVs, we used the following two criteria: 1) from the combined dataset, we excluded all SVs for which the breakpoint positions fell within 100-bp of a gap reported in the reference sequence, 2) we excluded all SVs for which more than 25% of the bases fell within a gap (Letaief et al., 2017). Following this early round of SVs filtering, we merged all SVs that displayed the same SV type/event (either duplication, deletion or inversion) with more than 90% overlap. Later, we set the SVs genotype of an individual as missing if it did not satisfy at least two of the following criteria: 1) CV in a sample should either be less than 0.8 times (for deletion) or greater than 1.2 times (for duplication) of the average coverage in the same sample, 2) the total number of paired-end evidence in each sample showing SV should be greater than or equal to three, 3) the total number of split-reads mapping in each sample showing SV should be at least one, 4) CV should be less than 0.8 times (for deletion) or greater than 1.2 times (for duplication) of the CV for the same genomic region in an individual with a single copy.

6.2.5 CNV detection using CNVnator

CNVnator (Abyzov et al., 2011) was also used to detect CNVs from whole genome sequencing data. This additional approach of CNV detection was employed for two reasons: 1) the overlapping results from CNVnator along with *smoove* approach were used to identify genes most likely encompassed by CNV, 2) to detect CNVs from WGS of ancient aurochs sample which was generated using single-end library resequencing. The details about the aurochs sample and downstream processing of the bam file are described in Park et al., (2015). In brief, the algorithm as implemented in CNVnator uses the principle that any given genomic regions harboring a CNV event displays differential read depth compared to a normal single-copy genomic region. Moreover, the algorithm also performs read depth correction based on the GC-content of the

genome. To avoid identification of false positive CNVs due to segmental duplications and low complexity regions in the genome, we filtered out CNVs for which the fraction of reads with low mapping quality was greater than fifty percent. Further, a CNV was also excluded if its breakpoints fell within 100-bp of a gap or when more than 25% of its length consisted of gaps reported in the reference sequence.

6.2.6 Annotation

Of the total SVs identified separately using *smoove* and CNVnator, we only retained those events (deletions and duplications) that showed overlap with each other. Later, we used the program Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016) to annotate the gene content of overlapped SVs identified for each approach—*smoove* and CNVnator. Finally, we only kept genes/regulatory regions for which the VEP predicted the similar consequences in both these approaches.

6.2.7 Comparison with SV reported for cattle in Databases

To validate the SVs identified in the study, they were compared with the Database of Genomic Variant archive (DGVa). For this purpose, we downloaded the “gvcf” file from the Ensembl database (ftp://ftp.ensembl.org/pub/release-93/variation/gvf/bos_taurus/) which contained SVs identified in the cattle genome in previous studies with DGVa.

6.3 Results

6.3.1 CNV statistics and sharing

A total of 29 whole genome sequences, with coverage ranging between 2.63x and 12.67x representing 23 European taurine, 4 African taurine and 2 zebu cattle were obtained, covering a broad range of bovine phenotypic diversity (Table S1). We identified CNVs, mainly using split reads and paired-end reads algorithms as implemented in LUMPY (Figure 6.1A). Additionally, to reduce the detection of false positive SV in our dataset, we also implemented filtering criteria based on the read depth information using an *in-house* python script. LUMPY as implemented in the *smoove* pipeline identified a total of 36,676 SV events across all samples. After applying various filtering criteria and merging the events showing 90 percent of overlap, the number of SV were reduced to 11,185.

About 35% of the total SVs were a singleton. We discovered a large variability in the number of different SV events. In general, deletion events were significantly over-represented compared to duplications and inversions (Table S2). Also, the number of SVs per individual varied from 681 to 3437, displaying large variability across cattle samples (Figure 6.1b). On average, African

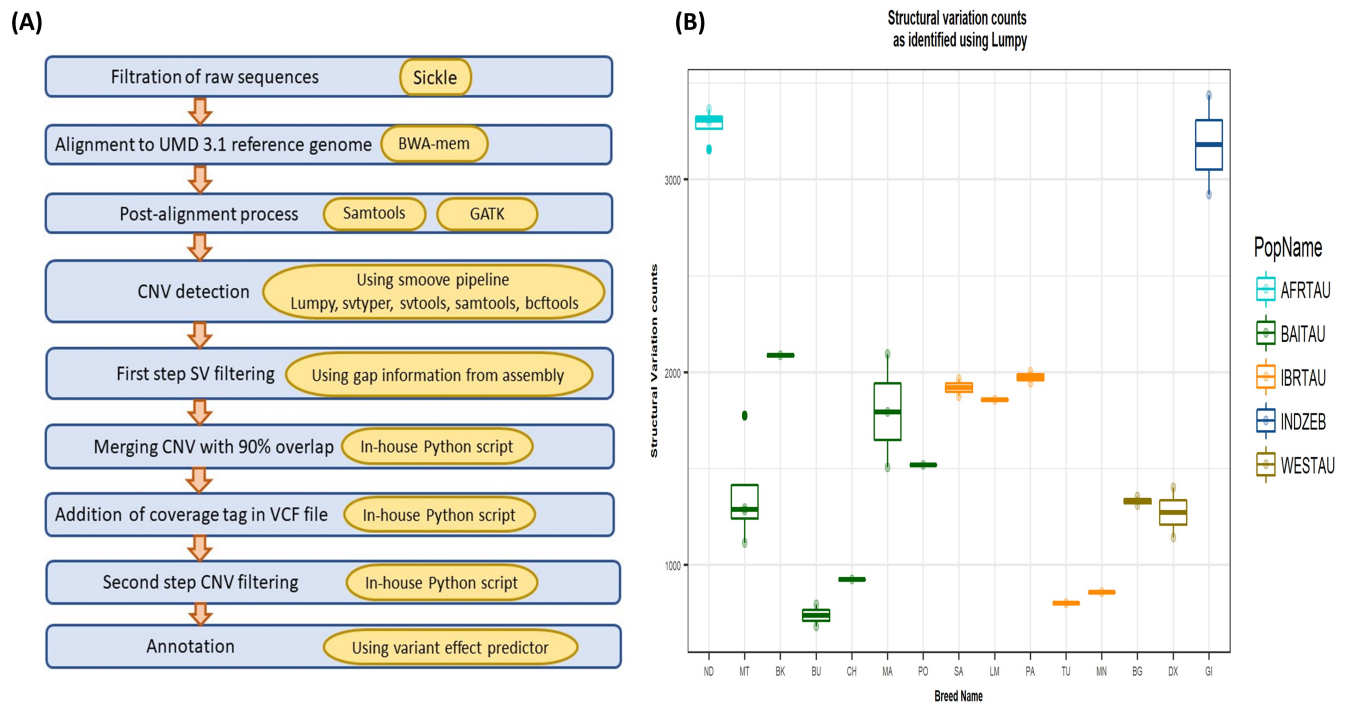


Figure 6.1: A). Bioinformatics workflow used to detect SV wrapping the pipeline implemented in *smooove*, B). Distribution of total SV counts per individual across different cattle populations. Abbreviations: AFRTAU-African taurine, BAITAU-Balkan and Italian cattle, IBRTAU- Iberian cattle, WESTAU- Western European cattle, INDZEB- Indian zebu.

(AFRTAU) and zebu (INDZEB) individuals displayed a higher number of events compared to European individuals.

A comparison with the SVs presents in the DGVA showed that about 40% of total SVs identified in the present study overlaps with the genomic positions of SV reported in the previous studies. However, we noted several SV discrepancies between DGVA and the present study. For instance, between 45.6 and 47.6 Mb on Chromosome 2, DGVA only reports an inversion, while we also observed many deletions in the same region in addition to a small inversion.

To investigate the distribution of SV among individuals, we calculated the proportion of SV sharing between every pair of individuals. To account for variability in total SV count across individuals, the proportion of SV sharing was calculated as the total number of shared SVs divided by the sum of total SVs identified in both individuals. As expected, the heatmap clustering results (Figure 6.2A) indicates a high degree of sharing of SVs between individuals of the same breed or/and similar geographic origin (Figure 6.2A). Next, we also performed a principal component analysis (PCA) to assess the clustering pattern based on SVs. As expected, our results presented in Figure 6.2B separated zebu from the taurine populations, and African taurine from European taurine populations. PCA, excluding African and zebu cattle, also

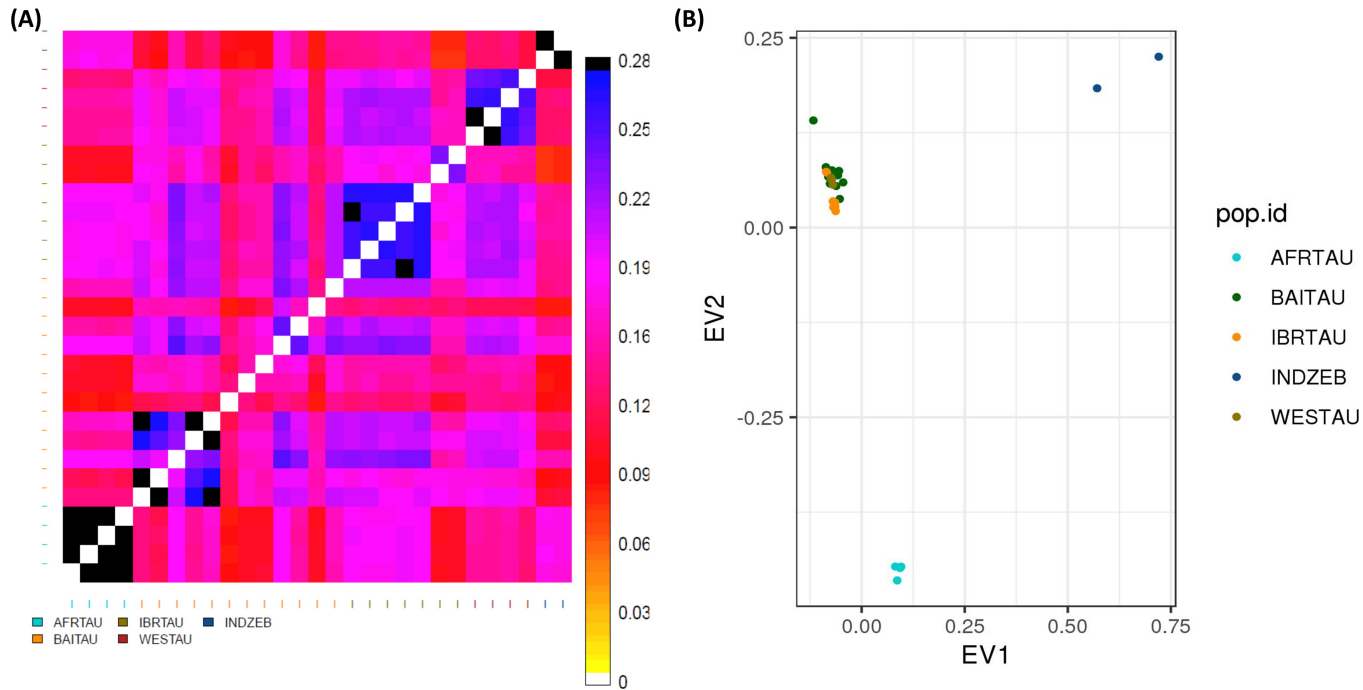


Figure 6.2: A). Heatmap plot shows the proportion of CNV sharing between each pair of individuals. B). Principal component analysis performed on genotyping SV; EV1 and EV2 represent the first and second principal component, respectively.

separated the clusters of Iberian and Italian cattle with Western European positioned at the center of the plot (Fig S2). However, we note that Italian individuals (two Buscha and one Chianina) with low coverage (between 2X and 6X) and two Iberian individuals (one Maronesa and one Tudanca) formed a cluster with western European cattle.

6.3.2 Functional annotation

One of the goals of this study was to identify genic regions encompassed by SV in a diverse set of cattle populations. However, accurate identification of such regions is difficult in case of low sequencing coverage data ($<10X$). Therefore, we developed a pipeline (Figure S1) combining the output from *smoove* and CNVnator to identify accurately the genic and regulatory regions encompassed by SVs (described in the method section). Functional annotation identified a total of 449 genes that displayed the same type of consequences as predicted by the VEP in the overlapped 441 SV detected across *smoove* and CNVnator outputs (Figure 6.3). For example, VEP predicted transcript ablation (or deletion of transcript feature) for various protein-coding genes located between 108.52 Mb and 109.29 Mb on Chromosome 6. This chromosomal segment encompasses SVs that were identified in many individuals and contained genes such as *PDE6B*, *SLC26A1* and *FGFRL1* (Table 1 and Table S3). VEP also predicted transcript ablations for

various other genes such as *GRM6*, *ARID1B*, *ALKBH5* and others which encode proteins with functions related to immunity and olfactory activities (Table 1 and Table S3). Furthermore, we also identified duplication events involving transcripts of various protein-coding genes such as *APOH*, *GLYAT* and *IFITM1* (Table 1 and Table S4). Several genes such as *TRAV16* and *TRAV17* displayed multiple types of consequences (ablation and amplification) across the cattle populations. Additionally, we also identified SVs near or in the regulatory regions within various genes such as *GALNTL6*, *AOX1*, and *LEPR* (Table 1 and Table S5).

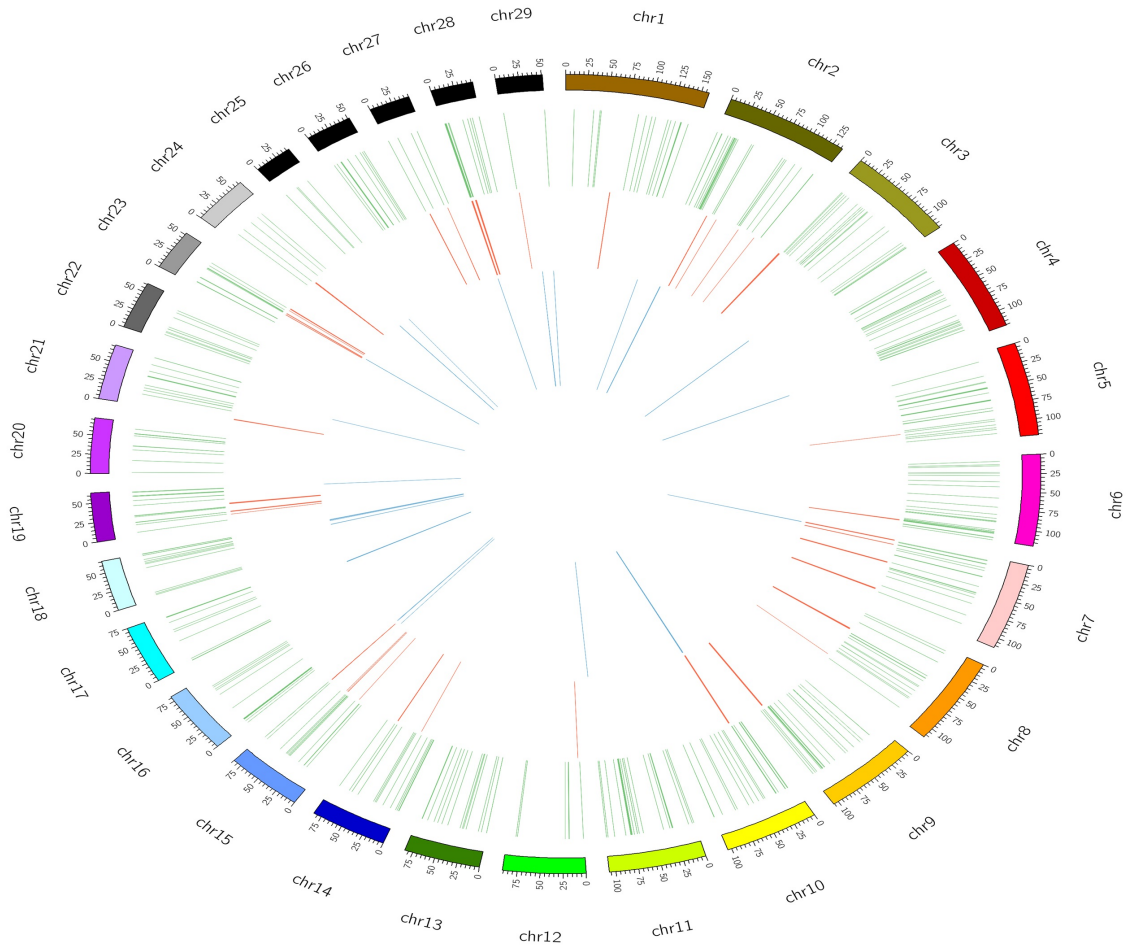


Figure 6.3: Circular plot showing the distribution of SVs identified within the regulatory regions or/ within the coding regions of genes on UMD 3.1 genome. The innermost layer (blue lines) represent the SVs for which VEP predicted transcript amplification. The intermediate layer (red lines) represent the SVs for which VEP predicted transcript ablation and the outermost layer (green lines) represent the SVs which encompasses the regulatory regions of various genes.

From the unique variants, that were identified after comparison with DGVa, we identified lineage-specific SV events in our dataset. Comparing two zebu samples against the pool of African and European taurine samples led to the identification of an intronic region of the *HMGA2* gene that was duplicated in both the zebu individuals but had a 2n copy number in the taurine individ-

uals (Figure 6.4). Interestingly, comparing four African taurine samples against the remaining samples led to the identification of several SV events that are specific to African taurine samples (Table S6). For instance, a duplication event involving the protein-coding *GALNT15* gene was identified only in African taurine individuals (Figure 6.5). Among European taurine samples, several *BAITAU* specific SV were detected in genes such as *HERC2* and *APOH*.

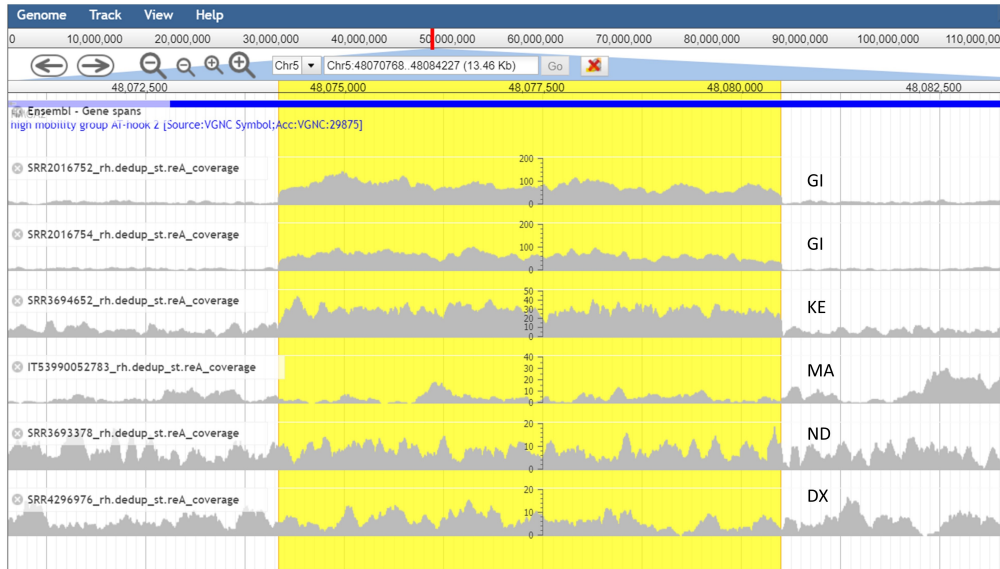


Figure 6.4: Jbrowse screenshot showing the coverage of zebu and taurine WGS alignment for duplication identified as lineage specific in zebu. Abbreviations: GI-Gir, KE-Kenana, MA-Maremmana, ND-N'Dama, DX-Dexter.

6.3.3 CNV in aurochs

After performing various quality filtering steps on CNV identified using CNVnator, a total of 537 deletions and 212 duplications were detected in the aurochs sample. Comparison of CNVs identified in aurochs with domestic cattle (in our dataset) led to the identification of three common genomic positions encompassing SV. Interestingly, of the three SVs, only one displayed the same type of event (a deletion) between aurochs and domestic cattle in our dataset. This deletion event spanned from 52.425 to 52.428 Mbp on Chromosome 23 (Figure S3). However, as the deletion events in the aurochs sample are difficult to ascertain given the low sequence coverage ($\sim 6X$) and fragmented nature of the DNA sequences from the ancient genome, we compared the duplicated CNVs identified in aurochs with duplicated CNVs reported for cattle in DGVa. Of the 212 duplications identified in aurochs (Table S7), 171 duplications overlap with duplications reported for cattle in DGVa. To rule out the spurious duplication events from the remaining 41 aurochs-specific duplications, we filtered out events that had gaps 1000 bp upstream or downstream of the events. Finally, we identified 24 duplicated events specific to the aurochs sample (Table S8).

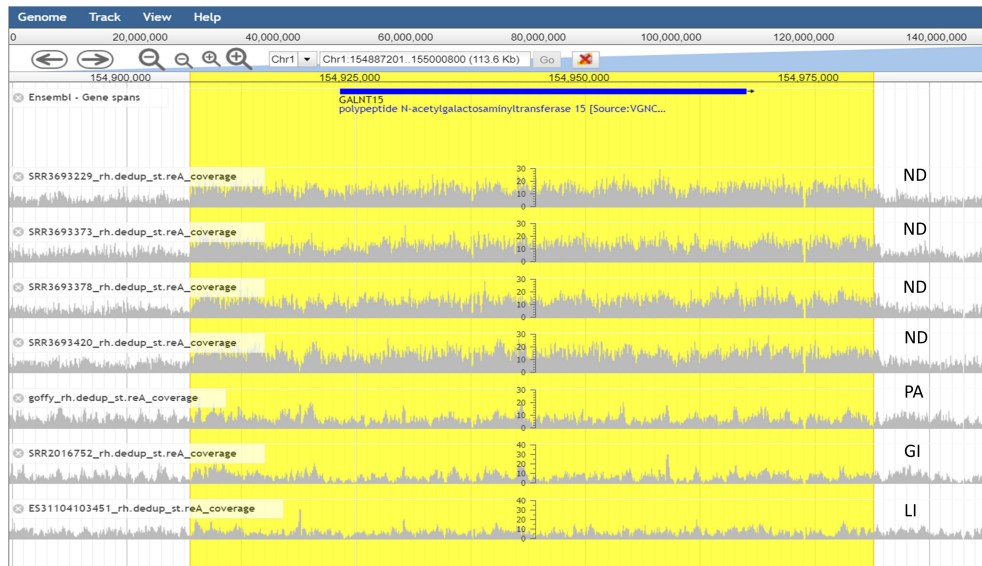


Figure 6.5: Jbrowse screenshot showing the coverage of zebu and taurine WGS alignment for duplication identified as lineage specific in African taurine. Abbreviations: GI-Gir, ND-N'Dama, PA-Pajuna, LI-Limia.

6.3.4 Discussion

Cattle populations across the world, exhibit a large phenotypic diversity which can be attributed to various genomic variations such as SNP, INDELs and SV. In this study, we used WGS data of taurine and indicine cattle to investigate SV profiles and identify various genes harboring SV.

Despite the advancement in next-generation sequencing and computation technology, the problem of accurately identifying SVs persists (Freeman et al., 2006; Zhan et al., 2011). For instance, Zhan et al. (2011) reported low overlap in CNVs identified using three different approaches—CGH array, WGS and, high-density SNP array- in the same set of individuals. Therefore, the authors suggested to only consider SVs identified from using multiple tools/approaches to reduce false positives. Hence, in our present study, we used LUMPY as implemented in the *smoove* pipeline to identify SVs. The algorithm in LUMPY uses discordant read-pairs (including split reads) information. Additionally, we also applied read-depth based filtering to obtain high-quality SV as the combination of read-depth and discordant read-pair mapping approaches has previously been shown to produce a high confidence CNV set (Zhou et al., 2017). Before downstream processing, we removed SVs that had more than 25 percent of bases located within gaps or SVs that had break-points located within 100 bp of a gap identified in UMD 3.1 assembly because alignments in such regions can be misinterpreted as deletions.

Like many previous studies, that used WGS data to identify SVs in cattle, we also report an abundance of deletion events compared to that of duplication and inversion events. Though there could be a possible biological explanation for the abundance of deletion events— such as non-allelic homologous recombination (NAHR) events producing more deletions than duplications

Table 6.1: Some examples of genes encompassed by SV identified in the study

Gene	Function/related to	UMD 3.1 coordinates of SV	Types of SV	Identified
<i>FGFRL1</i>	Formation of slow muscle fibers	Chr6:108528590-109293009	Deletions	Hou et al., (2011), Bous-saha et al., (2015), this study
<i>POLR2K</i>	synthesizing messenger RNA	Chr14:23735715-23737483	Deletions	This study
<i>ALKBH5</i>	modification of messenger RNA	Chr19:35022001-35044800	Deletions	This study
<i>LEPR</i>	Fat metabolism	Chr3:80120431-80124982	Deletions	This study
<i>HERC2</i>	Pigmentation in human	Chr2:644410-671787	Duplication	This study
<i>GLYAT</i>	Detoxification	Chr15:83435290-83493549	Duplications	Liu et al., Bickhart et al., (2016) and this study
<i>AOX1</i>	Detoxification	Chr2:89573026-89595342	Duplication	Hou et al., (2011), Bickhart et al., (2016) and this study
<i>APOH</i>	lipoprotein metabolism	Chr19:63210546-63270976	Duplication	This study

(Turner et al., 2008) — an algorithmic bias against duplications cannot be ruled out. Validation of SVs carried out by comparison with DGVa resulted in an overlap of about 40% of the total SVs identified in the present study. Interestingly, a previous study on 100 individuals from twelve different cattle breeds (Mielczarek et al., 2018) also reported a similar value of overlap with DGVa. These results show that despite the relatively small sample size investigated in our study, the inclusion of individuals from diverse cattle populations successfully captures a large proportion of previously reported variants.

We discovered, perhaps as expected, large differences in average SV counts per individual with highest values reported for zebu and African taurine individuals. This result is in contrast with the study of Bickhart et al., (2016), which reported comparable CNV counts for taurine and zebu individuals. Diversity in copy number variations has shown to be associated with population demography (Paudel et al., 2013; Upadhyay et al., 2017). For instance, using WGS data, Paudel

et al., (2013) reported a higher count of copy number variable regions in Asian pigs compared to European pigs, which they attributed to a difference in effective population size. Therefore, it is likely that the differences that we observed in SV counts among different cattle populations are due to population demography. Although part of the differences in SV counts can also be attributed to the fact that the UMD3.1 reference genome used for alignment is assembled from sequences of a Hereford (European taurine) cow. The pattern of population clustering using PCA separated individuals based on their geographical origin, except for three BAITAU and two Iberian individuals. It is worthwhile to note that these three BAITAU individuals had genome coverage less than 6X and that the two Iberian individuals are likely recent crossbreds.

In the present study, we identified various protein-coding genes and regulatory elements encompassing SVs which represents valuable resources for future studies aimed at identifying relationships between copy number variation and phenotypic variation in cattle (Table S3 to S5). To reduce the identification of genes covered by false positive SVs or inaccurate SV break-points, we only report the genes for which VEP predicted the same type of consequences of the SV identified by both CNVnator and *smoove*. Therefore, we suspect that with this approach, we might have missed many genes covered by true SVs. For instance, *smoove* identified a duplication of the entire *KIT* gene in African cattle, while CNVnator identified duplication in the regulatory region of the *KIT* gene in these individuals. Nevertheless, this approach identified 449 genes that were either covered by an SV or had an SV in their downstream or upstream sequences. Of the total number of genes identified as being encompassed by SVs, many have already been reported. For instance, the *GLYAT* gene (Bickhart et al., 2012, 2016b), which plays an important role in detoxification of metabolites, showed a duplication in Iberian and British individuals (Table 1). Apart from *KIT*, we also identified a duplication of part of the *HERC2* gene in two Maremmana cows, which has been associated with hair colour and skin pigmentation in human (Visser et al., 2012).

We also detected an SV in the regulatory region of *GALNTL6*, which has been associated with feed efficiency and growth traits in cattle (Doran et al., 2014; Seabury et al., 2017; Zhang et al., 2015). We also observed deletions in the third and sixth intron of the *LEPR* gene in two African taurine and one Iberian individual, respectively. The *LEPR* gene encodes a receptor for Leptin and its function is associated with fat metabolism and therefore, its role in regulating body weight is well described (Bolormaa et al., 2014). Interestingly, Shi et al. (2016) reported an association between copy number within the third intron of the *LEPR* gene and its expression. Further, deletions in the regulatory region of the *CAST* and *CAPN13* gene were also detected in three African taurine and one zebu individuals. Previous studies have associated these genes with meat tenderness and quality (Barendse et al., 2007; Casas et al., 2006; Tizioto et al., 2013). However, no CNV had been identified in this region of the *CAST* gene so far. Concerning adaptation and heat stress response, SV in the *DNAJ* gene families (*e.g.* *DNAJ12*, *DNAJ13*, *DNAJ18*), which act as a cofactor for the heat shock proteins, were also identified.

We also identified a breed-specific duplication event encompassing the *GALNT15* gene in all the samples of African taurine origin. Also, an indicine-specific duplication event encompassing an intron of the *HMGA2* gene was identified. We also validated this duplication event in additional samples of African zebu. Furthermore, this duplication event has previously been reported in

almost all zebu individuals (Bickhart et al., 2016). The *HMGA2* gene has been associated with body stature in human and various domestic animals (Goddard et al, 2011; Plassais et al., 2017; Bouwman et al., 2018). Comparison of duplications in domestic cattle and aurochs revealed that approx. 80% of the duplications identified in aurochs overlap with the duplications reported for cattle in DGVa. The functional annotation of these duplications revealed that the majority of these duplications cover genes with uncharacterized protein function and orthologs of the olfactory receptor and immunity-related genes (Table S9). Although such comparison was not performed for the deletions, one deletion that was identified as shared between ancient aurochs and domestic cattle (in the present study) encompasses an ortholog of an olfactory receptor-like gene. A possible selection on these shared events can also be inferred given the evolutionary advantage the process of olfaction and immunity provide to livestock species (Paudel et al., 2013). Nevertheless, these results suggest the ancestral origin for many SV events identified in modern cattle. Interestingly, Paudel et al., (2013) also reported a large overlap of copy number variable regions between wild and domesticated pigs.

We note that usually many difficulties are associated with accurate identification of SVs using short-read sequencing technologies. For example, short reads are difficult to map on the repetitive/complex regions of the genome which makes the identification of SV in these regions inaccurate. However, mapping in such regions can be accurately determined using reads generated by long-read sequencing technologies, such as the PacBio or Oxford nanopore platforms, which produce reads with an average length of about 15 kbp. Moreover, such long reads can even span entire SV events. Therefore, these will allow the identification of SV more accurately and comprehensively compared to short-read technologies (Huddleston et al., 2017; Viluma et al., 2017). Using a combination of SVs detection tools and stringent filtering criteria, we generated the SVs profile of diverse cattle populations. Further, functional annotation of these SV identified genes underlying a variety of traits related to coat color, metabolism and meat quality. In addition, we also identified lineage-specific SVs in various cattle populations. These results suggest the important contribution of SVs in cattle diversity. Moreover, a large number of duplicated events identified in an ancient aurochs sample overlapped with those identified in modern domestic cattle, indicating that these events might have been segregating in ancestral aurochs population. However, a larger number of sequenced aurochs samples are needed to explore this hypothesis further. To conclude, our study highlights the important contribution of SVs in relation to cattle diversity and evolution.

6.4 Funding

MRU benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate “EGS-ABG”.

6.5 Author Contributions

MRU, MAG and RPC conceived, designed, planned and directed the study. MRU and DM devised a pipeline to identify structural variations. MRU performed the downstream analysis. MRU, MAG, RPC, DM and GA interpreted the results. MRU prepared the figures and wrote the manuscript with input from all authors.

Chapter 7

General discussion

7.1 Introduction

European cattle display vast phenotypic diversity which can be attributed to genomic variations such as single nucleotide polymorphisms (SNPs) and structural variations (SVs). The distribution of these genomic variations in a population is heavily influenced by different population genomic forces such as migration, drift, and selection. In this thesis, genomic variations were characterized in traditional and primitive European cattle breeds using genome-wide SNPs. Specifically, hypotheses concerning gene flow from zebu, African taurine and wild local aurochs ancestry were investigated in detail. To understand the patterns of genomic variations comprehensively, I also characterized the structural variations in the genome of European cattle. In this final chapter, I will discuss the main findings of all the previous chapters in the context of existing literature and knowledge about the genetic structure, admixture, and variations in European cattle.

7.2 Patterns of genomic admixture

7.2.1 On gene flow between European and non-European cattle

The divergence between populations is directly proportional to the time since they shared a most recent common ancestor and differential selection pressure they experienced in their respective environments unless gene flow occurred in these populations. Indeed, the dynamics of population divergence is heavily influenced by gene exchange between isolated populations. In general, gene exchanges between previously isolated populations counter the divergence due to population scaled mutation rate and genetic drift. This demographic model, which is also known as Isolation with Migration (IM), has been investigated widely to explain the genomic divergence observed in a various population of livestock and wild species. For instance, studies have reported the presence of a high frequency of mtDNA haplotypes of Asian origin in various European pig breeds due to admixture. In fact, Bosse et al., (2014) also identified introgressed Asian pig haplotype in European domestic pigs which most probably contributed to increased fertility. These results are in good concordance with the historical record of the early nineteenth century which mention the import of Chinese pigs in Europe because of the renowned fertility of Chinese pigs. Another example is the introgression from Chinese pigs into European pigs of the regulatory gene variant at the porcine IGF2 gene that explains increased muscle growth (Van Laere et al., 2003). However, even though historical records associated with import/migration of zebu cattle are scant, the gene flow from indicine cattle in many European cattle breeds has been hypothesized. For example, based on the similarity of a β -globin variant, Pieragostini et al. (2000) proposed a contribution of zebu cattle in the gene pool of Podolica cattle. Furthermore, analyzing microsatellite markers in different Eurasian cattle breeds, Cymbron et al. (2005) reported that among all mainland European cattle breeds which they studied, Italian cattle breeds—particularly Maremmana and Modicana—followed by Greek cattle breed—Sykia—displayed the highest frequency of indicine population-associated alleles (PAA). They proposed a

Near Eastern origin for this indicine ancestry in Italian and Greek cattle breeds. This hypothesis was further supported by the identification of indicine mtDNA haplotypes in individuals of the Ukrainian Whitehead cattle breed (Kantanen et al., 2009). Further, analyzing genome-wide SNP data, McTavish et al. (2013) reported indicine ancestry in multiple southern European cattle breeds, and they also proposed a north-south gradient of indicine ancestry in Europe. Decker et al. (2014), however, refuted this hypothesis as they reported indicine ancestry only in three Italian cattle breeds—Chianina, Romagnola and Marchigiana. Nevertheless, all these studies lack in the genetic information of cattle breeds from the Balkan region which lies between Anatolia and Italy and therefore, may provide a more comprehensive understanding of indicine ancestry gradient in European cattle breeds.

In this thesis, I used genome-wide SNPs genotyped in different cattle breeds of Balkan and Italian regions (BAI) to characterize indicine ancestry in detail. Using unlinked SNPs and a haplotype-based approach, I show that indicine ancestry is a common feature of several BAI breeds. In chapter 2, I carried out standard population genomics analyses (such as ADMIXTURE and D-statistics) based on high-density SNP array data and proposed that high divergence of BAI breeds can be attributed to indicine ancestry. Interestingly, the signals of indicine ancestry were not observed in any of the Iberian cattle breeds that were investigated, confirming the previous hypothesis that indicine ancestry is uncommon in southern European breeds (Decker et al., 2014). Further, in chapter 3, I carried out a detailed characterization of indicine ancestry in European cattle and showed that different Italian cattle breeds as well as the breed called Busa, of Balkan origin—display a similar proportion of indicine ancestry in their genomes. These results could indicate that BAI breeds received this indicine ancestry from a common ancestor and subsequently, differentiated relatively recently. However, ADMIXTURE analysis is known to be affected by sample size, and moreover, several demographic scenarios often lead to same ADMIXTURE patterns as noted by Lawson et al. (2018). Similarly, the result of D-statistics does not necessarily imply gene flow between the lineages as a subdivision of ancestral populations, if this remains persistent for a long time, also leads to signals similar to recent gene flow (Theunert and Slatkin, 2017). However, sub-structure is unlikely to affect these results as Indian and European cattle have been domesticated independently. Nevertheless, based on the results of chapter 2 and chapter 3, I propose several models as shown in Figure 7.1 that can be tested on whole genome sequencing data (WGS) using a Bayesian approach for thorough investigation of demographic events in BAI breeds.

The fact that BAI breeds still display indicine ancestry in their genomes indicates the possibility that indicine genomic segments might be under selection because of some adaptive advantages they confer to BAI breeds. Indeed, this phenomenon—also known as ‘adaptive introgression’—whereby introgressed segments from distantly related populations provide increased fitness to the donor population, has been reported in many animal species (Hedrick, 2013). For instance, Song et al. (2011) identified a large genomic segment in a new world mouse population (*Mus musculus*) which has been introgressed from old world mice and contained the warfarin resistance gene *vkorc1* encoding the vitamin K epoxide reductase subcomponent 1. The BAI cattle breeds display many zebu-like traits such as adaptation to relatively hot climates and better general disease immunity. In fact, Modicana, which is an Italian cattle breed, displays

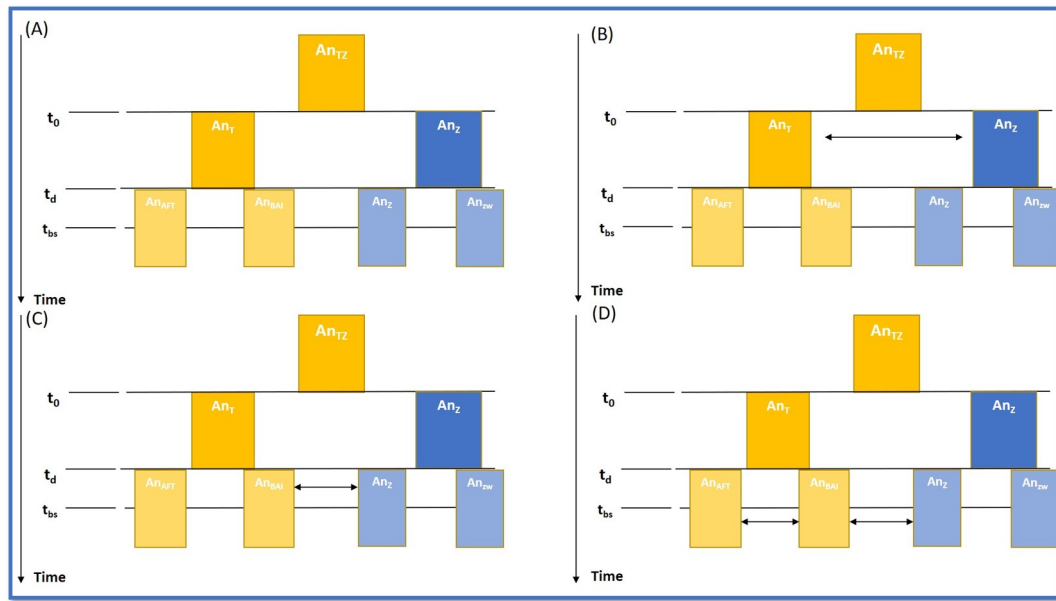


Figure 7.1: Schematic of the proposed demographic models to be tested on whole genome sequencing data. Double-headed arrows represent migration events that should be modelled as two continuous parameters. Barring the model (A), which represents null model without migration events, all other demographic models include migration events. The term “An” refers to the ancestral effective population size or simply, effective population size. Other abbreviations used as subscripts: TZ- term for the ancestors of taurine and zebu before they split, T-Taurine, Z-Zebu, AFT-African taurine, BAI- Balkan and Italian taurine, ZW: ancestral wild Zebu. The “ t_0 ” refers to the number of generations (back in time) in the past at which the ancestral taurine and zebu population separated. The “ t_d ” refers to the number of generations (back in time) at which the domestic cattle separated from their wild ancestors. The “ t_{bs} ” refers to the number of generations (back in time) at which the African cattle separated from the European domestic cattle.

bifid processes in the last thoracic vertebrae— traditionally considered as a zebu-specific characteristic (Grigson 2000). Therefore, an intensive sampling of various BAI breeds is needed to investigate this hypothesis of adaptive introgression.

Many studies analyzing uniparental markers such as mitochondrial DNA and Y chromosomal haplogroups as well as analyzing microsatellite and genome-wide SNP markers have identified African cattle ancestry in various southern European cattle breeds (Beja-Pereira et al., 2003; Cymbron et al., 2005, Ginja et al., 2010a, 2010b; Decker et al., 2014). In fact, Decker et al. (2014) reported indicine as well as African taurine cattle ancestry in central Italian cattle breeds of Chianiana, Romagnola, and Marchigiana. In Chapter 3, using a haplotype-based approach with genome-wide SNP markers, we proposed that other BAI cattle breeds like Busa and Maremmiana also display shared ancestry with non-European cattle which is quantitatively similar to central Italian cattle breeds. Moreover, ADMIXTURE analysis of high-density SNP array data also identified signals of shared ancestry between Iberian and African taurine cattle. These results could be interpreted as a legacy of the Moors who inhabited the Iberian Peninsula

between the 8th to 15th centuries and certainly would have brought livestock with them during their more than 600 years presence on the Peninsula. However, because European and African taurine cattle originated from the same domestication center, i.e., the Near East, the possibility of shared ancestry (without migration) cannot be ruled out. Shared genetic variation between relatively closely related populations but without migration has been observed for many species. For instance, by applying a Bayesian approach on microsatellite data, Sousa et al. (2012) showed that the observed genetic patterns in fish populations, which was attributed to admixture using a clustering-based approach, could be better explained by the demographic model with a population split but without admixture. However, introgression in BAI cattle breeds from a ghost population carrying both—African and indicine cattle—ancestry cannot be ruled out.

In this thesis, relationship between southern European cattle and East African zebu was also explored. Although, as described in chapter 2, no shared indicine ancestry between Iberian and Nellore (derived from Indian zebu) was observed, the inclusion of genotyping data of East African zebu in chapter 3 indicated shared ancestry between East African zebu and southern European cattle. However, as East African zebu itself is a cross-breed between African taurine and zebu, this signal of shared ancestry is difficult to interpret.

7.2.2 On gene flow between domestic European cattle and wild local aurochs

Backcrossing between the wild ancestor and its domesticated form is not uncommon in livestock species (Barbato et al., 2017; Frantz et al., 2015; Vilà et al., 2005; Frantz et al., 2013). For instance, Frantz et al., (2013) reported that wild local pigs and domesticated pigs in Eurasia interbred quite often, contrary to the general assumption of reproductive isolation between these two species. However, such events of interbreeding between domesticated cattle and local wild aurochs are highly debated. Before Park et al., (2015) first published the WGS data of British aurochs, researchers had used only uniparental markers (mtDNA and Y-chromosome SNPs) to investigate this research question (Achilli et al., 2008, 2009; Bollongino et al., 2008; Bonfiglio et al., 2010; Götherström et al., 2005; Svensson and Götherström, 2008). While mtDNA studies identified novel haplogroups in Italian cattle breeds such as Chianina and Romagnola, supporting some level of aurochs introgression in Italian cattle breeds (Bonfiglio et al., 2010), the results based on Y-chromosome analysis has remained inconclusive (Bollongino et al., 2008; Götherström et al., 2005). Park and colleagues (2015) analyzed WGS data of wild local aurochs in relation to worldwide cattle breeds, and they concluded, perhaps not surprisingly, that cattle breeds of Britain and Ireland share the highest level of genetic variants with the British aurochs sample among all the cattle breeds that they studied. Although they incorporated genetic data of more than 1200 animals, their dataset lacked in the genetic information of Iberian and some important primitive cattle breeds of BAI regions.

In this thesis, comparative analysis of genomic variants were performed between the British aurochs sample, which was used in the study of Park et al., (2015), and various primitive cattle breeds of Europe. The results as described in chapter 2 indicated instances of interbreeding between wild local aurochs and ancestors of domestic cattle. However, this gradient of derived

alleles across European cattle should be interpreted with caution as this analysis (D-statistics) can provide similar results even in case of a genetic structure in a population. Moreover, significant diversity existed among wild local aurochs as inferred by diverse mitochondrial haplogroups identified in ancient samples of wild aurochs (Bonfiglio et al., 2010). Therefore, the possibility of secondary geneflow between Italian domestic cattle and other distinct sub-population of local aurochs (not related to British aurochs) cannot be ruled out. Overall, our results not only reinforce the earlier findings of Park et al., (2015) but also provide an overview of the distribution of aurochs specific variants in major primitive cattle breeds of Europe.

In the future, availability of WGS data of ancient aurochs bones sampled from different parts in Europe and representing different time periods after the event of cattle domestication may provide more detailed insight into the level of introgression and possible adaptive advantage of these introgressed segments in extant European cattle breeds. Moreover, such studies also have the potential to provide insight into how livestock farming evolved over time, since the beginning of cattle domestication. For instance, a recent study (Bro-Jørgensen et al., 2018) analyzing mtDNA extracted from the horn of the last aurochs bull, identified the T3 haplogroup which is the most common haplogroup in domestic taurine cattle. Based on this result, it can be speculated that the last individuals of the surviving aurochs population might already have exchanged gene-flow with domesticated cattle before they went extinct.

7.3 Patterns of Genetic relatedness/structure and demographic history

Knowledge of genetic relatedness, demographic history and genetic status of a population play a decisive role in conservation management. This information as sometimes recorded in historical literature is often biased or not available. Genetic markers such as SNPs or microsatellites serve as powerful tools to retrieve unambiguous breed information that can reliably be used to design a conservation program. The results as described in chapter 2-4 provided detailed insights into the relationship among European cattle populations and demographic history using genome-wide SNP markers.

7.3.1 Genetic relatedness/structure

The information about genetic structure enables the assignment of individuals to their genetic origin and to identify admixed individuals in a population (Herrero-Medrano et al., 2013; Negrini et al., 2009). In this thesis, unlinked SNP marker-based analyses such as PCA, ADMIXTURE, estimating genetic distance, and haplotype-based approaches as implemented in CHROMOPAINTER and the fineStructure pipeline, were performed to assess genetic structure of European cattle. Generally, high-density SNP arrays suffer from ascertainment bias which can, sometimes, distort inferences about population structure (Albrechtsen et al., 2010). Diversity-related statistics for Swedish traditional cattle breeds as reported in Chapter 4 of this thesis

indicates that the Genomic Profiler High-Density Bovine150K (GGP HD150K) array can reliably be used to assess population structure and genetic diversity in European traditional cattle breeds. A similar assessment was not made for the Illumina Bovine777K array as a relatively low number of samples per breed was genotyped using this array, and therefore, it may have provided biased estimations of the diversity statistics. However, as a large number of markers from this array has been validated with $MAF > 0.05$ across various Eurasian cattle breeds, it can be safely assumed that estimation of genetic structure was unbiased.

The results presented in chapter 2 clearly differentiate cattle populations based on their origin. In fact, genetic structure/population split as recovered in PCA can be interpreted as a legacy of ancient migration events involving Neolithic farmers and their livestock through the Danube River route and the Mediterranean Sea route. In chapter 3 and 4, population clustering using a haplotype-based approach not only successfully retrieved the clustering pattern of unlinked SNP based analyses (PCA and ADMIXTURE) but also provided additional information about sub-structures within a population. For instance, in Chapter 4, the fineStructure-inferred tree clearly identified sub-structure within a population of Swedish mountain cattle breeds that corresponds to the farms from which the samplings were carried out. These results suggest that a haplotype-based approach as implemented in fineStructure is a powerful tool which can even identify sub-structures in a population.

7.3.2 Demographic history

Selection and demography both play significant roles in how genomic variants are distributed among populations. While selection changes the allele frequency of local variants, demographic changes affect variants across the entire genome (Bosse et al., 2015). Moreover, current effective population sizes are of major interest in conservation management. However, due to the lack of pedigree information of local cattle breeds, this information is often difficult to retrieve. Therefore, in this thesis, inferences of demographic history and recent changes in effective population size were carried out using the analysis based on runs of homozygosity (ROH) and linkage disequilibrium (LD). The presence of ROH in a genome indicates relatedness between the parents of an individual. However, the length of ROH decreases over time as recombination events break down ROH into smaller pieces. Therefore, usually, the presence of small ROH in a genome indicates ancestral relatedness, while long ROH indicates inbreeding, e.g., consanguineous mating as the haplotypes have not had enough time to break-down (Bosse et al., 2012). Similarly, LD-based methods also infer past and recent effective population size where small values of r^2 at small genetic distances and rapid decay of r^2 values indicate vast haplotype diversity (large N_e), and large values of r^2 across large interval of genetic distances indicates small N_e (Hayes et al., 2003; Tenesa et al., 2007).

In chapter 2-4, the inferences of demographic events were drawn using the ROH profile of different European cattle populations. I show that the abundance of ROH counts in British cattle breeds indicates the combined effect of genetic isolation and long selection history. Further, low nucleotide diversity outside ROH as estimated using WGS data indicated a relatively small number of founding individuals. Conversely, the presence of only a relatively few ROH segments

and high nucleotide diversity outside ROH in several BAI individuals indicated large ancestral founding population or highly diverse ancestral population. Similarly, the abundance of long ROH in several individuals of Iberian and BAI cattle indicated frequent mating between related individuals which can be attributed to a recent reduction in effective population size. Indeed, the recent reduction in the effective population size of Mirandesa and Maltese as reported in Chapter 2-3 is in good agreement with historical and scientific literature concerning these breeds. For example, it was reported that last pure-bred Maltese bull was culled in 1990 and since then, semen of Chianina bulls, which showed Maltese-cattle like feature, were used to recover the population size of Maltese cattle (Lancioni et al., 2016).

In Chapter 4, we investigated demographic history of Swedish traditional cattle populations most of which had never been studied before. The results showed that despite a reduction in population size recorded until the end of the 20th Century, Swedish mountain cattle still has a large effective population size. Additionally, the analyses also suggested that despite genetic isolation the breeders have managed to avoid frequent mating between related individuals in the southern Swedish cattle breeds—Väne cattle and Ringamåla cattle.

7.3.3 Status of genetic diversity

In this thesis, I characterized the genetic diversity of European cattle breeds using data sets obtained by both WGS and high-density SNP arrays. Both these data suggested that southern European cattle breeds—especially BAI cattle—have relatively higher genetic diversity compared to Iberian and North-Western European cattle. This observation can be attributed partly to the complex admixture history of BAI cattle. Moreover, its proximity to the center of domestication means that a founder effect might not have been as strong as it may have been in Western Europe and Iberia. Moreover, estimates of genetic diversity were in concordance with a documented breed history for each of the breeds that were analyzed. For example, breeds with genetic isolation and the recent reduction in N_e like the Maltese and Mirandesa have low genetic diversity. Furthermore, in chapter 4, diversity statistics suggested that Swedish mountain cattle display high genetic diversity, while another subpopulation of SMC—Fjällnära—display high variation in genetic diversity.

7.4 Patterns of Structural variations

SVs such as insertion, deletions, inversions, translocations, and duplications are important classes of genetic variations. These variations can drive mammalian adaptive evolution. In livestock, studies have associated SVs with coat colour and other morphological traits. Moreover, like other genomic markers such as SNPs and microsatellites, the distribution of SVs in a genome is affected by demography and selection. However, compared to SNPs, SVs affect large portions of a genome. Additionally, they may contribute to individual fitness by influencing mRNA and protein expression levels, and therefore, subjected to selection. For instance, genes such as *CATHL4* and *ULBP17* which have been associated with parasitic infection, display a

difference in copy number between indicine and taurine cattle (Bickhart et al., 2012). Therefore, defining the distribution and nature of SVs in cattle genomes is crucial to understand the underlying genetic factors responsible for the observed phenotypic diversity between different cattle breeds. In the following paragraphs, I discuss the distribution of SVs in the cattle genome and some important genes encompassed by SVs that we identified in this study. Additionally, I also discuss the strengths and limitations of the tools that were for SVs identification in this thesis.

7.4.1 Structural variations and demography

In chapter 5, copy number variations (CNVs) in cattle genomes were identified using signal intensity data of bovine high-density SNP arrays. We show that, on average, BAI and British cattle display a significantly higher number of CNVs and non-redundant CNV regions (CNVRs) compared to Dutch and Alpine cattle. We also suggest that differential selection pressure and drift effects between cattle breeds can lead to differential CNV counts. However, to validate this hypothesis, additional samples representing different cattle populations need to be genotyped. Nevertheless, this observation is in agreement with a recent study by Mielczarek et al. (2018), in which they reported a significant inter-as well as intra-population variability in copy number loci between different European cattle populations. Similarly, in chapter 6, using WGS data, we also reported higher SV counts in African and Indicine cattle compared to European cattle. A study (Paudel et al., 2013) analyzing CNVs in Eurasian pig populations also reported higher CNV counts in Asian pigs compared to European pigs which they attributed to higher effective population size. On the contrary, Bickhart et al. (2016) observed comparable SV counts across different European and Asian cattle breeds. Therefore, part of the differences between SV counts across cattle populations can be attributed to the fact that the UMD3.1 reference genome, which was used for sequence alignment in this thesis, is assembled from sequences of a Hereford (European taurine) cow.

It has been shown that population inferences based on the pattern of SVs and SNPs produce identical results in geographically distinct populations (Jakobsson et al., 2008). In chapter 5, we show that CNVRs data successfully clustered individuals belonging to the breeds that displayed low genetic diversity using SNP data (such as English longhorn and Maltese). However, hierarchical clustering failed to cluster the individuals based on the geographical similarities, indicating the effect of small sample size and sharing of high frequent CNVRs. Moreover, the possibility of false positive CNVs distorting the sharing of CNVRs cannot be excluded.

7.4.2 Structural variation and functional annotation

Many CNVs affecting phenotypic traits related to coat colour and morphology have been identified in livestock as well as in companion animals (Durkin et al., 2012; Jakobsson et al., 2008; Salmon Hillbertz et al., 2007). For instance, Salmon Hillbertz et al., (2007) identified a 133-kb duplication in the genome of Ridgeback dogs which encompass three fibroblast growth factor (FGF) genes and causes hair ridge and predisposition to dermoid sinus. In chapter 5 and 6,

we identified SVs encompassing various genes related to important livestock traits. Additionally, in both these studies, we also reported over-representation of genes related to immunity and olfaction processes. In chapter 5, we identified and validated the structural variant (Cs29) encompassing the *KIT* gene in English Longhorn cattle. This variant was first identified in Belgian blue cattle and was shown to be associated with coat-colour sidedness (Durkin et al., 2012). Later, Brenig et al. (2013) identified the same variant in White Park and Galloway cattle. In addition, they also suggested a dose-dependent effect of Cs29 in these breeds. Interestingly, English Longhorn cattle also display considerable variation in coat colour, i.e., such as red, brown, grey or white. Therefore, it is likely that such a dose-dependent effect of the Cs29 variant might be responsible for coat colour variation in English Longhorn cattle.

In this thesis, SVs were identified in various genes related to metabolism, meat quality, and immunity-related traits. For instance, in chapter 6, we described SVs encompassing genes such as *CAST* and *CAPN13* that are associated with meat quality and tenderness (Barendse et al., 2007; Casas et al., 2006; Tizoto et al., 2013). However, to verify their effect on gene expression requires that transcriptome data be generated from relevant tissues.

In human, studies have suggested that a large number of SVs are shared across different populations (Sjödín and Jakobsson, 2012). SV can arise independently in a population and, if selected upon, can spread in a population. Studies have identified population-specific SVs in many cattle populations (Bickhart et al., 2016; Xu et al., 2016). In chapter 6, we identified several population-specific SV in African and Zebu cattle populations. Moreover, several novel SVs were also identified in primitive cattle breeds. For instance, a novel SV was identified in the gene *HERC2*. This gene has been associated with pigmentation in human (Visser et al., 2012). However, as I mentioned earlier in the section, verification of such novel SVs as identified in this thesis requires that many samples with recorded phenotypes related to coat colour and body conformation traits should be investigated using gene expression data.

7.4.3 Structural variation in ancient aurochs sample

In recent times, the advancement in experimental and bioinformatics approaches has led to sequencing and analysis of hundreds of ancient genomes (Orlando et al., 2015), which in turn, have transformed our understanding of population genomics forces leading to speciation and adaptation. Studies involving ancient genomes in livestock and humans have shown how the genetic make-up of populations has changed substantially over a short period of time owing to the selection pressure (Lazaridis et al., 2016; Orlando et al., 2013, 2015; Somel et al., 2016). Moreover, studies of ancient genomes also allow researchers to trace back the age of functional alleles across time. However, the fragmentation of ancient DNA due to post-mortem changes, limits the read length of DNA molecules between 60 to 150 bp length, which is shorter than the read length generated by Illumina sequencing technology (Miller et al. 2008; Briggs et al. 2009). Moreover, this fragmented nature of ancient DNA (aDNA) also prevents sequencing using paired-end approaches. Therefore, often single-end sequencing has been preferred to sequence aDNA and subsequently, read-depth approaches have been used to identify SVs (Lin et al., 2015; Sudmant et al., 2015). In this thesis, SVs were identified in aDNA prepared from an aurochs

sample using a read-depth approach as implemented in CNVnator. We reported that about 80% of the total duplications identified in the aurochs sample are still segregating among modern cattle. In fact, we also identified one shared deletion between ancient aurochs and the studied cattle breeds which likely have the same break-points. Therefore, it can be hypothesized that many of the SV between aurochs and modern cattle are identical by descent. Moreover, It is likely that such SVs might be under selection because of the adaptive advantage they confer. In human, a recent study has identified a shared deletion event between ancient Neanderthals and modern non-African human populations which the authors attributed to introgression from the Neanderthals (Sudmant et al., 2015). Similarly, we hypothesized that secondary introgression from aurochs in European cattle might have led to frequency differences of “introgressed SVs” between European and non-European taurine. However, such study awaits sequence data from ancient aurochs sample with substantially much higher coverage than those currently available.

7.4.4 Challenges of SVs identification in livestock

SVs can be identified from various types of data generated by WGS, comparative genomic hybridization (CGH) and SNP arrays (Alkan et al., 2011). However, studies have reported a low agreement in the SV identified from different data sources in the same individual (Pinto et al., 2011; Zhan et al., 2011). For instance, Zhan et al. (2011) identified SVs using three different platforms (WGS, SNP array, and CGH) in the same individual and observed only a maximum of 23% overlap among these platforms. Moreover, studies have also reported low agreement between different CNV callers used on the same platform (Legault et al., 2015; Pinto et al., 2011). Indeed, different algorithms used for SVs identification have their own strengths and limitations. For example, *Lumpy*, which uses split-read and discordant reads to identify SVs, can identify small SV events reliably compared to the large events in repetitive regions of a genome. Moreover, in other platforms such as SNP array, a large fluctuation in signal intensity data due to relatively bad DNA quality can lead to the identification of false positive SVs. Furthermore, sometimes SV identification algorithms are optimized using data only generated in humans, which makes the comprehensive SVs identification in non-model organisms difficult. Therefore, selection of proper tools and optimizing post-filtering strategies to generate reliable and reproducible SV set in livestock is a major challenge for researchers.

Across all the SVs identification platforms, the quality and quantity of SVs heavily rely on a good reference genome assembly. For example, in SNP array, the hybridization probes to capture the variants of interest are designed from the reference genome, while alignment against the reference genome is often the first step in re-sequenced data produced using WGS approaches. Therefore, it is essential that the reference genome is as complete, correctly assembled and error-free as possible so that SV can be reliably identified. Unfortunately, often the reference assembly in livestock genomes are incomplete with relatively high errors, which may cause misinterpretation of the underlying sequences involved in SV. For example, Zimin et al. (2012) identified 39 Mb of sequences which were incorrectly assembled as segmental duplications in the *Btau4.1* cattle reference assembly. Using an SNP array platform, Zhou et al. (2016) identified 9 frequent

false-positive copy number variable regions which were attributed to assembly errors. In fact, in chapter 5, we also reported that underlying probes covering the same regions show different signal intensities between cows and bulls, supporting the findings of Zhou et al., (2016). Also, in the same chapter, we identified an abundance of CNVs between 72 and 74 Mb region of chromosome 12, which partly can be attributed to assembly errors as the size of chromosome 12 in the new cattle genome build (ARS-UCD 1.2) is about 2 Mb shorter compared to that of UMD3.1. Therefore, working with incomplete genome-builds for SVs identification requires that results should be interpreted with caution.

Balanced SVs such as translocations and inversions, which do not change the overall copy number of the sequences, are difficult to identify with the current sequencing technologies. Therefore, this likely is one of the reasons for the underrepresentation of such events in genomic data obtained from livestock studies. Nevertheless, studies have shown that events such as translocations can have a significant impact on phenotypic diversity like, e.g., the *KIT* gene translocation affecting coat colour in Belgian blue and Brown-swiss (Durkin et al., 2012). In chapter 5, we also confirmed that this translocation is quite frequently present in British cattle breeds. However, the lack of phenotypic information meant that the effect of this translocation in these breeds could not be investigated. In fact, linking phenotypes with underlying genotypes is the most crucial goal of livestock genomics. Therefore, detailed and extensive phenotypic information from large numbers of individuals is essential to allow the proper understanding of underlying genotypes. To this end, it is also required that proper genome annotations are available for the genome assemblies for livestock.

7.5 Genomic characterization of European cattle: combining information from SNP arrays and whole genome sequences

Conservation management of a population requires a thorough understanding of the pattern of admixture, demography and genetic diversity. Moreover, scanning the genomic sequences in a population may unravel the genomic basis of adaptation. Native local cattle breeds have inhabited their respective environments for many centuries. However, limited availability of literature related to breed history for some populations may act as a hurdle in their conservation management. The distribution of genomic variations in a population provides reliable information about breed history. However, a question may arise regarding the type of genomic variations that should be used for genetic characterization. Although genotyping microsatellite markers in animals is more economical than SNP arrays, it provides only partial information concerning demographic history assessed through LD and ROH (Herrero-Medrano, 2013). Inferences based on SNP arrays and WGS indicated similar demographic history for European cattle populations. These results show that despite the small ascertainment bias, the currently available BovineHD genotyping arrays are very useful in deriving statistics related to genetic diversity and demography.

The results obtained from the studies described in this thesis demonstrate demographic history and admixture as two of the most important forces driving the distribution of genomic variations in cattle populations. Moreover, I also demonstrated that substantial genetic diversity exists among European cattle population which can be attributed to the founder effects involving migration of Neolithic farmers as well as gene-flow from non-European taurine cattle populations. For instance, BAI cattle breeds displayed high heterozygosity as well as an abundance of common and unique SV. These results partly can be attributed to the fact that among all the cattle breeds studied in this thesis, geographically, BAI cattle breeds are the closest to the centre of domestication.

Most of the native cattle breeds are still reared in small farms using conventional management. However, differences in breeding strategy between farms can lead to heterogeneous population structure. For instance, in chapter 4, sub-structure was identified in the two Swedish cattle breeds-Fjällnåra and Ringamålako. Moreover, It was also demonstrated that cross-breeding between local cattle breeds is also a prominent factor contributing to genetic diversity. Conversely, low genetic diversity due to genetic isolation in Iberian cattle breeds—Mirandesa and Cachena—and Swedish cattle breeds—Väneko—requires conservation efforts. I propose cross-breeding with individuals from phenotypically similar breeds might be a sustainable approach to conserve the breeds at risk, for instance, Ringamålako in case of Väneko. This might enhance the genetic diversity in such genetically isolated breeds. In fact, such conservation steps have already been carried out in the Maltese cattle breed, where Chianina bull has been used to increase the genetic diversity in this breed.

In this thesis, using SNP array as well as a whole genome sequencing approach, many common as well as novel SVs were identified. These results could indicate that native cattle breeds harbor unique genomic variants which might play an important role in adaptation. Moreover, in chapter 6, novel SVs have been exclusively identified in African taurine and Indian zebu. These results could indicate that SVs plays a vital role in population differentiation. However, determining the break-points of SV events was a major challenge in the studies performed in this thesis. Perhaps, in the future, the availability of sequence data produced by long read sequencing approaches may help resolve this issue.

7.6 Concluding remarks

This thesis provided detailed insights into how demographic changes and admixture patterns have contributed to genomic variation among European cattle breeds. The results in this thesis suggest a contribution of non-European taurine and ancestral wild aurochs populations, which warrants further investigation concerning adaptive introgression. Moreover, the results related to genetic diversity and population structure are valuable for conservation management of native cattle breeds. In this thesis, I also identified novel and lineage-specific structural variations which can be targeted by future association studies.

References

- Achilli, A. et al. (2008). Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Current Biology*, 18, R157–R158.
- Albrechtsen, A., Nielsen, F. C., & Nielsen, R. (2010). Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. *Molecular Biology and Evolution*, 27, 2534–2547.
- Aldred, P. M. R., Hollox, E. J., & Armour, J. A. L. (2005). Copy number polymorphism and expression level variation of the human α -defensin genes DEFA1 and DEFA3. *Human Molecular Genetics*, 14, 2045–2052.
- Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19, 1655–1664.
- Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12, 363–376.
- Anderung, C., Bouwman, A., Persson, P., Carretero, J. M., Ortega, A. I., Elburg, R., Smith, C., Arsuaga, J. L., Ellegren, H., & Gotherstrom, A. (2005). Prehistoric contacts over the Straits of Gibraltar indicated by genetic analysis of Iberian Bronze Age cattle. *Proceedings of the National Academy of Sciences*, 102, 8431–8435.
- Bahbahani, H. et al. (2017). Signatures of Selection for Environmental Adaptation and Zebu \times Taurine Hybrid Fitness in East African Shorthorn Zebu. *Frontiers in Genetics*, 8, 68.
- Balme, D. B. (1965). *History of animals translated from Historia animalium by Aristotle*. London : Heinemann ; Cambridge, Mass. : Harvard University Press, 1965-1991.
- Bank, R., Hettema, E., Muijs, M., Pals, G., Arwert, F., Boomsma, D., & Pronk, J. (1992). Variation in gene copy number and polymorphism of the human salivary amylase isoenzyme system in Caucasians. *Human Genetics*, 89, 213–222.
- Barbato, M., Hailer, F., Orozco-terWengel, P., Kijas, J., Mereu, P., Cabras, P., Mazza, R., Pirastru, M., & Bruford, M. W. (2017). Genomic signatures of adaptive introgression from European mouflon into domestic sheep. *Scientific Reports*, 7, 7623.
- Barbato, M., Orozco-terWengel, P., Tapio, M., & Bruford, M. W. (2015). SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Frontiers in Genetics*, 6, 109.
- Barendse, W., Harrison, B. E., Hawken, R. J., Ferguson, D. M., Thompson, J. M., Thomas, M. B., & Bunch, R. J. (2007). Epistasis Between Calpain 1 and Its Inhibitor Calpastatin

- Within Breeds of Cattle. *Genetics*, 176, 2601–2610.
- Bartosiewicz, L. (2011). The Hungarian Grey cattle: a traditional European breed. *Animal Genetic Resources Information*, 21, 49–60.
- Baum, B. R. (1989). PHYLIP: Phylogeny Inference Package. Version 3.2. *The Quarterly Review of Biology*, 64, 539–541.
- Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P., & Ramachandran, S. (2016). pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*, 32, 2817–2823.
- Beja-Pereira, A., Alexandrino, P., Bessa, I., Carretero, Y., Dunner, S., Ferrand, N., Jordana, J., Laloe, D., Moazami-Goudarzi, K., Sanchez, A., & Cañon, J. (2003). Genetic characterization of Southwestern European bovine breeds: A historical and biogeographical reassessment with a set of 16 microsatellites. *Journal of Heredity*, 94, 243–250.
- Beja-Pereira, A. et al. (2006). The origin of European cattle: evidence from modern and ancient DNA. *Proceedings of the National Academy of Sciences*, 103, 8113–8118.
- Ben Sassi, N., Gonzalez-Recio, O., de Paz-del Río, R., Rodríguez-Ramilo, S. T., & Fernández, A. I. (2016). Associated effects of copy number variants on economically important traits in Spanish Holstein dairy cattle. *Journal of Dairy Science*, 99, 6371–6380.
- Bhoumik, A., & Ronai, Z. (2008). ATF2: A transcription factor that elicits oncogenic or tumor suppressor activities. *Cell Cycle*, 7, 2341–2345.
- Bickhart, D. M. et al. (2012). Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Research*, 22, 778–790.
- Bickhart, D. M., & Liu, G. E. (2014). The challenges and importance of structural variation detection in livestock. *Frontiers in Genetics*, 5, 37.
- Bickhart, D. M. et al. (2016). Diversity and population-genetic properties of copy number variations and multicopy genes in cattle. *DNA Research*, 23, 253–262.
- Bodo, I., Gera, I., & Koppany, G. (2004). *The hungarian grey cattle breed book*. Budapest: ASSOCIATION OF THE HUNGARIAN GREY CATTLE BREEDERS.
- Bokonyi, S., Halapy, L., & Tringham, R. (1974). *History of domestic mammals in central and eastern Europe*. Akade'miai Kiado'.
- Bollongino, R., Burger, J., Powell, A., Mashkour, M., Vigne, J. D., & Thomas, M. G. (2012). Modern Taurine Cattle Descended from Small Number of Near-Eastern Founders. *Molecular Biology and Evolution*, 29, 2101–2104.
- Bollongino, R., Elsner, J., Vigne, J. D., & Burger, J. (2008). Y-SNPs do not indicate hybridisation between European aurochs and domestic cattle. *PLoS ONE*, 3, e3418.
- Bolormaa, S., Pryce, J. E., Reverter, A., Zhang, Y., Barendse, W., Kemper, K., Tier, B., Savin, K., Hayes, B. J., & Goddard, M. E. (2014). A Multi-Trait, Meta-analysis for Detecting Pleiotropic Polymorphisms for Stature, Fatness and Reproduction in Beef Cattle. *PLoS Genetics*, 10, e1004198.

- Bonfiglio, S., Achilli, A., Olivieri, A., Negrini, R., Colli, L., Liotta, L., Ajmone-Marsan, P., Torroni, A., & Ferretti, L. (2010). The Enigmatic Origin of Bovine mtDNA Haplogroup R: Sporadic Interbreeding or an Independent Event of *Bos primigenius* Domestication in Italy? *PLoS ONE*, 5, e15760.
- Bosse, M., Megens, H.-J., Frantz, L. A. F., Madsen, O., Larson, G., Paudel, Y., Duijvesteijn, N., Harlizius, B., Hagemeijer, Y., Crooijmans, R. P. M. A., & Groenen, M. A. M. (2014). Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nature Communications*, 5, 4392.
- Bosse, M., Megens, H.-J., Madsen, O., Paudel, Y., Frantz, L. A. F., Schook, L. B., Crooijmans, R. P. M. A., & Groenen, M. A. M. (2012). Regions of Homozygosity in the Porcine Genome: Consequence of Demography and the Recombination Landscape. *PLoS Genetics*, 8, e1003100.
- Boussaha, M. et al. (2015). Genome-Wide Study of Structural Variants in Bovine Holstein, Montbéliarde and Normande Dairy Breeds. *PLOS ONE*, 10, e0135931.
- Bradley, D. G., MacHugh, D. E., Cunningham, P., & Loftus, R. T. (1996). Mitochondrial diversity and the origins of African and European cattle. *Proceedings of the National Academy of Sciences*, 93, 5131–5135.
- Brenig, B., Beck, J., Floren, C., Bornemann-Kolatzki, K., Wiedemann, I., Hennecke, S., Swalve, H., & Schütz, E. (2013). Molecular genetics of coat colour variations in White Galloway and White Park cattle. *Animal Genetics*, 44, 450–453.
- Bro-Jørgensen, M. H., Carøe, C., Vieira, F. G., Nestor, S., Hallström, A., Gregersen, K. M., Etting, V., Gilbert, M. T. P., & Sinding, M.-H. S. (2018). Ancient DNA analysis of Scandinavian medieval drinking horns and the horn of the last aurochs bull. *Journal of Archaeological Science*, 99, 47–54.
- Browett, S., McHugo, G., Richardson, I. W., Magee, D. A., Park, S. D. E., Fahey, A. G., Kearney, J. F., Correia, C. N., Randhawa, I. A. S., & MacHugh, D. E. (2018). Genomic Characterisation of the Indigenous Irish Kerry Cattle Breed. *Frontiers in Genetics*, 9, 51.
- Browning, S. R., & 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.
- Broxham, E. T., Kugler, W., & Medugorac, I. (2015). A case study on strains of Busa cattle structured into a metapopulation to show the potential for use of single-nucleotide polymorphism genotyping in the management of small, cross-border populations of livestock breeds and varieties. *Frontiers in Genetics*, 6, 73.
- Busing, F. M. T. A., Meijer, E., & Leeden, R. V. D. (1999). Delete-m Jackknife for Unequal m. *Statistics and Computing*, 9, 3–8.
- Campbell, B. M. S. (2009). The Great Famine: Northern Europe in the Early Fourteenth Century. By William Chester Jordan. Princeton: Princeton University Press, 1996. Pp. 318. \$29.95. *The Journal of Economic History*, 57, 725–726.
- Casas, E., White, S. N., Wheeler, T. L., Shackelford, S. D., Koohmaraie, M., Riley, D. G., Chase, C. C., Johnson, D. D., & Smith, T. P. L. (2006). Effects of calpastatin and μ -calpain

- markers in beef cattle on tenderness traits. *Journal of Animal Science*, 84, 520–525.
- Cañas-Álvarez, J. J., González-Rodríguez, A., Munilla, S., Varona, L., Díaz, C., Baro, J. A., Altarriba, J., Molina, A., & Piedrafitra, J. (2015). Genetic diversity and divergence among Spanish beef cattle breeds assessed by a bovine high-density SNP chip. *Journal of Animal Science*, 93, 5164–5174.
- Cañón, J. et al. (2001). Genetic diversity measures of local European beef cattle breeds for conservation purposes. *Genetics Selection Evolution*, 33, 311.
- Chen, C., Qiao, R., Wei, R., Guo, Y., Ai, H., Ma, J., Ren, J., & Huang, L. (2012). A comprehensive survey of copy number variation in 18 diverse pig populations and identification of candidate copy number variable genes associated with complex traits. *BMC Genomics*, 13, 733.
- Chen, S. et al. (2009). Zebu Cattle Are an Exclusive Legacy of the South Asia Neolithic. *Molecular Biology and Evolution*, 27, 1–6.
- Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., Marth, G. T., Quinlan, A. R., & Hall, I. M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods*, 12, 966–968.
- Chomczynski, P., & Sacchi, N. (2006). The single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction: twenty-something years on. *Nature Protocols*, 1, 581–585.
- Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., Bassett, A. S., Seller, A., Holmes, C. C., & Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research*, 35, 2013–2025.
- Conolly, J., Manning, K., Colledge, S., Dobney, K., & Shennan, S. (2012). Species distribution modelling of ancient cattle from early Neolithic sites in SW Asia and Europe. *The Holocene*, 22, 997–1010.
- Cymbron, T., Freeman, A. R., Isabel Malheiro, M., Vigne, J. D., & Bradley, D. G. (2005). Microsatellite diversity suggests different histories for Mediterranean and Northern European cattle populations. *Proceedings of the Royal Society B: Biological Sciences*, 272, 1837–1843.
- Cymbron, T., Loftus, R. T., Malheiro, M. I., & Bradley, D. G. (1999). Mitochondrial sequence variation suggests an African influence in Portuguese cattle. *Proceedings of the Royal Society B: Biological Sciences*, 266, 597–603.
- 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., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27, 2156–2158.
- Daniel, H., Lionel, G., Hervé, M., Peters, J., & Segui, M. S. (2005). Identifying early domestic cattle from Pre-Pottery Neolithic sites on the Middle Euphrates using sexual dimorphism.

- book section 9. (pp. 86–96). Oxbow Books.
- Decker, J. E. et al. (2014). Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. *PLoS Genetics*, *10*, e1004254.
- 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.
- Di Lorenzo, P. et al. (2018). Mitochondrial DNA variants of Podolian cattle breeds testify for a dual maternal origin. *PLOS ONE*, *13*, e0192567.
- Di Trana, A., Sepe, L., Di Gregorio, P., Di Napoli, M. A., Giorgio, D., Caputo, A. R., & Claps, S. (2015). The Role of Local Sheep and Goat Breeds and Their Products as a Tool for Sustainability and Safeguard of the Mediterranean Environment. In *The Sustainability of Agro-Food and Natural Resource Systems in the Mediterranean Basin* (pp. 77–112). Springer International Publishing.
- Doran, A. G., Berry, D. P., & Creevey, C. J. (2014). Whole genome association study identifies regions of the bovine genome and biological pathways involved in carcass trait performance in Holstein-Friesian cattle. *BMC Genomics*, *15*, 837.
- Dorian, J. G., & Ruvinsky, A. (2014). *The genetics of cattle*. Wallingford, Oxfordshire: CAB International.
- Dumas, L., Kim, Y. H., Karimpour-Fard, A., Cox, M., Hopkins, J., Pollack, J. R., & Sikela, J. M. (2007). Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Research*, *17*, 1266–1277.
- Durkin, K. et al. (2012). Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature*, *482*, 81–84.
- D’Andrea, M., Pariset, L., Matassino, D., Valentini, A., Lenstra, J. A., Maiorano, G., & Pilla, F. (2011). Genetic characterization and structure of the Italian Podolian cattle breed and its relationship with some major European breeds. *Italian Journal of Animal Science*, *10*, e54.
- Edwards, C. J. et al. (2007). Mitochondrial DNA analysis shows a Near Eastern Neolithic origin for domestic cattle and no indication of domestication of European aurochs. *Proceedings of the Royal Society B: Biological Sciences*, *274*, 1377–1385.
- Edwards, C. J., Ginja, C., Kantanen, J., Pérez-Pardal, L., Tresset, A., Stock, F., Gama, L. T., Penedo, M. C. T., Bradley, D. G., Lenstra, J. A., & Nijman, I. J. (2011). Dual Origins of Dairy Cattle Farming – Evidence from a Comprehensive Survey of European Y-Chromosomal Variation. *PLoS ONE*, *6*, e15922.
- Edwards, C. J., MacHugh, D. E., Dobney, K. M., Martin, L., Russell, N., Horwitz, L. K., McIntosh, S. K., MacDonald, K. C., Helmer, D., Tresset, A., Vigne, J.-D., & Bradley, D. G. (2004). Ancient DNA analysis of 101 cattle remains: limits and prospects. *Journal of Archaeological Science*, *31*, 695–710.
- Edwards, C. J. et al. (2010). A Complete Mitochondrial Genome Sequence from a Mesolithic Wild Aurochs (*Bos primigenius*). *PLoS ONE*, *5*, e9255.

- Epstein, H. (1971). *The origin of the domestic animals of Africa*. Africana Publishing Corporation.
- Fadista, J., Thomsen, B., Holm, L.-E., & Bendixen, C. (2010). Copy number variation in the bovine genome. *BMC Genomics*, *11*, 284.
- Fao (2015). *The second report on the state of the world's FAO commission on genetic resources for food and agriculture assessments • 2015*.
- Faust, G. G., & Hall, I. M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*, *30*, 2503–2505.
- Felius, M. (1995). *Cattle breeds : an encyclopedia*. Doetinchem, Netherlands: Misset.
- Felius, M., Beerling, M.-L., Buchanan, D., Theunissen, B., Koolmees, P., & Lenstra, J. (2014). On the History of Cattle Genetic Resources. *Diversity*, *6*, 705–750.
- Ferdinando, C., & Donato, M. (2001). Bovino grigio allevato in italia: origine. Nota 1: il bovino macrocero. *Taurus speciale*, *13*, 89–99.
- Ferenčaković, M., Hamzić, E., Gredler, B., Solberg, T. R., Klemetsdal, G., Curik, I., & Sölkner, J. (2012). Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations. *Journal of Animal Breeding and Genetics*, *130*, 286–293.
- Frantz, A. C., Zachos, F. E., Kirschning, J., Cellina, S., Bertouille, S., Mamuris, Z., Koutsogiannouli, E. A., & Burke, T. (2013). Genetic evidence for introgression between domestic pigs and wild boars (*Sus scrofa*) in Belgium and Luxembourg: a comparative approach with multiple marker systems. *Biological Journal of the Linnean Society*, *110*, 104–115.
- Frantz, L. A. F., Schraiber, J. G., Madsen, O., Megens, H.-J., Cagan, A., Bosse, M., Paudel, Y., Crooijmans, R. P. M. A., Larson, G., & Groenen, M. A. M. (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.
- Freeman, J. L. (2006). Copy number variation: New insights in genome diversity. *Genome Research*, *16*, 949–961.
- Gao, Y., Jiang, J., Yang, S., Hou, Y., Liu, G. E., Zhang, S., Zhang, Q., & Sun, D. (2017). CNV discovery for milk composition traits in dairy cattle using whole genome resequencing. *BMC Genomics*, *18*, 265.
- Gautier, M., Faraut, T., Moazami-Goudarzi, K., Navratil, V., Foglio, M., Grohs, C., Boland, A., Garnier, J. G., Boichard, D., Lathrop, G. M., Gut, I. G., & Eggen, A. (2007). Genetic and haplotypic structure in 14 European and African cattle breeds. *Genetics*, *177*, 1059–1070.
- Gautier, M., Laloë, D., & Moazami-Goudarzi, K. (2010). Insights into the Genetic History of French Cattle from Dense SNP Data on 47 Worldwide Breeds. *PLoS ONE*, *5*, e13038.
- Gautier, M., Moazami-Goudarzi, K., Leveziel, H., Parinello, H., Grohs, C., Rialle, S., Kowalczyk, R., & Flori, L. (2016). Deciphering the Wisent Demographic and Adaptive Histories from Individual Whole-Genome Sequences. *Molecular Biology Evolution*, *33*, 2801–2814.
- Gazave, E. et al. (2011). Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Research*, *21*, 1626–1639.

- Gifford-Gonzalez, D., & Hanotte, O. (2011). Domesticating Animals in Africa: Implications of Genetic and Archaeological Findings. *Journal of World Prehistory*, 24, 1–23.
- Ginja, C. et al. (2013). Analysis of conservation priorities of Iberoamerican cattle based on autosomal microsatellite markers. *Genetics Selection Evolution*, 45, 35.
- Ginja, C., Penedo, M. C. T., Melucci, L., Quiroz, J., Martínez López, O. R., Revidatti, M. A., Martínez-Martínez, A., Delgado, J. V., & Gama, L. T. (2010a). Origins and genetic diversity of New World Creole cattle: inferences from mitochondrial and Y chromosome polymorphisms. *Animal Genetics*, 41, 128–141.
- Ginja, C., Telo Da Gama, L., & Penedo, M. C. T. (2010b). Analysis of STR Markers Reveals High Genetic Structure in Portuguese Native Cattle. *Journal of Heredity*, 101, 201–210.
- Gonzalez, E. (2005). The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility. *Science*, 307, 1434–1440.
- Goudet, J. (2005). hierfstat, a package for r to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, 5, 184–186.
- Green, R. E. et al. (2010). A Draft Sequence of the Neandertal Genome. *Science*, 328, 710–722.
- Grigson, C. (1991). An African origin for African cattle? some archaeological evidence. *The African Archaeological Review*, 9, 119–144.
- Götherström, A., Anderung, C., Hellborg, L., Elburg, R., Smith, C., Bradley, D. G., & Ellegren, H. (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.
- Haak, W. et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522, 207–211.
- Hall, S., & Clutton-Brock, J. (1988). *Two Hundred Years of British Farm Livestock*. British Museum, London, UK.
- Han, J. L., Yang, M., Yue, Y. J., Guo, T. T., Liu, J. B., Niu, C. E., & Yang, B. H. (2015). Analysis of agouti signaling protein (*ASIP*) gene polymorphisms and association with coat color in Tibetan sheep (*Ovis aries*). *Genetics and Molecular Research*, 14, 1200–1209.
- Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., & McCarroll, S. A. (2015). Large multiallelic copy number variations in humans. *Nature Genetics*, 47, 296–303.
- Hartl, G. B., Göltenboth, R., Grilltsch, M., & Willing, R. (1988). On the biochemical systematics of the bovini. *Biochemical Systematics and Ecology*, 16, 575–579.
- Hassanin, A., & Ropiquet, A. (2004). Molecular phylogeny of the tribe Bovini (Bovidae, Bovinae) and the taxonomic status of the Kouprey, *Bos sauveli* Urbain 1937. *Molecular Phylogenetics and Evolution*, 33, 896–907.
- Haubold, B., Pfaffelhuber, P., & Lynch, M. (2010). mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Molecular Ecology*, 19, 277–284.

- Hayes, B. J. (2003). Novel Multilocus Measure of Linkage Disequilibrium to Estimate Past Effective Population Size. *Genome Research*, 13, 635–643.
- Hedrick, P. W. (2013). Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, 22, 4606–4618.
- Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A Genetic Atlas of Human Admixture History. *Science*, 343, 747–751.
- Herrero-Medrano, J., Megens, H.-J., Groenen, M. A. M., Bosse, M., Pérez-Enciso, M., & Crooijmans, R. P. M. A. (2014). Whole-genome sequence analysis reveals differences in population management and selection of European low-input pig breeds. *BMC Genomics*, 15, 601.
- Herrero-Medrano, J. M., Megens, H. J., Crooijmans, R. P., Abellana, J. M., & Ramis, G. (2012). Farm-by-farm analysis of microsatellite, mtDNA and SNP genotype data reveals inbreeding and crossbreeding as threats to the survival of a native Spanish pig breed. *Animal Genetics*, 44, 259–266.
- Hiemstra, S. J., Haas, Y. d., Mäkit-Tanila, A., & Gandini, G. (2010). *Local cattle breeds in Europe : development of policies and strategies for self-sustaining breeds*. Wageningen: Wageningen Academic Publishers.
- Hofmanová, Z. et al. (2016). Early farmers from across Europe directly descended from Neolithic Aegeans. *Proceedings of the National Academy of Sciences*, 113, 6886–6891.
- Hou, Y. et al. (2012). Fine mapping of copy number variations on two cattle genome assemblies using high density SNP array. *BMC Genomics*, 13, 376.
- Hou, Y., Liu, G. E., Bickhart, D. M., Cardone, M. F., Wang, K., Kim, E.-s., Matukumalli, L. K., Ventura, M., Song, J., VanRaden, P. M., Sonstegard, T. S., & Van Tassell, C. P. (2011). Genomic characteristics of cattle copy number variations. *BMC Genomics*, 12, 127.
- Hu, Q., Ma, T., Wang, K., Xu, T., Liu, J., & Qiu, Q. (2012). The Yak genome database: an integrative database for studying yak biology and high-altitude adaption. *BMC Genomics*, 13, 600.
- Huddleston, J. et al. (2016). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*, 27, 677–685.
- Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, 14, 68–73.
- Jakobsson, M. et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451, 998–1003.
- Jiang, L., Jiang, J., Wang, J., Ding, X., Liu, J., & Zhang, Q. (2012). Genome-Wide Identification of Copy Number Variations in Chinese Holstein. *PLoS ONE*, 7, e48732.
- Jiang, L., Jiang, J., Yang, J., Liu, X., Wang, J., Wang, H., Ding, X., Liu, J., & Zhang, Q. (2013). Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. *BMC Genomics*, 14, 131.
- Jones, E. R. et al. (2015). Upper Palaeolithic genomes reveal deep roots of modern Eurasians.

- Nature Communications*, 6.
- Jorge, W. (1974). Chromosome Study of Some Breeds of Cattle. *Caryologia*, 27, 325–329.
- Joshi, N. A., & N, F. J. (). Sickel: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33), .
- Kadri, N. K. et al. (2016). Coding and noncoding variants in HFM1, MLH3, MSH4, MSH5, RNF212, and RNF212B affect recombination rate in cattle. *Genome Research*, 26, 1323–1332.
- Kantanen, J. (2000). Genetic diversity and population structure of 20 north European cattle breeds. *Journal of Heredity*, 91, 446–457.
- Kantanen, J., Edwards, C. J., Bradley, D. G., Viinalass, H., Thessler, S., Ivanova, Z., Kiselyova, T., Činkulov, M., Popov, R., Stojanović, S., Ammosov, I., & Vilkki, J. (2009). Maternal and paternal genealogy of Eurasian taurine cattle (*Bos taurus*). *Heredity*, 103, 404–415.
- Khaja, R., Macdonald, J., Zhang, J., & W Scherer, S. (2006). Methods for identifying and mapping recent segmental and gene duplications in eukaryotic genomes. In M. Bina (Ed.), *Gene mapping, discovery, and expression: methods and protocols* (pp. 9–20). volume 338.
- Kieffer, N. M., & Cartwright, T. C. (1968). Sex Chromosome Polymorphism in Domestic Cattle. *Journal of Heredity*, 59, 35–36.
- Kijas, J. W., Barendse, W., Barris, W., Harrison, B., McCulloch, R., McWilliam, S., & Whan, V. (2011). Analysis of copy number variants in the cattle genome. *Gene*, 482, 73–77.
- Kim, E.-S., Cole, J. B., Huson, H., Wiggans, G. R., Van Tassell, C. P., Crooker, B. A., Liu, G., Da, Y., & Sonstegard, T. S. (2013). Effect of Artificial Selection on Runs of Homozygosity in U.S. Holstein Cattle. *PLoS ONE*, 8, e80813.
- Kim, J. et al. (2017). The genome landscape of indigenous African cattle. *Genome Biology*, 18.
- Kirin, M., McQuillan, R., Franklin, C. S., Campbell, H., McKeigue, P. M., & Wilson, J. F. (2010). Genomic Runs of Homozygosity Record Population History and Consanguinity. *PLoS ONE*, 5, e13996.
- Korbel, J. O. et al. (2007). Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science*, 318, 420.
- Kron, G. (2002). Archaeozoology and the Productivity of Roman Livestock Farming. In *MBAH2* (pp. 53–73.). volume 21.
- L. Janecek, L., Honeycutt, R., M. Adkins, R., & K. Davis, S. (1996). Mitochondrial Gene Sequences and the Molecular Systematics of the Artiodactyl Subfamily Bovinae. *Molecular Phylogenetics and Evolution*, 6, 107–119.
- de Lagran, I. (2014). Recent Data and Approaches on the Neolithization of the Iberian Peninsula. *European Journal of Archaeology*, 18, 429–453.
- Lancioni, H. et al. (2016). Survey of uniparental genetic markers in the Maltese cattle breed reveals a significant founder effect but does not indicate local domestication. *Animal Genetics*, 47, 267–269.
- Lari, M. et al. (2011). The Complete Mitochondrial Genome of an 11,450-year-old Aurochsen

- (*Bos primigenius*) from Central Italy. *BMC Evolutionary Biology*, 11, 32.
- Lawson, D. J., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, 9, 3258.
- Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of Population Structure using Dense Haplotype Data. *PLoS Genetics*, 8, e1002453.
- Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, 15, R84.
- Lazaridis, I. et al. (2016). Genomic insights into the origin of farming in the ancient Near East. *Nature*, 536, 419–424.
- Legault, M.-A., Girard, S., Lemieux Perreault, L.-P., Rouleau, G. A., & Dubé, M.-P. (2015). Comparison of Sequencing Based CNV Discovery Methods Using Monozygotic Twin Quartets. *PLOS ONE*, 10, e0122287.
- Leslie, S. et al. (2015). The fine-scale genetic structure of the British population. *Nature*, 519, 309–314.
- Letaief, R. et al. (2017). Identification of copy number variation in French dairy and beef breeds using next-generation sequencing. *Genetics Selection Evolution*, 49, 77.
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997*, 1303.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079.
- Li, M. H., & Kantanen, J. (2010). Genetic structure of Eurasian cattle (*Bos taurus*) based on microsatellites: clarification for their breed classification. *Animal Genetics*, 41, 150–158.
- Li, N., & Stephens, M. (2003). Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, 165, 2213.
- Lien, S., Kantanen, J., Olsaker, I., Holm, L. E., Eythorsdottir, E., Sandberg, K., Dalsgard, B., & Adalsteinsson, S. (1999). Comparison of milk protein allele frequencies in Nordic cattle breeds. *Animal Genetics*, 30, 85–91.
- Lin, Y.-L., Pavlidis, P., Karakoc, E., Ajay, J., & Gokcumen, O. (2015). The Evolution and Functional Impact of Human Deletion Variants Shared with Archaic Hominin Genomes. *Molecular Biology and Evolution*, 32, 1008–1019.
- Liu, G. E., Brown, T., Hebert, D. A., Cardone, M. F., Hou, Y., Choudhary, R. K., Shaffer, J., Amazu, C., Connor, E. E., Ventura, M., & Gasbarre, L. C. (2010a). Initial analysis of copy number variations in cattle selected for resistance or susceptibility to intestinal nematodes. *Mammalian Genome*, 22, 111–121.
- Liu, G. E. et al. (2010b). Analysis of copy number variations among diverse cattle breeds. *Genome Research*, 20, 693–703.
- Liu, J., Zhang, L., Xu, L., Ren, H., Lu, J., Zhang, X., Zhang, S., Zhou, X., Wei, C., Zhao, F.,

- & Du, L. (2013). Analysis of copy number variations in the sheep genome using 50K SNP BeadChip array. *BMC Genomics*, *14*, 229.
- Loftus, R. T., MacHugh, D. E., Bradley, D. G., Sharp, P. M., & Cunningham, P. (1994). Evidence for two independent domestications of cattle. *Proceedings of the National Academy of Sciences*, *91*, 2757–2761.
- Lu, D., Miller, S., Sargolzaei, M., Kelly, M., Vander Voort, G., Caldwell, T., Wang, Z., Plastow, G., & Moore, S. (2013). Genome-wide association analyses for growth and feed efficiency traits in beef cattle1. *Journal of Animal Science*, *91*, 3612–3633.
- Lupski, J. R. (2007a). An evolution revolution provides further revelation. *BioEssays*, *29*, 1182–1184.
- Lupski, J. R. (2007b). Genomic rearrangements and sporadic disease. *Nature Genetics*, *39*, S43–S47.
- Ma, L., O’Connell, J. R., VanRaden, P. M., Shen, B., Padhi, A., Sun, C., Bickhart, D. M., Cole, J. B., Null, D. J., Liu, G. E., Da, Y., & Wiggans, G. R. (2015). Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis. *PLOS Genetics*, *11*, e1005387.
- MacEachern, S., McEwan, J., & Goddard, M. (2009). Phylogenetic reconstruction and the identification of ancient polymorphism in the Bovini tribe (Bovidae, Bovinae). *BMC Genomics*, *10*, 177.
- Martinez et al. (2012). Genetic Footprints of Iberian Cattle in America 500 Years after the Arrival of Columbus. *Plos One*, *7*, e49066.
- Martins, H., Oms, F. X., Pereira, L., Pike, A. W. G., Rowsell, K., & Zilhão, J. (2015). Radio-carbon Dating the Beginning of the Neolithic in Iberia: New Results, New Problems. *Journal of Mediterranean Archaeology*, *28*, 105–131.
- Martín-Burriel, I. et al. (2011). Genetic diversity, structure, and breed relationships in Iberian cattle. *Journal of Animal Science*, *89*(4), 893–906.
- Mason, I. L. (1984). *Evolution of domesticated animals*. Prentice Hall Press.
- Mastrangelo, S. et al. (2018). Conservation status and historical relatedness of Italian cattle breeds. *Genetics Selection Evolution*, *50*, 35.
- Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., O’Connell, J., Moore, S. S., Smith, T. P. L., Sonstegard, T. S., & Van Tassell, C. P. (2009). Development and Characterization of a High Density SNP Genotyping Assay for Cattle. *PLoS ONE*, *4*, e5350.
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*, 122.
- McTavish, E. J., Decker, J. E., Schnabel, R. D., Taylor, J. F., & Hillis, D. M. (2013). New World cattle show ancestry from multiple independent domestication events. *Proceedings of the National Academy of Sciences*, *110*, E1398–E1406.
- Medugorac, I. et al. (2017). Whole-genome analysis of introgressive hybridization and charac-

- terization of the bovine legacy of Mongolian yaks. *Nature Genetics*, 49, 470.
- Medugorac, I., Medugorac, A., Russ, I., Veit-Kensch, C. E., Taberlet, P., Luntz, B., Mix, H. M., & Forster, M. (2009). Genetic diversity of European cattle breeds highlights the conservation value of traditional unselected breeds with high effective population size. *Molecular Ecology*, 18, 3394–410.
- Medvedev, P., Stanciu, M., & Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, 6, S13–S20.
- Mesbah-Uddin, M., Guldbrandtsen, B., Iso-Touru, T., Vilkki, J., De Koning, D.-J., Boichard, D., Lund, M. S., & Sahana, G. (2017). Genome-wide mapping of large deletions and their population-genetic properties in dairy cattle. *DNA Research*, 25, 49–59.
- Mielczarek, M., Fraszczak, M., Nicolazzi, E., Williams, J. L., & Szyda, J. (2018). Landscape of copy number variations in *Bos taurus*: individual – and inter-breed variability. *BMC Genomics*, 19, 410.
- Moller, M. J., Chaudhary, R., Hellmén, E., Höyheim, B., Chowdhary, B., & Andersson, L. (1996). Pigs with the dominant white coat color phenotype carry a duplication of the KIT gene encoding the mast/stem cell growth factor receptor. *Mammalian Genome*, 7, 822–830.
- Mona, S., Catalano, G., Lari, M., Larson, G., Boscato, P., Casoli, A., Sineo, L., Di Patti, C., Pecchioli, E., Caramelli, D., & Bertorelle, G. (2010). Population dynamic of the extinct European aurochs: genetic evidence of a north-south differentiation pattern and no evidence of post-glacial expansion. *BMC Evolutionary Biology*, 10, 83.
- Monlong, J., Cossette, P., Meloche, C., Rouleau, G., Girard, S. L., & Bourque, G. (2018). Human copy number variants are enriched in regions of low mappability. *Nucleic Acids Research*, 46, 7236–7249.
- Montinaro, F., Busby, G. B. J., Pascali, V. L., Myers, S., Hellenthal, G., & Capelli, C. (2015). Unravelling the hidden ancestry of American admixed populations. *Nature Communications*, 6, 6596.
- Murgiano, L., Jagannathan, V., Calderoni, V., Joechler, M., Gentile, A., & Drögemüller, C. (2014). Looking the Cow in the Eye: Deletion in the NID1 Gene Is Associated with Recessive Inherited Cataract in Romagnola Cattle. *PLoS ONE*, 9, e110628.
- Negrini, R., Nicoloso, L., Crepaldi, P., Milanese, E., Colli, L., Chegdani, F., Pariset, L., Dunner, S., Leveziel, H., Williams, J. L., & Ajmone Marsan, P. (2009). Assessing SNP markers for assigning individuals to cattle populations. *Animal Genetics*, 40, 18–26.
- Norris, B. J., & Whan, V. A. (2008). A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep. *Genome Research*, 18, 1282–1293.
- Nozawa, M., & Nei, M. (2007). Evolutionary dynamics of olfactory receptor genes in *Drosophila* species. *Proceedings of the National Academy of Sciences*, 104, 7122–7127.
- Orlando, L., Gilbert, M. T. P., & Willerslev, E. (2015). Reconstructing ancient genomes and epigenomes. *Nature Reviews Genetics*, 16, 395–408.
- Orlando, L. et al. (2013). Recalibrating Equus evolution using the genome sequence of an early

- Middle Pleistocene horse. *Nature*, 499, 74–78.
- 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.
- Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient Admixture in Human History. *Genetics*, 192, 1065–1093.
- Patterson, N., Price, A. L., & Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS Genetics*, 2, e190.
- Paudel, Y., Madsen, O., Megens, H.-J., Frantz, L. A. F., Bosse, M., Bastiaansen, J. W. M., Crooijmans, R. P. M. A., & Groenen, M. A. M. (2013). Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics*, 14, 449.
- Pellecchia, M., Negrini, R., Colli, L., Patrini, M., Milanese, E., Achilli, A., Bertorelle, G., Cavalli-Sforza, L. L., Piazza, A., Torroni, A., & Ajmone-Marsan, P. (2007). The mystery of Etruscan origins: novel clues from *Bos taurus* mitochondrial DNA. *Proceedings of the Royal Society B: Biological Sciences*, 274, 1175–1179.
- Perry, G. H. (2008). The evolutionary significance of copy number variation in the human genome. *Cytogenetic and Genome Research*, 123, 283–287.
- Perry, G. H. et al. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, 39, 1256–1260.
- Pezer, Z., Harr, B., Teschke, M., Babiker, H., & Tautz, D. (2015). Divergence patterns of genic copy number variation in natural populations of the house mouse (*Mus musculus domesticus*) reveal three conserved genes with major population-specific expansions. *Genome Research*, 25, 1114–1124.
- Pickrell, J. K., & Pritchard, J. K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics*, 8, e1002967.
- Pielberg, G., Olsson, C., Syvänen, A. C., & Andersson, L. (2002). Unexpectedly high allelic diversity at the KIT locus causing dominant white color in the domestic pig. *Genetics*, 160, 305–311.
- Pieragostini, E., Scaloni, A., Rullo, R., & Di Luccia, A. (2000). Identical marker alleles in Podolic cattle (*Bos taurus*) and Indian zebu (*Bos indicus*). *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 127, 1–9.
- Pinto, D. et al. (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology*, 29, 512–520.
- Pirooznia, M., Goes, F. S., & Zandi, P. P. (2015). Whole-genome CNV analysis: advances in computational approaches. *Frontiers in Genetics*, 06, 138.
- Pitt, D., Sevane, N., Nicolazzi, E. L., MacHugh, D. E., Park, S. D. E., Colli, L., Martinez, R., Bruford, M. W., & Orozco-terWengel, P. (2019). Domestication of cattle: Two or three

- events? *Evolutionary Applications*, (pp. 123–136).
- Poulsen, N. A., Glantz, M., Rosengaard, A. K., Paulsson, M., & Larsen, L. B. (2017). Comparison of milk protein composition and rennet coagulation properties in native Swedish dairy cow breeds and high-yielding Swedish Red cows. *Journal of Dairy Science*, *100*, 8722–8734.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*, 945.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (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., Berry, D. P., McParland, S., & Bradley, D. G. (2012). Runs of homozygosity and population history in cattle. *BMC Genetics*, *13*, 70.
- Raven, L.-A., Cocks, B. G., & Hayes, B. J. (2014). Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. *BMC Genomics*, *15*, 62.
- Redon, R. et al. (2006). Global variation in copy number in the human genome. *Nature*, *444*, 444–454.
- Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature*, *461*, 489–494.
- Reyer, H., Hawken, R., Murani, E., Ponsuksili, S., & Wimmers, K. (2015). The genetics of feed conversion efficiency traits in a commercial broiler line. *Scientific Reports*, *5*, 16387.
- Ritz, L. R., Glowatzki-Mullis, M. L., MacHugh, D. E., & Gaillard, C. (2000). Phylogenetic analysis of the tribe Bovini using microsatellites. *Animal Genetics*, *31*, 178–185.
- Rupp, R., Hernandez, A., & Mallard, B. A. (2007). Association of Bovine Leukocyte Antigen (BoLA) DRB3.2 with Immune Response, Mastitis, and Production and Type Traits in Canadian Holsteins. *Journal of Dairy Science*, *90*, 1029–1038.
- Saether, N. H., Boe, K. E., & Vangen, O. (2006). Differences in grazing behaviour between a high and a moderate yielding Norwegian dairy cattle breed grazing semi-natural mountain grasslands. *Acta Agriculturae Scandinavica, Section A - Animal Science*, *56*, 91–98.
- Salmon Hillbertz, N. H. C. et al. (2007). Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. *Nature Genetics*, *39*, 1318–1320.
- Salque, M., Bogucki, P. I., Pyzel, J., Sobkowiak-Tabaka, I., Grygiel, R., Szmyt, M., & Evershed, R. P. (2013). Earliest evidence for cheese making in the sixth millennium BC in northern Europe. *Nature*, *493*, 522–525.
- Sargentini, C., Riccardo, B., Pablo, D.-R., Alessandro, G., Andrea, M., Paola, L., Cazzola, P., Simona, B., & Tiziana, C. (2007). Onset of puberty in Maremmana heifers. *Italian Journal of Animal Science*, *6*, 385–394.
- Sasaki, S., Watanabe, T., Nishimura, S., & Sugimoto, Y. (2016). Genome-wide identification of

- copy number variation using high-density single-nucleotide polymorphism array in Japanese Black cattle. *BMC Genetics*, 17, 26.
- Schaich, H., Rudner, M., & Konold, W. (2010). Short-term impact of river restoration and grazing on floodplain vegetation in Luxembourg. *Agriculture, Ecosystems Environment*, 139, 142–149.
- Scheu, A., Hartz, S., Schmölcke, U., Tresset, A., Burger, J., & Bollongino, R. (2008). Ancient DNA provides no evidence for independent domestication of cattle in Mesolithic Rosenhof, Northern Germany. *Journal of Archaeological Science*, 35, 1257–1264.
- Scheu, A., Powell, A., Bollongino, R., Vigne, J. D., Tresset, A., Cakirlar, C., Benecke, N., & Burger, J. (2015). The genetic prehistory of domesticated cattle from their origin to the spread across Europe. *BMC Genetics*, 16, 54.
- Schibler, J., Elsner, J., & Schlumbaum, A. (2014). Incorporation of aurochs into a cattle herd in Neolithic Europe: single event or breeding? *Scientific Reports*, 4, 5798.
- Schibler, L. (2000). Fine Mapping Suggests that the Goat Polled Intersex Syndrome and the Human Blepharophimosis Ptosis Epicanthus Syndrome Map to a 100-kb Homologous Region. *Genome Research*, 10, 311–318.
- Schliep, K. P. (2010). phangorn: phylogenetic analysis in R. *Bioinformatics*, 27, 592–593.
- Schuster-Bockler, B., Conrad, D., & Bateman, A. (2010). Dosage sensitivity shapes the evolution of copy-number varied regions. *PLoS One*, 5, e9474.
- Seabury, C. M. et al. (2017). Genome-wide association study for feed efficiency and growth traits in U.S. beef cattle. *BMC Genomics*, 18, 386.
- Sermyagin, A. A. et al. (2018). Whole-genome SNP analysis elucidates the genetic structure of Russian cattle and its relationship with Eurasian taurine breeds. *Genetics Selection Evolution*, 50, 37.
- Sharp, A. J., Cheng, Z., & Eichler, E. E. (2006). Structural Variation of the Human Genome. *Annual Review of Genomics and Human Genetics*, 7, 407–442.
- Sharp, A. J. et al. (2005). Segmental Duplications and Copy-Number Variation in the Human Genome. *The American Journal of Human Genetics*, 77, 78–88.
- Shi, T., Xu, Y., Yang, M., Huang, Y., Lan, X., Lei, C., Qi, X., Yang, X., & Chen, H. (2015). Copy number variations at LEPR gene locus associated with gene expression and phenotypic traits in Chinese cattle. *Animal Science Journal*, 87, 336–343.
- Silva, V. H. d., Regitano, L. C. d. A., Geistlinger, L., Pértille, F., Giachetto, P. F., Brassaloti, R. A., Morosini, N. S., Zimmer, R., & Coutinho, L. L. (2016). Genome-Wide Detection of CNVs and Their Association with Meat Tenderness in Nelore Cattle. *PLOS ONE*, 11, e0157711.
- Sindi, S. S., Önal, S., Peng, L. C., Wu, H.-T., & Raphael, B. J. (2012). An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biology*, 13, R22.
- Sjödén, P., & Jakobsson, M. (2012). Population Genetic Nature of Copy Number Variation. In

- Methods in Molecular Biology* (pp. 209–223). Springer New York.
- Somel, M. et al. (2016). Archaeogenomic analysis of ancient Anatolians : first genetic indication for Neolithic cultural diffusion in the Near East. *85th Annual Meeting of the American-Association-of-Physical-Anthropologists, APR 13-16, 2016, Atlanta, GA, 159*, 297–298.
- Song, Y., Endepols, S., Klemann, N., Richter, D., Matuschka, F.-R., Shih, C.-H., Nachman, M. W., & Kohn, M. H. (2011). Adaptive Introgression of Anticoagulant Rodent Poison Resistance by Hybridization between Old World Mice. *Current Biology*, *21*, 1296–1301.
- Soubrier, J. et al. (2016). Early cave art and ancient DNA record the origin of European bison. *Nature Communications*, *7*, 13158.
- Sousa, V. C., Beaumont, M. A., Fernandes, P., Coelho, M. M., & Chikhi, L. (2011). Population divergence with or without admixture: selecting models using an ABC approach. *Heredity*, *108*, 521–530.
- Spielmann, M., & Klopocki, E. (2013). CNVs of noncoding cis-regulatory elements in human disease. *Curr Opin Genet Dev*, *23*, 249–56.
- Stankiewicz, P., & Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends in Genetics*, *18*, 74–82.
- Stankiewicz, P., & Lupski, J. R. (2010). Structural Variation in the Human Genome and its Role in Disease. *Annual Review of Medicine*, *61*, 437–455.
- Stanley, P. (1995). *Robert Bakewell and the Longhorn breed of cattle*. Ipswich :: Farming Press.
- Stothard, P., Choi, J.-W., Basu, U., Sumner-Thomson, J. M., Meng, Y., Liao, X., & Moore, S. S. (2011). Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. *BMC Genomics*, *12*, 559.
- Sudmant, P. H. et al. (2015). Global diversity, population stratification, and selection of human copy-number variation. *Science*, *349*, aab3761.
- Svensson, E., & Gotherstrom, A. (2008). Temporal fluctuations of Y-chromosomal variation in *Bos taurus*. *Biology Letters*, *4*, 752–754.
- Swedish Board of, A. (2011a). *Lantrasbevarande. Biologisk mångfald för framtiden*. Report, . URL: http://www2.jordbruksverket.se/webdav/files/SJV/trycksaker/Pdf_jo/jo11-16.pdf.
- Swedish Board of, A. (2011b). *Lantraser Vill du vara med* . Report, . URL: https://www2.jordbruksverket.se/webdav/files/SJV/trycksaker/Pdf_ovrigt/ovr238.pdf.
- Tapio, I., VÄRv, S., Bennewitz, J., Maleviciute, J., Fimland, E., Grislis, Z., Meuwissen, T. H. E., Miceikiene, I., Olsaker, I., Viinalass, H., Vilkki, J., & Kantanen, J. (2006). Prioritization for Conservation of Northern European Cattle Breeds Based on Analysis of Microsatellite Data. *Conservation Biology*, *20*, 1768–1779.
- Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., & Visscher, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Research*, *17*, 520–526.

- Theunert, C., & Slatkin, M. (2017). Distinguishing Recent Admixture from Ancestral Population Structure. *Genome Biology and Evolution*, 9, 427–437.
- Thomas, R. (2005). Zooarchaeology, Improvement and the British Agricultural Revolution. *International Journal of Historical Archaeology*, 9, 71–88.
- Tizioto, P. C. et al. (2013). Genome scan for meat quality traits in Nelore beef cattle. *Physiological Genomics*, 45, 1012–1020.
- Tresset, A. (2003). French Connections II: of cows and men. In E. M. E. N. Ian Armit, & S. Derek (Eds.), *Neolithic Settlement in Ireland and Western Britain* book section 3. (pp. 18–30). Oxford, UK: Oxbow Books.
- Troy, C. S., MacHugh, D. E., Bailey, J. F., Magee, D. A., Loftus, R. T., Cunningham, P., Chamberlain, A. T., Sykes, B. C., & Bradley, D. G. (2001). Genetic evidence for Near-Eastern origins of European cattle. *Nature*, 410, 1088.
- Turner, D. J., Miretti, M., Rajan, D., Fiegler, H., Carter, N. P., Blayney, M. L., Beck, S., & Hurles, M. E. (2007). Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nature Genetics*, 40, 90–95.
- Upadhyay, M., da Silva, V. H., Megens, H.-J., Visker, M. H. P. W., Ajmone-Marsan, P., Bâlțeanu, V. A., Dunner, S., Garcia, J. F., Ginja, C., Kantanen, J., Groenen, M. A. M., & Crooijmans, R. P. M. A. (2017). Distribution and Functionality of Copy Number Variation across European Cattle Populations. *Frontiers in Genetics*, 8, 108.
- Upadhyay, M. R. et al. (2016). Genetic origin, admixture and population history of aurochs (*Bos primigenius*) and primitive European cattle. *Heredity*, 118, 169–176.
- Utsunomiya, Y., Bomba, L., Lucente, G., Colli, L., Negrini, R., Lenstra, J., Erhardt, G., Garcia, J., & Ajmone-Marsan, P. (2014). Revisiting AFLP fingerprinting for an unbiased assessment of genetic structure and differentiation of taurine and zebu cattle. *BMC Genetics*, 15, 47.
- Van Laere, A.-S. et al. (2003). A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature*, 425, 832–836.
- Vilà, C., Seddon, J., & Ellegren, H. (2005). Genes of domestic mammals augmented by backcrossing with wild ancestors. *Trends in Genetics*, 21, 214–218.
- Visser, M., Kayser, M., & Palstra, R. J. (2012). HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Research*, 22, 446–455.
- Viluma, A., Mikko, S., Hahn, D., Skow, L., Andersson, G., & Bergström, T. F. (2017). Genomic structure of the horse major histocompatibility complex class II region resolved using PacBio long-read sequencing technology. *Scientific Reports*, 7, 45518.
- van Vuure, C. (2005). *Retracing the aurochs history, morphology and ecology of an extinct wild ox*. Pensoft Pub.
- Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., Hakonarson, H., & Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, 17,

- 1665–1674.
- Wang, M. D., Dzama, K., Rees, D. J. G., & Muchadeyi, F. C. (2016). Tropically adapted cattle of Africa: Perspectives on potential role of copy number variations. *Animal Genetics*, 47, 154–164.
- Warburton, P. E., Hasson, D., Guillem, F., Lescale, C., Jin, X., & Abrusan, G. (2008). Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics*, 9, 533.
- Wendorf, F., Close, A., & Schild, R. (1989). Early domestic cattle and scientific methodology. In L. Krzyżaniak, & M. Kobusiewicz (Eds.), *Proceedings of the International Symposium organized by the Archaeological Commission of the Polish Academy of Sciences* (pp. 61–67). volume 2 of *Studies in African archaeology* (Poznań, Poland).
- Wright, E. (2013). *The history of the European aurochs (Bos primigenius) from the Middle Pleistocene to its extinction: an archaeological investigation of its evolution, morphological variability and response to human exploitation*. Doctoral.
- Xu, L., Cole, J. B., Bickhart, D. M., Hou, Y., Song, J., VanRaden, P. M., Sonstegard, T. S., Van Tassell, C. P., & Liu, G. E. (2014). Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. *BMC Genomics*, 15, 683.
- Xu, L., Hou, Y., Bickhart, D. M., Zhou, Y., Hay, E. H. a., Song, J., Sonstegard, T. S., Van Tassell, C. P., & Liu, G. E. (2016). Population-genetic properties of differentiated copy number variations in cattle. *Scientific Reports*, 6, 23161.
- Yoshida, T., Furuta, H., Kondo, Y., & Mukoyama, H. (2011). Association of BoLA-DRB3 alleles with mastitis resistance and susceptibility in Japanese Holstein cows. *Animal Science Journal*, 83, 359–366.
- Yurchenko, A., Yudin, N., Aitnazarov, R., Plyusnina, A., Brukhin, V., Soloshenko, V., Lhasaranov, B., Popov, R., Paronyan, I. A., Plemashov, K. V., & Larkin, D. M. (2017). Genome-wide genotyping uncovers genetic profiles and history of the Russian cattle breeds. *Heredity*, 120, 125–137.
- Zarrei, M., MacDonald, J. R., Merico, D., & Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics*, 16, 172–183.
- Zhan, B., Fadista, J., Thomsen, B., Hedegaard, J., Panitz, F., & Bendixen, C. (2011). Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. *BMC Genomics*, 12, 557.
- Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. *Annual Review of Genomics and Human Genetics*, 10, 451–481.
- Zhang, L., Jia, S., Plath, M., Huang, Y., Li, C., Lei, C., Zhao, X., & Chen, H. (2015). Impact of Parental *Bos taurus* and *Bos indicus* Origins on Copy Number Variation in Traditional Chinese Cattle Breeds. *Genome Biology and Evolution*, 7, 2352–2361.
- Zhang, L. et al. (2014a). Detection of copy number variations and their effects in Chinese bulls.

BMC Genomics, 15, 480.

- Zhang, Q., Ma, Y., Wang, X., Zhang, Y., & Zhao, X. (2014b). Identification of copy number variations in Qinchuan cattle using BovineHD Genotyping Beadchip array. *Molecular Genetics and Genomics*, 290, 319–327.
- Zhang, Z. D., Du, J., Lam, H., Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). Identification of genomic indels and structural variations using split reads. *BMC Genomics*, 12, 375.
- Zheng, X., Gogarten, S. M., Lawrence, M., Stilp, A., Conomos, M. P., Weir, B. S., Laurie, C., & Levine, D. (2017). SeqArray—a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics*, 33, 2251–2257.
- Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., Weir, B. S., & Korkman, N. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 28, 3326–3328.
- Zhou, B., Ho, S., Zhang, X., Pattni, R., Haraksingh, R., & Urban, A. (2017). Whole-genome sequencing analysis of copy number variation (CNV) using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *bioRxiv[PREPRINT]*, .
- Zhou, Y., Utsunomiya, Y. T., Xu, L., Hay, E. H. a., Bickhart, D. M., Sonstegard, T. S., Van Tassell, C. P., Garcia, J. F., & Liu, G. E. (2016). Comparative analyses across cattle genders and breeds reveal the pitfalls caused by false positive and lineage-differential copy number variations. *Scientific Reports*, 6, 29219.
- Zhu, C. et al. (2016). Genome-wide detection of CNVs in Chinese indigenous sheep with different types of tails using ovine high-density 600K SNP arrays CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Scientific Reports*, 6, 974–984.
- Zilhao, J. et al. (1993). The Spread of Agro-Pastoral Economies across Mediterranean Europe: A View from the Far West. *Journal of Mediterranean Archaeology*, 6, 5–63.
- Zimin, A. V. et al. (2009). A whole-genome assembly of the domestic cow, *Bos taurus* Identification of copy number variations and common deletion polymorphisms in cattle. *Genome Biology*, 10, R42.
- Zimin, A. V. et al. (2012). Mis-Assembled “Segmental Duplications” in Two Versions of the *Bos taurus* Genome Mitochondrial genomes of extinct aurochs survive in domestic cattle. *PLoS ONE*, 7, e42680.

Summary

Summary

A large diversity exists among European cattle breeds which can be attributed to population genomics forces such as migration, demography, and selection. These forces determine the distribution of variations such as single nucleotide polymorphisms (SNP) and structural variations (SV) in the genome. While the effect of these population genomics forces, as well as the distribution of genomic variations, have been studied extensively in commercial European cattle breeds, very few studies have focused on traditional and primitive cattle breeds of Europe. Therefore, this thesis aimed at providing a comprehensive overview of genomic admixture, demography, and variations in European cattle breeds using genome-wide SNP markers.

In chapter 2, the genome-wide SNP data indicated that many European cattle breeds display zebu and wild local aurochs ancestry in their genome. Therefore, the high divergence of Balkan and Italian cattle breeds can be attributed to zebu ancestry present in their genome. In this chapter, comprehensive overview of shared genomic variants was provided between wild local aurochs and different European cattle populations. These results suggested the possibility of several instances of intermingling between wild local aurochs and ancestors of modern European cattle. Additionally, runs of homozygous genotypes also indicated that several Iberian and Italian cattle breeds have undergone a recent reduction in effective population size.

In chapter 3, we collated the genotyping data of various African and zebu cattle breeds available in public database to characterize African taurine and zebu ancestry in more detail. Additionally, we also made use of individual whole genome sequencing information to perform a comparative evaluation of genetic diversity in European cattle. The results showed that Italian cattle display high genetic diversity which can be attributed to the diverse founder population. In line with the previous studies, the influence of African taurine ancestry in Iberian cattle was also reported in this chapter. Further, we show that like central Italian cattle breeds, Busha also displays complex non-European ancestry—African taurine and zebu—, probably indicating the common origin of this ancestry in various Balkan and Italian cattle populations.

In chapter 4, genetic relatedness, diversity and demographic history of native Swedish cattle breeds were studied using genome-wide SNP markers. The results indicated that these cattle breeds could be classified into two genetic clusters: Swedish mountain cattle breeds (including Bohus Polled) and horned cattle breeds from southern Sweden. Interestingly, we also identified sub-substructure within a Fjallanra population which corresponded to their farm of origin. Moreover, the results also indicated a relatively high genetic diversity in Swedish mountain cattle. Additionally, low genetic diversity was observed in Vaneko and Ringamalako which can be attributed to their genetic isolation. Comparative evaluation of genotyping data of various northern-western European cattle breeds indicated little contribution in gene pool of native Swedish cattle.

In chapter 5, identification and distribution of copy number variations (CNVs) were assessed using signal intensity from SNP genotyping array. The results indicated differences in CNV profile between different European cattle populations, indicating the effect of demography and selection. Moreover, enrichment analysis showed that CNVs are enriched in the genomic regions

related to olfactory processes and immunity. Additionally, structural variation involving *KIT* genes, which has been associated with coat-color sidedness in Belgian blue cattle in the previous study, was also identified in English longhorn samples, indicating its role in coat color diversity in the population.

In chapter 6, three different approaches—paired-end read, split reads and read-depth—were employed to identify SVs in individual whole genome sequences of cattle. Similar to chapter 5, differences in SVs profile was observed between African taurine, zebu and European taurine, indicating the effect of demography. However, these results in parts can also be attributed to reference genome which is assembled from European taurine individual. The analysis also identified lineage-specific SVs in different cattle populations, indicating their role in population differentiation. Additionally, CNVs were also identified from ancient aurochs whole genome sequences using the read-depth approach. The results identified a high sharing of duplication event between aurochs sample and modern European cattle, which can be attributed to recent evolutionary split between these two populations. Finally, the approaches identified many SV associated with traits related to meat quality, coat color and metabolism, which can be targeted by future association studies.

Finally, in chapter 7, I discuss the findings of all the previous chapters in relation to what is known so far with respect to genomic admixture, diversity, and demography of European cattle. I also discuss the strength and limitations of the approaches employed in the thesis. I also put forward hypotheses of a demographic scenario involving complex admixture pattern in Balkan and Italian cattle breeds. I also discuss the possibility that zebu ancestry could be playing an important role by providing increased fitness to some European cattle. I also discuss the need of additional ancient aurochs samples representing different time periods and geographical ranges to answers the questions related to cattle domestication in Europe. I conclude by highlighting the practical implication of the information related to genetic diversity and structure in conservation management of native cattle breeds.

Samenvatting

Samenvatting

Er bestaat een grote diversiteit tussen Europese runderrassen welke toegeschreven kan worden aan diverse factoren die van invloed zijn op het genoom, zoals migratie, demografie en selectie. Deze factoren bepalen de verdeling van de waargenomen variaties in een genoom zoals puntmutaties (“single nucleotide polymorphism”, SNP) en structurele variaties (SV). Terwijl het effect van deze factoren op het genoom alsmede de verdeling van de genoomvariatie uitvoerig zijn bestudeerd in commerciële Europese runderrassen, hebben slechts enkele studies zich gefocust op traditionele en primitieve runderrassen in Europa. Daarom is er in dit proefschrift getracht om met behulp van genoom-wijde SNP merkers een uitvoerig overzicht te verkrijgen van de demografie en genomische vermenging en variatie binnen Europese runderrassen.

In hoofdstuk 2, laat ik zien dat het genoom van veel Europese runderrassen, sporen vertoont van het Europese oeros en zebu. Ik veronderstel dat de hoge mate van divergentie van Balkan en Italiaanse runderrassen toegewezen kan worden aan het inkruisen van zebu in het verleden. In dit hoofdstuk, geef ik ook een uitgebreid overzicht van overeenkomstige genetische varianten tussen het wilde lokale oeros en diverse Europese runderpopulaties. Deze resultaten suggereren de mogelijkheid dat in meerdere gevallen een kruising tussen het wilde oeros en voorouders van de moderne Europese rundveerassen heeft moeten hebben plaatsgevonden. Verder toon ik ook aan dat verschillende Iberische en Italiaanse rundveerassen een recente reductie in effectieve populatiegrootte hebben ondergaan.

In hoofdstuk 3, combineer ik de genotypeerde data van diverse Afrikaanse en zebu runderrassen, waarvoor data voorhanden is in publieke databases en mijn data van de Europese runderrassen, en kom zo op een meer gedetailleerde karakterisatie van de mate van de genetische inbreng van zebu en Afrikaanse runderen in deze Europese rundveerassen. Daarnaast, heb ik gebruik gemaakt van complete genoomsequenties van individuele dieren voor een vergelijkende evaluatie van de genetische diversiteit van Europese runderen. Hiermee toon ik aan dat Italiaanse runderen de hoogste genetische diversiteit laten zien, wat toegeschreven kan worden aan een diversiteit voorouder populatie. De resultaten beschreven in dit hoofdstuk tonen ook een Afrikaanse taurine oorsprong aan in Iberische runderen en dit is in lijn met bevindingen uit eerdere studies. Verder laat ik zien dat, vergelijkbaar met Centraal Italiaanse runderrassen, ook Busa een complexe niet-Europese, Afrikaans taurine en zebu, oorsprong laat zien, mogelijk als gevolg van een gezamenlijke oorsprong van de diverse Balkan en Italiaanse runderen.

Hoofdstuk 4 beschrijft een studie naar de verwantschap, diversiteit en demografische geschiedenis van lokale Zweedse runderrassen op basis van genoom-wijde SNP merkers. De resultaten wijzen erop dat deze rassen geassocieerd kunnen worden in twee genetische clusters: Zweedse bergrunderassen (inclusief Bohus Polled) en gehoornde runderrassen van Zuid Zweden. Interessant is ook dat ik in staat was om een substructuur aan te tonen binnen de Fjallanra populatie, welke correspondeert met de boerderij waarvan de dieren afkomstig zijn. De resultaten tonen verder een relatief hoge genetische diversiteit aan van de Zweedse bergrunderen. Dit in tegenstelling tot de lage genetische diversiteit van de rassen Vaneko en Ringamalako, die verklaard kan worden uit de genetische isolatie van deze rassen. Een vergelijking van de genotypeerde data van

diverse noordwest Europese runderrassen toont een geringe bijdrage aan de genetisch diversiteit van lokale Zweedse runderen.

In hoofdstuk 5, beschrijf ik de identificatie en distributie van CNVs (afkorting voor de Engelse term “copy number variation”). De resultaten, gebaseerd op de signaal intensiteit van SNP genotyeerde data, tonen verschillen aan in de CNV profielen van verschillende Europese runderrassen waarschijnlijk het gevolg van verschillen in demografie en selectie. Een verrijkingsanalyse toont aan dat er meer CNVs gevonden worden in genoomregio’s gerelateerd aan reukzin en afweermechanismen. In het Engelse ras “English Longhorn” vond ik een structurele variatie rond het *KIT* gen, die eerder gevonden was in het runderras Belgische Blauwe. Deze structurele variatie speelt waarschijnlijk een rol bij de specifieke kleurvarianties binnen dit ras.

Hoofdstuk 6 beschrijft de resultaten van drie verschillende methoden voor het opsporen van structurele variaties in het genoom van runderen gebaseerd op complete genoom sequenties. Vergelijkbaar met de resultaten beschreven in hoofdstuk 5, worden verschillen in SV profielen gevonden tussen zebu, Afrikaanse en Europese runderen die wijzen op het effect van verschillen in demografie. Echter, deze verschillen kunnen voor een deel ook het resultaat zijn van het gebruik van het referentiegenoom, dat afkomstig is van een Europees individu. In de analyse, worden ook een aantal ras specifieke SV gevonden binnen diverse runderrassen, wat een rol hiervan suggereert in populatie differentiatie. Ook in het oeros sample konden CNVs aangetoond worden en deze resultaten tonen een hoge mate van overeenkomst met CNVs van moderne Europese runderen. Dit kan verklaard worden uit de vrij recente splitsing tussen deze twee populaties. De verschillende methoden resulteerden uiteindelijk in een groot aantal SV waarvan velen geassocieerd blijken te zijn met kenmerken als vleeskwaliteit, kleur van de vacht en metabolisme, allen goede targets voor toekomstige associatiestudies.

In hoofdstuk 7, tenslotte, bediscussieer ik alle resultaten die in de voorgaande hoofdstukken zijn beschreven in relatie tot de huidige kennis m.b.t. genomische vermenging, diversiteit en demografie van Europese runderen. Ik bediscussieer ook de sterkte en beperkingen van de diverse methoden zoals beschreven in dit proefschrift. Ik postuleer een aantal hypothesen voor demografische scenario’s gebaseerd op complexe patronen van genomische vermenging van Balkan en Italiaanse runderrassen. Daarnaast bediscussieer ik de mogelijkheid dat genetische variatie afkomstig van zebu, een belangrijke rol speelt in een verhoogde fitness van sommige Europese runderen. Verder benadruk ik de noodzaak voor additionele DNA samples van het oeros die verschillende tijdsfasen en geografische gebieden vertegenwoordigen. Deze zijn nodig om een beter inzicht te verkrijgen van het domesticatieproces van het rund in Europa. Ik sluit het hoofdstuk af met het benadrukken van de praktische toepassing van de informatie m.b.t. genetische diversiteit en populatiestructuur voor het behoud en beheer van lokale runderrassen.

Sammanfattning

Sammanfattning

En stor fenotypisk variation förekommer hos Europeiska nötkreatursraser, som populations-genetiskt kan hänföras till effekterna av migration, demografi och selektion. Dessa faktorer påverkar fördelningen av genetiska varianter, såsom SNP-markörer och strukturella variationer (SV), mellan olika populationer. Effekten av dessa faktorer samt distributionen av genetiska varianter är väl dokumenterade i kommersiella europeiska nötkreatursraser. Däremot är kunskapen begränsad om hur situationen är hos traditionella och primitiva europeiska nötkreatursraser. Denna avhandling syftar därför till att ge en ökad kunskap och en överblick om genomisk och genetisk variation hos de traditionella europeiska nötboskapsraserna med användning av SNP-markörer som täcker hela genomet.

I kapitel 2 visar jag att många europeiska nötboskapsraser har genvarianter som härstammar från både zebu och lokala uroax. Jag föreslår att den höga divergens som finns hos boskapsraser från Balkan och Italien kan hänföras till zebu-inslag. I detta kapitel ger jag också en övergripande beskrivning av genomisk variation mellan vilda lokala uroax och olika europeiska nötkreaturspopulationer. Dessa resultat indikerar ett signifikant genflöde mellan vilda lokala uroax och de tidigare populationerna av domesticerade europeiska nötkreatur. Dessutom visar jag att flera iberiska och italienska nötboskapsrasers effektiva populationsstorlekar nyligen har minskat.

I kapitel 3 samlade jag genotypdata för olika afrikanska nötkreatursraser samt raser av zebutyp som finns tillgängliga i publika databaser för att mer i detalj karakterisera graden av taurint (*bos taurus*)- respektive zebu (*bos indicus*)-ursprung i dessa boskapsraser. Dessutom använde jag genomsekvenser från specifika individer för att göra en jämförande utvärdering av genetisk mångfald i europeiska nötkreatursraser. Jag visar att italienska nötkreatur har hög genetisk variation som kan hänföras till att ett stort antal individer domesticerades. Inflytandet av afrikanska taurina nötkreatur som tidigare rapporterats hos iberiska nötkreatur bekräftades i detta kapitel. Även busa-djur från Balkan uppvisar ett komplext ursprung med genetiskt bidrag från både afrikanska taurin och zebu, vilket indikerar att vissa nötkreatursraser från Balkan och Italien har ett gemensamt ursprung.

I kapitel 4 studerades genetiskt släktskap, variation och demografisk historia hos alla kända svenska lantraser. Vi genotypade hela arvsmassan med en uppsättning av SNP markörer. Resultaten visade att dessa nötkreatursraser kan klassificeras i två genetiska grupper: svenska fjällkoraser (inklusive bohuskulla) och nötkreatursraser från södra Sverige (väneko och ringamålako) vilka båda har horn. Intressant nog kunde jag också påvisa understrukturer inom rasen fjällnära ko som motsvarade olika ursprungsgårdar. Dessutom indikerade resultaten också relativt hög grad av genetisk variation hos svenska fjällkor. Däremot observerades endast låg grad av genetisk variation hos både väneko och ringamålako vilket kan hänföras till deras genetiska isolering. Analyserna visade dessutom att de svenska lantraserna saknar eller har mycket lågt släktskap med andra analyserade nordvästeuropeiska nötkreatursraser och därmed har en unik uppsättning genetiska varianter.

I kapitel 5 identifierades kopietsvariation (copy number variation; CNV) och hur dessa är dis-

tribuerade hos olika europeiska nötkreatursraser. Det gjordes genom att analysera signalintensitet från SNP-genotypning av arvsmassan. Resultaten påvisade skillnader i CNV-profil mellan olika europeiska nötkreaturspopulationer, vilket indikerar effekter av demografi och selektion. CNV visades vara vanliga i genomiska regioner som innehåller gener med funktioner relaterade till olfaktoriska processer och immunsystemet. Dessutom identifierades hos rasen engelsk långhornad boskap strukturella variationer som involverar *KIT*-gener, som tidigare har förknippats med en specifik pälsfärg hos rasen belgisk blå.

I kapitel 6 användes tre olika analysmetoder av helgenomsekvensdata (paired end read, split read, samt läsdjup), för att identifiera strukturell variation (SV) i enskilda helgenomsekvenser av nötkreatur. I likhet med vad som observerades i kapitel 5 upptäcktes även skillnader i SV-profil mellan afrikansk taurin, zebu och europeisk taurin, vilket indikerar effekten av olika demografi. Resultaten kan delvis ha påverkats av att referensgenomet är från en europeisk taurin individ. Analysen identifierade också linjespecifika SV i olika boskapspopulationer, vilket indikerar att SV har en betydelse vid differentiering av populationer. Dessutom identifierades även CNVs i sekvenser som kommer från arkeologiska uroxe-prover. Resultaten visade mycket stora likheter mellan moderna europeiska nötkreatur och uroxen vilket bäst kan förklaras av den relativt korta tid som har förflutit sedan domesticeringen samt ett visst genflöde mellan populationerna innan uroxen dog ut. Sammanfattningsvis identifierades många SV relaterade till köttkvalitet, pälsfärg och metabolism. Resultat som kan vara värdefulla att beakta vid framtida genetiska associeringsstudier.

Slutligen diskuterar och sammanfattar jag i kapitel 7 resultaten från alla tidigare kapitel i förhållande till vad som hittills är känt rörande genomiska varianter och deras demografi hos europeiska nötboskap, d.v.s. hur populationer har korsats och därigenom påverkat varandra genom effekterna av s.k. admixture. Jag diskuterar också styrkan och begränsningarna i de metoder som används i avhandlingen. Jag lägger fram hypoteser om ett demografiskt komplext scenario där nötkreatur från Balkan och Italien har korsats och vilka effekter detta har haft. Jag diskuterar också möjligheten att inkorsning av zebu kan ha spelat en viktig roll genom att ge ökad fitness i vissa europeiska nötkreatursraser. Jag diskuterar också behovet av att sekvensera ytterligare uroxe-prover som representerar olika tidsperioder och geografiska områden för att svara på frågor som rör domesticering i Europa. Jag konkluderar med att belysa vilken praktisk implikation kunskapen om genetisk variation och struktur har i bevarandet av nötkreaturslantraser i olika länder i Europa.

Acknowledgements

Acknowledgements

I would like to take this opportunity to express my gratitude to all those people who have contributed to this thesis as well as supported me during this wonderful journey of my PhD. John Donne aptly remarked that “no man is an island”, indeed, human being can only thrive when he is a part of community. I feel very lucky that for the period of four years I was a part of this ABGC community at Wageningen University and Research (WUR). Moreover, as this PhD project was a collaborative effort of WUR and Swedish University of Agricultural Sciences (SLU), Uppsala, I also got opportunity to carry out part of my research in Department of Animal Breeding and Genetics at SLU which has enriched my professional expertise.

I am grateful to my WUR supervisors Prof. Dr. Martien Groenen and Dr. Richard Crooijmans for their guidance and constructive criticism on my scientific work which have shaped my scientific attitude. In fact, I also appreciate their patience which they have shown at a time when I was unable to meet deadlines. Without their faith in me, this thesis might not have seen the light of day. Overall, they have played a major role in nurturing the researcher out of me. I am also grateful to my SLU supervisors Prof. Dr. Göran Andersson and Dr. Sofia Mikko for their supervision and advices on analysis and wet laboratory procedures that I carried out during my PhD.

I am thankful to all the present and past members of weekly genomics meeting group—Ole, Hendrik-Jan, Kyle, Juanma, Mirte, Vinicius, Martijn, Chiara, Langqing, Zhou, Henri, Lim—for bearing my presentation and their scientific suggestions. Although I have learnt one thing or another from everyone in the group, I would like to single out Martijn from whom I learnt a lot in bioinformatics. He is the person I would give credit to for improving my Python skills. I also thank Vinicius for his help in R programming. My special thanks to Mirte for her scientific suggestions and advices in carrying out population genomics analysis. I also would like to thank other member of ABGC community—Gwen, Bert, Lissette, Ada, Maya—for their continuous support during my PhD. Thank you to the researchers of ABG dept at SLU—Susanne E., Anna, Susanne G., Naveed, Gabriela—for their support during my PhD duration in Uppsala. Special thanks to Sandrine and Jovana for their guidance and suggestions regarding mobility and other matters related to European Graduate School in Animal Breeding and Genetics.

Thank you to all my friends in Wagenigen, especially, Priyanka di, Sidhuji and Tanvi, for the delicious dinner that we often prepared and had together. Also, I would cherish the time that we spent together playing “dumb-charades” and board games. I would also like to thank the members of “chess club” Wageningen with whom I spent every Thursday evening of the first six months after I arrived in Wageningen. Thanks to “freepik.com” for providing graphics for the thesis cover.

I am grateful to my ma and papa for their unconditional love and support throughout my career. I am also grateful to my grandma for her love and care. Special thanks to my sister Namrata and my brother-in-law Kunal for their love and support. In the end, I would like to thank my love Nehal for her constant support, care and company through thick and thin of my life so far.

Curriculum vitae

Curriculum vitae

About the author

Maulik Upadhyay was born on 23rd May 1988 in Junagadh district of India. After completing his bachelor's degree in veterinary sciences and animal husbandry (B.V.Sc. & A.H.) from Anand Agricultural University (AAU), he got enrolled for master's programme in animal breeding and genetics in the same university. During his masters, he also worked as a teaching assistant in the practical courses of Biostatistics and Animal Genetics which were offered to the undergraduate students. In his master's thesis, he analyzed whole genome sequencing data generated by Roche GS-FLX Titanium and Ion Torrent PGM sequencer to identify single nucleotide polymorphisms (SNPs) from buffalo genome. After completing his masters, he worked as a bioinformatician in the project of "Single Nucleotide Polymorphisms detection (SNP) in the coding region of the genome and its association with feed conversion ratio in broilers". In September 2014, he started his joint PhD under the aegis of the European Graduate School in Animal Breeding and Genetics (EGS-ABG). While most of his PhD research works were carried out at Wageningen (Netherlands), he had the opportunity to spend eight months at Uppsala (Sweden). During his PhD, he worked on characterizing genomic admixture and variation in native and primitive cattle breeds of Europe. The results of his PhD are presented in this thesis entitled "Genomic variation across European cattle: contribution of gene flow".

Peer-reviewed journal publications

M. R. Upadhyay, C. Bortoluzzi, M. Barbato, P. Ajmone-Marsan, L. Colli, J.A. Lenstra, C. Ginja, T. Sonstegard, M. Bosse, M.A.M. Groenen and R.P.M.A. Crooijmans, Deciphering the pattern of genetic diversity and admixture using Genome-wide SNPs in Southern European cattle (2019), *Evolutionary Applications*. doi:10.1111/eva.12770.

M. R. Upadhyay, V.H. da Silva, H.J. Megens, M.H.P.W. Visker, P. Ajmone-Marsan, V.A. Bălteanu, S. Dunner, J.F. Garcia, C. Ginja, J. Kantanen, M.A.M. Groenen and R.P.M.A. Crooijmans, Distribution and Functionality of Copy Number Variation across European Cattle Populations (2017), *Frontiers in Genetics*, 8. doi: 10.3389/fgene.2017.00108.

M. R. Upadhyay, W. Chen, J.A. Lenstra, C.R. Goderie, D.E. MacHugh, S.D. Park, D.A. Magee, D. Matassino, F. Ciani, H.J. Megens, J.A.M. van Arendonk, P. Ajmone-Marsan, V.A. Bălteanu, S. Dunner, J.F. Garcia, C. Ginja, J. Kantanen, M.A.M. Groenen and R.P.M.A. Crooijmans, Genetic origin, admixture and population history of aurochs (*Bos primigenius*) and primitive European cattle (2017), *Heredity*, 118(2), 169–176.

R. B. Onzima, **M. R. Upadhyay**, H.P. Doekes, L.F. Brito, M. Bosse, E. Kanis, M.A.M. Groenen and R.P.M.A. Crooijmans, Genome-wide Characterization of Selection Signatures and Runs of Homozygosity in Ugandan Goat Breeds (2018), *Frontiers in Genetics*.doi:10.3389/gene.2018.00318

R. B. Onzima,**M. R. Upadhyay**, R. Mukiibi, E. Kanis, M.A.M. Groenen and R.P.M.A. Crooijmans, Genome-wide population structure and admixture analysis reveals weak differentiation among Ugandan goat breeds (2018), *Animal Genetics*, 49, 59-70. doi: 10.1111/age.12631.

T.M. Shah, N.V. Patel, A.B. Patel, **M.R. Upadhyay**, A. Mohapatra, K.M. Singh, S.D. Deshpande, O.G. Joshi, A genome-wide approach to screen for genetic variants in broilers (*Gallus gallus*) with divergent feed conversion ratio (2016), *Molecular Genetics and Genomics*, 291,1715–1725.

A.C.Patel, T.K.Jishaa, D. Upadhyay, R. Parikh, **M.R. Upadhyay**, R. Thaker, S. Das, J.V. Solanki, D.N. Rank, Molecular characterization of camel breeds of Gujarat using microsatellite markers (2015), *Livestock Science*, 181, 85-88.

M. R. Upadhyay, A.B. Patel, R.B. Subramanian, T.M. Shah, S.J. Jakhesara, V.D. Bhatt, P.G. Koringa, D.N. Rank & C.G. Joshi, Single nucleotide variant detection in Jaffrabadi buffalo (*Bubalus bubalis*) using high-throughput targeted sequencing (2015), *Frontier in Life Sciences*, 8(2), 192-199.


Manuscripts under review

M. R. Upadhyay, S. Eriksson, S. Mikko, E. Strandberg, M.A.M. Groenen and R.P.M.A. Crooijmans, G. Andersson and A.M. Johansson, Genomic relatedness and diversity of Swedish native cattle breeds. *Genetics Selection Evolution* (2019).

M. R. Upadhyay, M.F.L. Derks, G. Andersson, M.A.M. Groenen and R.P.M.A. Crooijmans, Comparative evaluation of structural variations in taurine and indicine cattle using individual whole genome sequences. *BMC Genomics* (2019).

M. Barbato, M.D. Corvo,**M. R. Upadhyay**, E. Kim, F. Hailer, L. Colli, R. Negrini, R.P.M.A. Crooijmans, T.Sonstegard, P. Ajmone-Marsan, Signals of adaptive introgression in white cattle breeds from Central Italy. *Evolutionary Applications* (2019).

Training and Supervision Plan

Training and Supervision Plan (TSP)				 Graduate School WIAS	
Section 3. EDUCATION AND TRAINING (minimum 30 credits)					
A. The Basic Package		year		credits *	
WIAS Introduction Day (mandatory)		2014		0.3	
Course on philosophy of science and/or ethics (mandatory)		2014		1.5	
European Graduate school of Animal Breeding and genetics (Introduction week)		2014		2.0	
Subtotal Basic Package					4
B. Disciplinary Competences		year		credits	
Introduction to Phylogenetics analysis using R		2014		2.0	
The sustainability concept in Animal Breeding		2015		2.0	
Data management and Planning		2015		0.4	
Emerging technologies in Animal Breeding		2017		1.5	
Population Genomics: background and tools		2017		3.0	
Code club meeting of Wageningen University and Research		2017-2018		1.0	
HPC (High performance computing) Advanced Course		2018		0.2	
Summer school France		2018		0.8	
Linear Models in Animal Breeding		2018		3.0	
HPC Basic course		2018		0.2	
WIAS Course Statistics for the Life Sciences		2018		2.0	
M.Sc. Level course					
Programming in Python		2014		6.0	
Subtotal Disciplinary Competences					22
C. Professional Competences		year		credits	
Course on essential skills (Frank Little) (recommended)		2014		1.2	
Techniques for Writing and Presenting a Scientific Paper (TWP)		2015		1.2	
Effective behavior in your professional surroundings (EB)		2016		1.3	
project and Time management		2016		1.0	
Writing a grant proposal		2017		2.0	
Searching and Organizing literature		2018		0.6	
How to beat Procastination		2018		0.3	
Subtotal Professional Competences					8
D. Presentation Skills (maximum 4 credits)		year		credits	
< title of presentation, name of conference/seminar, place, date, oral / poster >					
1). Genetic structure of primitive European cattle breeds and aurochs	Wageningen PhD symposium	Wageningen, 26th April, 2016	Oral	2016	1.0
	Mini-symposium on the search for genotypic, phenotypic & ecotypic equivalents of an extinct species	Uppsala, Oct.6, 2016	Oral	2016	1.0
2). Genetic structure of primitive European cattle breeds and aurochs					
3). Genetic differentiation of primitive cattle breeds illustrate aurochs admixture	50th Population Genetics Group Meeting	Cambridge, 4th-7th January, 2017	Oral	2017	1.0
4). Characterization of copy number variation in European cattle	International Conference on Animal Genetics, 2017	Dublin, 16th-21st July, 2017	Poster	2017	1.0
5). Inferring pattern of genomics admixture in Southern European cattle breeds	Joint Congress on Evolutionary Biology, 2018 DrosEU Summer School on Adaptation Genomics	Montpellier, France, 23'rd-24th August, 2018	Poster	2018	0.0
Subtotal presentations					4
E. Teaching competences (max 6 credits)		year		credits	
Supervising thesis (B.Sc.)		2015		1.0	
Tutorship		2015		1.8	
Subtotal Teaching competences					3
Education and Training Total (minimum 30 credits)*					40

Data availability and supplementary material

Data availability and supplementary material

All the genotyping and signal intensity data have been submitted in DRYAD with links provided in the manuscripts published based on the chapter 2 and chapter 5 of this thesis.

The supplementary material for the chapter 2 and chapter 5 are available through the journal websites:

chapter 2: (<https://www.nature.com/articles/hdy201679#supplementary-information>)

chapter 5: (<https://www.frontiersin.org/articles/10.3389/fgene.2017.00108/full#h11>)

The supplementary materials for the unpublished articles are uploaded on google drive, please visit this link to download the supplementary materials:

https://drive.google.com/file/d/1JxUqg25NKEG1ZRXnLGvot8Bb3n4_QMfV/view

Colophon

Colophon

The work performed in Chapter 4 was financed by the funds of Swedish Research Foundation (FORMAS).

The author was supported by the European Commission (within the framework of the Erasmus-Mundus joint doctorate “EGS-ABG”) and Taurus foundation (Stichting Taurus), Nijmegen, The Netherlands.

The cover of this thesis was designed by Nehal Ranabhatt and Razia Sultana.

The thesis was printed by Digiforce — Proefschriftmaken.nl, De Limiet 26, 4131NC, Vianen, the Netherlands.