Genome-wide identification of signatures of positive selection in African admixed zebu cattle

Hussain Bahbahani, BSc (Hons), MRes

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

March 2015

Declaration

I hereby declare that this thesis has not been previously presented or submitted for examination to this or any other university. References to other people's work are duly acknowledged. The work described herein has been carried out by me except otherwise stated.

Hussain Bahbahani
University of Nottingham
March 2015

Abstract

The small East African shorthorn zebu (EASZ) is one of the cattle types in East Africa. It is an ancient, stabilised zebu-taurine admixed cattle population. In comparison to the recently introduced, highly productive, exotic taurine cattle, EASZ are preferred by local farmers due to their superior adaptability to the local tropical African environment. They require minimal veterinary care, may sustain themselves on poor feed quality and are thermotolerant. Understanding the genetic control of their adaptations will help sustainable breeding improvement programs, both within populations and through crossbreeding. An important step to achieve this goal is characterising the genome of this indigenous cattle population and identifying signatures of positive selection.

Several advanced tools are now available for genomic studies in cattle, including genome-wide single nucleotide polymorphism (SNP) arrays and next-generation sequencing of the full genome and the exome. Here, low- and high-density, genome-wide SNP data were first analysed to identify candidate signatures of positive selection in EASZ using allele frequencies, linkage disequilibrium and composite tests. Because of limited genome coverage and SNPs' ascertainment bias issues, the full genome (ABI SOLiD sequence data) of 10 pooled EASZ samples was also investigated for candidate sweep regions by assessing the pooled heterozygosity (*Hp*) of 100 kb sliding genome windows. The full exome sequences of an additional 10 EASZ were also explored to identify signals of copy number variants (CNVs) in their coding sequences and to estimate the allele frequencies of specific candidate causative mutations. Mitochondrial DNA (mtDNA) sequences of African cattle, including 13 EASZ samples, were also analysed to investigate their diversity, as well as to identify signatures of positive selection.

Analysing the EASZ genome with the Illumina BovineSNP50 BeadChip v.1, using two Extended Haplotype Homozygosity (EHH)-based (intra-population *iHS* and inter-population *Rsb*) analyses and an inter-population F_{ST} approach, identified 24 candidate genome regions under positive selection. Genes associated with adaptation to the African environment (e.g., reproduction and fertility and immune response) were found within these regions. A total of 18 regions had previously been identified in tropical-adapted cattle, while four regions were shown to be under positive selection in commercial breeds. Out of these 24 regions, six showed a substantial (≥ +/- 1 standard deviation from the mean) excess (one region) or deficiency (five regions) in the Asian zebu ancestry, thereby inferring the possible origin of the selected haplotypes.

The genome coverage and the SNPs' ascertainment bias issues were addressed in Chapters 3 and 4 by using the Illumina BovineHD BeadChip and the full EASZ genome sequence to detect signatures of positive selection on EASZ autosomes and the BTA X sex chromosome. A total of 101 and 165 autosomal candidate sweep regions were identified *via* genome-wide SNP analyses and full genome *Hp* analysis, respectively. Out of these regions, 35 were common between the two groups. The analyses of high-density, genome-wide SNP data of zebu cattle populations from Uganda and Nigeria allowed us to classify 15 and seven regions as East African zebu-sharing and East and West African zebu-sharing candidate regions, respectively. Furthermore, the sizes of these regions were fine mapped up to ~ 93 kb. In the EASZ BTA X sex chromosome (Chapter 4), 20 candidate regions (six by *Rsb*, two by *iHS*, and twelve by *Hp*) were identified. Four of these regions were also detected in zebu cattle populations from Uganda or Nigeria (two in each population). Seven regions demonstrated a substantial deviation from the mean zebu ancestry, thereby suggesting the possible origins of the selected haplotypes (three indicine and four taurine origins).

Genes and quantitative trait loci (QTL) associated with adaptive traits (e.g., reproduction, immunity and heat stress) were found within the identified candidate regions. Several SNPs and indels (insertion/deletion) were detected using the 10 pooled EASZ full genome sequences. Four autosomal and five BTA X non-synonymous variants were considered as possible candidate causative mutations targeted by selection, although none were fixed in our

population. Qualitative signals of CNV (multiple copies) in 17 autosomal and two BTA X genes within the candidate regions were identified using the EASZ exome sequences.

The full mtDNA sequences of 13 EASZ samples were affiliated to three T1 sub-haplogroups (T1a, T1b and T1b1). Overall, no selective advantage was found to be associated with taurine mtDNA over the zebu type, which is in agreement with a male-mediated introgression of Asian zebu cattle into African taurine cattle; this explains the sole presence of taurine mtDNA in African cattle. Purifying selection was found to be the main selective pressure on bovine mtDNA, with less selective constrains on the *ATP6* and *ATP8* genes. Interestingly, in the *Cox-2* gene of the T1b/ T1b1 sub-haplogroups (together the most common sub-haplogroups in African cattle), a single nucleotide mutation leading to an amino acid change may confer a selective advantage, which calls for further data and analyses.

Brought into Africa and shaped by their environment, African cattle represent a unique livestock genetic resource for sustainable food production on the continent. This thesis has revealed that selection pressures have uniquely shaped the genome of African cattle. It is just a first milestone. Now, we need to identify the causative mutation(s), which will require full and targeted genome sequencing of a large number of animals and populations. Then, the door will be opened to use the traditional marker-assisted introgression approach and/or genetic engineering techniques, such as transcription activator-like effector nucleases (TALENs) and CRISPR (clustered, regularly interspaced, short, palindromic repeats)/Cas9 (CRISPR-associated) system, to introduce the favourable mutations into new ‘composite’ cattle for the benefit of farmers and inhabitants of the African continent.

Published papers

BAHBAHANI, H. & HANOTTE, O. 2015. Genetic resistance: tolerance to vector-borne diseases, prospect and challenges of genomics. *OIE Scientific and Technical Review*, 34 (1)**,** 185-197.

BAHBAHANI, H., CLIFFORD, H., WRAGG, D., MBOLE-KARIUKI, M. N., VAN TASSELL, C. P., SONSTEGARD, T., WOOLHOUSE, W. & HANOTTE, O. 2015. Signatures of positive selection in East African Shorthorn Zebu: A genome-wide single nucleotide polymorphism analysis. *Scientific Reports*. *In press*.

Acknowledgements

This thesis would not be completed without the efforts of many people which deserve my deepest gratitude. I would like first to start with the chief of this work, my supervisor Professor Olivier Hanotte. Thank you for every moment you spent teaching me how to be a scientist. I cannot forget the stressful moments in this journey which without you standing next to me I would not be able to pass them. Beside the PhD work, you gave me the main tools to face my future life.

Thanks go to the people who supplied me with the cattle SNP data; Dr. Mary Ndila Mbole-Kariuki (AU-IBAR, Nairobi) for the EASZ SNPs data, Dr. Tad Sonstegard (USDA-ARS, Maryland) for the reference cattle SNPs data, Dr. Heather Huson (USDA-ARS, Maryland) for the Ugandese cattle SNPs data, Dr. Oyekanmi Nash (NABDA, Nigeria) and Dr Christopher Mukasa (Ahmadu Bello University, Nigeria) for the Nigerian cattle SNPs data. I would like to thank Dr. Martin Blyth and Dr. Sunir Malla from the Deep Seq facility at the University of Nottingham for their efforts in generating the full EASZ genome and exome sequence data. I also would like to thank Mr. Harry Clifford for his input in writing the scripts used in Chapter 2 during his master degree at University of Nottingham. A big thank you to Professor Mark Woolhouse from University of Edinburgh for all the discussions we had regarding the work in this thesis. Many thanks to Dr. Joram Mowacharo for his work in sequencing the full EASZ mtDNA, which is the nucleus of the fifth chapter of this thesis. I also would like to thank my annual assessor Dr. Sara Goodacre (University of Nottingham) for our continuous discussions that shaped this thesis massively. I do not want to forget my lab mates, especially Dr. David Wragg and Dr. Faisal Almathen, for every idea they introduced to improve this thesis.

Last but not least, I would like to thank my whole family and the two most important persons in my life; my parents. These two persons have a great belief on me. They were always the shelter I would look for when life become so hard. Every word they encouraged me with was a real push to continue these four years.

I would like to end by thanking Kuwait University for funding this work with a generous PhD scholarship.

Table of Contents

Acronyms and abbreviations

AG	Adamawa Gudali
AO	Ankole
AZ	Azawak
BC	Before Christ
BJ	Bunaji
BP	Before present
bp	Base pair
BTA	*Bos taurus taurus* chromosome
CNV	Copy number variation
CRISPR/Cas	Clustered, regularly interspaced, short, palindromic repeats/ CRISPR-associated
EASZ	East African shorthorn zebu
GATK	Genome Analysis Toolkit
GIR	Gir
HD	High density
HOL	Holstein-Friesian
IBS	Identity-by-state
iHS	Integrated Haplotype Score
indel	insertion/deletion
ILRI	International Livestock Research Institute
JER	Jersey
kb	Kilo base pair
KR	Karamojong zebu
LD	Linkage disequilibrium
MAF	Minor allele frequency
Mb	Mega base pair
MT	Muturu
mtDNA	Mitochondrial DNA
NDM	N’Dama
NEL	Nellore
NG	Nanda
NGR	Nigeria
OR	Red bororo
QC	Quality control
qRT-PCR	Quantitative real-time polymerase chain reaction
QTL	Quantitative trait loci
SD	Standard deviation
SDOC	Standardised depth of coverage
SNP	Single nucleotide polymorphism
SO	Sokoto Gudali
TALENs	Transcription activator-like effector nucleases
UGN	Uganda
WD	Wadara
YK	Yakanaji
ZS	Serere zebu

Chapter one

Opening scene

History of cattle

Cattle types and their geographical distribution

Bos is one of the five genera of the Bovidae family, which also contains *Bubalus*, *Syncerus*, *Bibos* and *Bison*. Ten species are currently recognized within the *Bos* genus, including *Bos taurus*, which is divided into two subspecies: *Bos taurus taurus* (taurine cattle) and *Bos taurus indicus* (zebu or indicine cattle) (Epstein, 1971, Epstein and Mason, 1984, Loftus *et al*., 1994, MacHugh *et al*., 1997) (Figure 1.1).

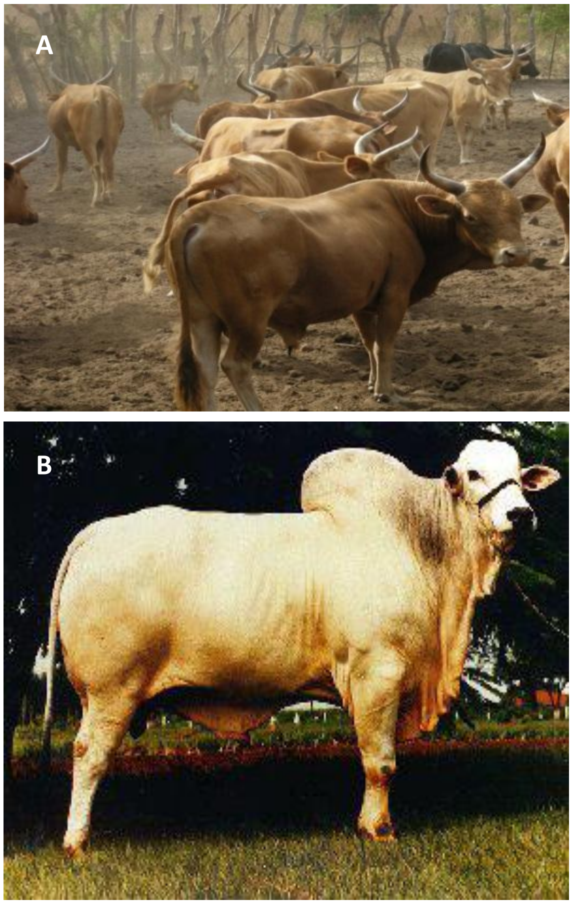

Figure 1.1: The humpless *Bos taurus taurus* cattle (African taurine: N'Dama) (A). The humped *Bos taurus indicus* cattle (Asian zebu: Nellore) (B).

Morphologically, these two cattle types are mainly differentiated by the presence of a thoracic hump and large dewlap in zebu cattle. Based on horn size, cattle have been divided into long-horned and short-horned cattle. Examples of long-horned cattle include the African taurine (N'Dama) and African zebu (Ankole). The European taurine (Holstein-Friesian) and Asian zebu (Nellore) are both two examples of the short-horned type (Epstein, 1971). The taurine and zebu karyotypes are identical, with 29 autosomal pairs and one sex chromosome pair, with the exception of the Y chromosome, which is sub-metacentric in taurine cattle and acrocentric in zebu cattle (Kieffer and Cartwright, 1968). In Africa, ancient zebu-taurine crossbreeds are often called sanga (Epstein, 1971). This type of cattle, which includes Sheko, Tuli and Nguni cattle, are mainly characterized by their small cervico-thoracic hump (Epstein, 1971, Frisch *et al.*, 1997).

Excluding commercial breeds that often have a worldwide distribution (e.g., Holstein-Friesian) (FAO, 2007b), indigenous cattle types/populations show specific geographic distributions (Figure 1.2) (FAO, 2007a). They are found in all habitats, except in extreme environments such as deserts (e.g., the Sahara), polar and circumpolar areas.

A distinct geographic distribution pattern shaped by human history (e.g., centres of domestication, human migration and trading) and environmental adaptation is observed in the two main cattle types (zebu and taurine). The Indian subcontinent, South America, and most of Africa are mostly populated with zebu cattle. In the more temperate conditions between Western Europe and Eastern Asia, taurine cattle are widespread (Bradley *et al.*, 1998). Taurine cattle (e.g., Baladi, N'Dama and Muturu) are also present in Northern Africa and the humid and sub-humid regions of West Africa (Epstein, 1971, Rege and Tawah, 1999, Porter, 2002).

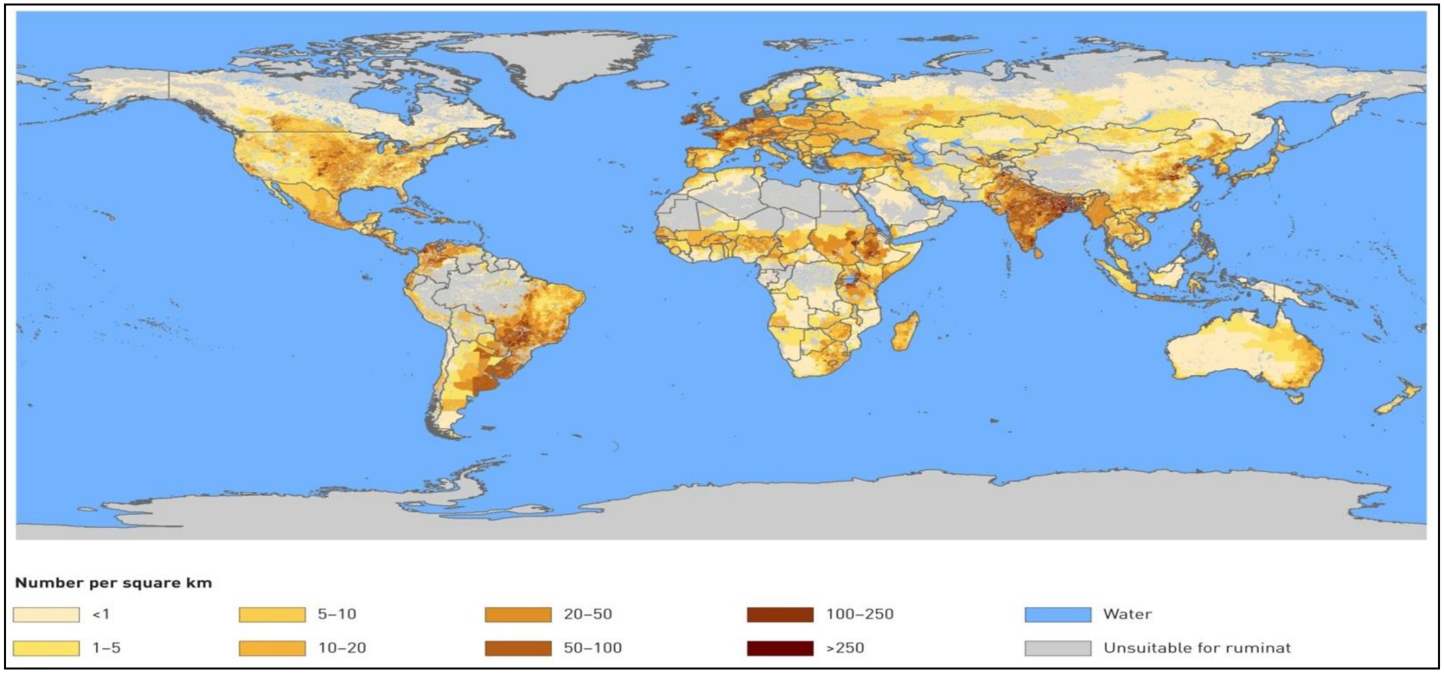

Figure 1.2: The worldwide distribution of cattle and their densities (FAO, 2007a).

Cattle domestication and genetic variation

The geographic origin and the number of centre(s) of domestication for cattle are still debated. Molecular and archaeological evidence supports two or three domestication centres for cattle (the Near East, the Indian subcontinent and North Africa) (Adametz, 1920, Grigson, 1991, Loftus *et al*., 1994, Bradley *et al*., 1996, MacHugh *et al*., 1997, Stock and Gifford-Gonzalez, 2013).

The two cattle subspecies were domesticated from different ancestral auroch *Bos primigenius* populations. They have been divided into three continental subspecies: *Bos primigenius namadicus* (Asia), *Bos primigenius opisthonomus* (Africa) and *Bos primigenius primigenius* (Europe). The last auroch was killed in 1627 in Poland (Epstein, 1971).

Initially, it was proposed that both taurine and zebu cattle arose from the wild ancestor *Bos primigenius namadicus via* a single domestication event around 10,000 years BP in the Near East (Epstein, 1971, Epstein and Mason, 1984, Payne, 1991). This was disproved by Loftus *et al*. (1994) through mtDNA analysis of taurine and zebu cattle from three different continents (Europe, Africa and Asia). In this study, there was a clear sequence dichotomy between Indian zebu and taurine cattle from Africa and Europe, with an estimated divergence date of at least 200,000 years BP, well before the Neolithic period. This clearly indicates that zebu and taurine cattle were domesticated separately.

A subsequent mtDNA analysis by Bradley *et al*. (1996) indicated that within the taurine, European and African continental-specific mtDNA D-loop lineages are present with a divergence dated to 22,000–26,000 years BP. Therefore, these two domesticated lineages could have expanded from two different putative centres of domestication in the Near East (5,000 BP) and in North Africa (9,000 years BP). Autosomal microsatellite analysis (MacHugh *et al*., 1997) further supported a divergence between the zebu and taurine genomes about 700,000 years BP, as well as a divergence between African and European taurine cattle 200,000 years BP.

More recently, mtDNA analyses of several Asian zebu cattle populations have shown that the previously defined zebu haplogroups I1 and I2 (Baig *et al*., 2005, Lai *et al*., 2006) demonstrated higher nucleotide diversity in Indian subcontinent populations than in other parts of Asia (Chen *et al*., 2010). In parallel with previous archaeological evidence (Meadow, 1993), this supports the hypothesis that modern zebu cattle were domesticated from their ancestor auroch *Bos primigenius namadicus* on the Indian subcontinent, likely in the Indus Valley.

The Near East as the centre of domestication for taurine cattle has been supported archeologically by the discovery of ancient faunal remains of domestic and wild cattle at Çatal Hüyük in Anatolia by 7,800 years BP (Perkins, 1969). Additionally, higher mtDNA diversity has been observed in taurine cattle from the Near East, with several haplogroups (T, T2, T3 and, rarely, T1), than in European (haplogroup T3) and African (haplogroup T1) cattle (Troy *et al*., 2001) (Figure 1.3). The phylogenetic topology of these haplogroups, i.e., the star-like pattern, suggests a past population expansion. This further validates the view that taurine cattle were domesticated in the Near East from the wild auroch *Bos primigenius* and subsequently migrated to Europe and Africa.

The scarcity of the African mitochondrial haplogroup (T1) in the putative taurine domestication centre, the Near East (Figure 1.3), has raised the question of a separate African taurine domestication event in North Africa. This hypothesis was introduced previously by some archaeological studies (Adametz, 1920, Wendorf, 1998) and discussed later by Bradley *et al*. (1996) based on mtDNA analysis.

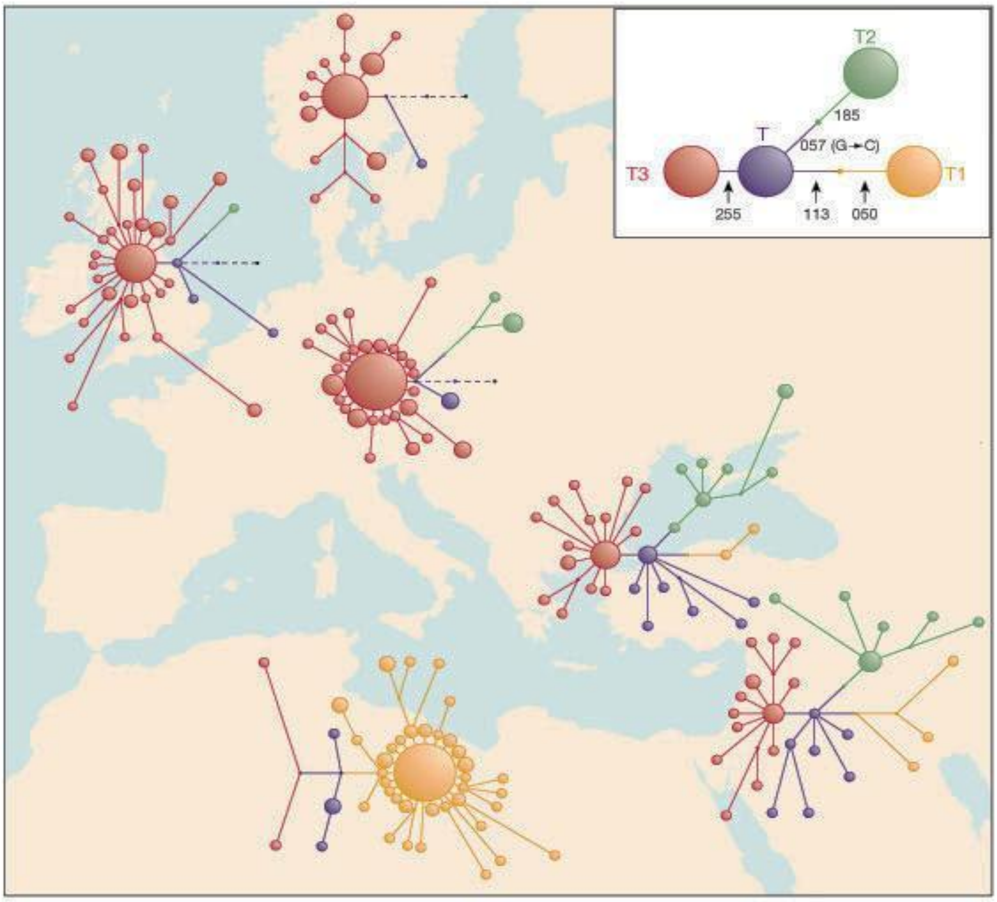

Figure 1.3: The pattern of domestic taurine cattle mtDNA haplogroup diversity in Africa, Europe and the Near East (Troy *et al*., 2001).

However, recent molecular evidence has cast doubt on a possible African centre of origin for domestic cattle (Achilli *et al*., 2008, Bonfiglio *et al*., 2012). Indeed, a full mtDNA analysis has shown that not only a small number of mutations separate the European and African D-loops, but also the full African mtDNA variation is embedded within the diversity of European and Near East mtDNA variation.

African cattle and East African shorthorn zebu

The history of African cattle started with the migration of taurine populations from their centre(s) of domestication in the Near East to Northeast Africa through Egypt about 5,000 year BC, as supported by pictorial representations and archaeological remains (Gifford-Gonzalez and Hanotte, 2011). Following this introduction, these taurine moved westward and southward across the African continent (Hanotte *et al*., 2002). Subsequent to this migration, zebu cattle were introduced to Africa from the east, through the Horn of Africa, from their centre of domestication on the Indian subcontinent. This introduction took place in two waves (~2000 BC and ~700 AD), the second of which was likely associated with Arab trading and the Islamization of the continent (Hanotte *et al*., 2002). Due to the sole presence of taurine mtDNA in African zebu cattle (Loftus *et al*., 1994, Bradley *et al*., 1996, Salim *et al*., 2014), and the spread of an indicine Y chromosome allele in several African taurine populations (Hanotte *et al*., 2000), it is proposed that the introgression of zebu cattle to Africa was mainly male-derived. More recently, after the second wave of zebu introduction to Africa (~700 AD), the movement of cattle with a predominantly zebu genetic background was accelerated towards the western and southern parts of the continent. This was likely facilitated by rinderpest outbreaks that occurred in Africa at the end of the 19th century, and for which Asian zebu cattle are believed to be more resistant (compared with taurine cattle) (Epstein and Mason, 1984, MacHugh *et al*., 1997, Hanotte *et al*., 2002). Some African taurine populations inhabiting high tsetse fly challenge areas, e.g., N'Dama from Guinea, have not shown any zebu introgression (Bradley *et al*., 1994, Hanotte *et al*., 2000). This is due to the susceptibility of zebu cattle to the disease transmitted by these flies, trypanosomosis, to which African taurine cattle are more tolerant (Murray *et al*., 1982). The influx of zebu ancestry into the local African taurine cattle peaks at the zebu entry point to Africa, i.e., East Africa, while demonstrating a gradual decline westward and southward (MacHugh *et al*., 1997, Hanotte *et al*., 2002, Decker *et al*., 2014).

Several cattle populations have been recognised in East Africa, e.g., Boran from Ethiopia, Kenana and Butana from Sudan, Karamajong from Uganda and Nandi from Kenya. They are phenotypically zebu, while genetically they have been shown to be of a zebu-taurine admixture (MacHugh *et al*., 1997, Hanotte *et al*., 2000, Rege *et al*., 2001, Gibbs *et al*., 2009). These cattle show a degree of adaptation to the different environmental pressures in East Africa, such as resistance to *Rhipicephalus appendiculatus* tick infestation (Latif *et al*., 1991a, Latif and Pegram, 1992), tolerance to poor forage and water (Western and Finch, 1986), and the ability to cope with thermal stress (Gaughan *et al*., 1999, Hansen, 2004).

East African indigenous zebu cattle are classified into small East African shorthorn zebu (SEASZ) and large East African shorthorn zebu (LEASZ). The SEASZ are the most common, and they are distributed across the humid and sub-humid agro-ecological zones of East Africa, whilst the LEASZ are generally restricted to drier areas (Rege *et al*., 2001). In Kenya, EASZ are differentiated into various populations according to tribal boundaries and socio-economic cultures, e.g., Kavirondo zebu in the Luo and Luhya communities, and Teso zebu in the Teso community (Rege *et al*., 2001). These cattle are mainly used for draft purposes, manure and milk production (Rege *et al*., 2001). A recent genome-wide autosomal single nucleotide polymorphism (SNP) analysis of Kenyan SEASZ revealed an even genomic admixture of around 84% zebu and 16% taurine ancestry across animals, believed to be partially shaped by selection and genetic drift (Mbole-Kariuki *et al*., 2014).

Characterisation of the cattle genome

Development of genomic tools

Evaluating the diversity and the genetic structure of various cattle populations has evolved over multiple stages until reaching today's full genome coverage stage. Initially, scientists relied on chromosome karyotypes to classify cattle. Because of the distinct Y chromosome karyotypes in the two cattle subspecies, acrocentric on zebu and sub-metacentric in taurine cattle (Kieffer and Cartwright, 1968, Stranzinger *et al*., 1987), some African zebu cattle were either defined as having zebu male ancestry, e.g., Malawi zebu, or taurine male ancestry, e.g., Tuli cattle (Meghen *et al*., 1994, Frisch *et al*., 1997).

Other markers have been used to study the population structures of different cattle populations, e.g., protein electrophoresis and Restriction Fragment

Length Polymorphism (RFLP). Both of these classes of genetic markers show low variation to effectively conduct population genetic analyses (Baker and Manwell, 1980, Theilmann *et al*., 1989).

Microsatellites, tandem repeats of very short (one to six base pair) nucleotide motifs, are widely used genomic markers in cattle population studies. Because of the high level of polymorphisms typically observed at microsatellite loci, they have been used to define the evolutionary relationships between cattle subspecies, the population levels of genomic admixture, migration history, as well as to map genomic quantitative trait loci (QTL), within species (MacHugh *et al*., 1997, Hanotte *et al*., 2002, Hanotte *et al*., 2003). Although microsatellite markers have demonstrated great success in increasing our understanding of the population structure and history of cattle, their relatively limited coverage of the bovine genome is a drawback.

Integrating recent advances in genomic tools into the bovine full genome characterization has opened new avenues for scientists to further analyse the genetic background of cattle populations. One of these advances is genotyping of the bovine genome with SNPs. These variants are arranged in customised arrays that are commonly referred to as DNA SNP chips. Two examples of these arrays for cattle, which are designed by Illumina, are the low SNP density array (BovineSNP50 Genotyping BeadChip, versions 1 and 2), which genotypes more than 54,000 SNPs (Matukumalli *et al*., 2009), and the higher density array (BovineHD Genotyping BeadChip), which genotypes more than 777,000 SNPs (Rincon *et al*., 2011). Both of these arrays genotype SNPs based on the dual-colour, single-base extension Infinium HD assay (Matukumalli *et al*., 2009). SNPs genotyped by these two arrays were validated mainly in commercial taurine cattle breeds for the purpose of genome-wide association analyses. This ascertainment bias to European taurine breeds, which is more substantial in the lower density array (Matukumalli *et al*., 2009), can make this tool less powerful in analysing the genome of zebu cattle and non-European indigenous taurine cattle populations. However, several research groups have successfully utilised this tool in different types of genomic analyses, such as the detection of genome regions with signatures of selection in several noncommercial indigenous cattle populations (Gautier *et al.*, 2009, Gautier and Naves, 2011, Utsunomiya *et al.*, 2013, Perez O'Brien *et al.*, 2014), the identification of genetically differentiated genome regions between zebu and taurine cattle (Porto-Neto *et al.*, 2013), or to determine the genomic structure of indigenous EASZ cattle (Mbole-Kariuki *et al.*, 2014). In addition to the Illumina version of these SNP arrays, Affymetrix has also designed equivalent SNP arrays, the Axiom Genome-Wide BOS 1 Array and the ultra-high-density Affymetrix BOS 1 pre-screening assay (AFFXB1P). Both of these arrays are based on an assay called molecular inversion probe hybridization (Fan *et al.*, 2010, Rincon *et al.*, 2011, Ramey *et al.*, 2013).

DNA sequencing is now considered to be the optimum tool that can be used to fully characterise the bovine genome. The automated Sanger DNA sequencing process (Sanger *et al.*, 1977) is highly efficient for sequencing targeted regions of the bovine genome, e.g., the D-loop of bovine mtDNA (Loftus *et al.*, 1994, Bradley *et al.*, 1996, Troy *et al.*, 2001, Chen *et al.*, 2010). However, the lengths of the obtained DNA sequences are relatively small (typically less than 1,000 bp).

Several next-generation sequencing platforms, such as Roche/454, Illumina and SOLiD, have been developed since 2008. Each of these platforms implements a specific sequencing chemistry and pipeline, as reviewed by Metzker (2010). The on-going improvement of these platforms, in terms of read length, sequencing speed and genotyping accuracy, is opening the door to the re-sequencing and *de novo* sequencing genomes at the population level in a cost-effective manner. These recently advanced technologies have been used in different livestock populations to identify genome regions bearing footprints of selection (Rubin *et al.*, 2010, Rubin *et al.*, 2012, Liao *et al.*, 2013), regions with copy number variation (CNV) (Rubin *et al.*, 2012, Bickhart *et al.*, 2012), and to understand their demographical history (Groenen *et al.*, 2012).

In addition to full genome sequencing, it is also possible to narrow down the sequencing target to the protein-coding regions of the genome, i.e., exome sequencing. This helps to reduce the cost associated with sequencing the whole genome, while increasing the depth of sequence coverage for the targeted regions (Teer and Mullikin, 2010, Singleton, 2011, Cosart *et al*., 2011). This approach has led to substantial progress in identifying causative mutations associated with different genetic defects, e.g., ocular birth defects in humans (Raca *et al*., 2010) and hereditary perinatal weak calf syndrome in Japanese Black cattle (Hirano *et al*., 2013). Recently, exome sequencing in Brown Swiss and Holstein cattle identified a single non-synonymous causative mutation in the structural maintenance of chromosome 2 (SMC2) gene associated with a recessive fertility defects haplotype, Holstein Haplotype 3 (HH3) (McClure *et al*., 2014).

Genomic approaches to detect signatures of positive selection

Different approaches at the genome-wide level are commonly used to identify footprints of selection, particularly positive selection and, to a lesser extent, balancing selection (Biswas and Akey, 2006, Oleksyk *et al*., 2010). Some are implemented within populations (intra-population approaches), while others require the comparison of genome-wide data from two populations (inter-population approaches). They are also varied in terms of the type of genomic signals they investigate, as well as the age of selection they detect (Table 1.1).

Table 1.1: Examples of approaches used to identify genomic signatures of positive selection, their population level, and their estimated time of selection. *All estimates are for human lineages, assuming a generation interval of 25 years (adopted from Olyksyk *et al.*, 2010).

Type of genomic signal	**Genomic analysis**	**Population level**	**Estimated age of selection (generation)***
Increased function-altering mutation rate	ω = Dn/Ds (Neilsen and Yang, 1998)	Intra-population	> 40,000
Genetic diversity reduction (selective sweep)	ZHp (Rubin *et al.*, 2010), SNP heterozygosity (Olyksyk *et al.*, 2008)	Intra-population	< 8,000
Change in allele frequency spectrum	Tajima's D (Tajima, 1989), Fu and Li's D-test (Fu and Li, 1993), Fay and Wu's H-test (Fay and Wu, 2000)	Intra-population	< 8,000
Population genomic differentiation	FST (Wright, 1951)	Inter-population	< 3,000
Extended haplotype homozygosity (EHH)	LRH (Sabeti *et al.*, 2002), XP-EHH (Sabeti *et al.*, 2007), iHs (Voight *et al.*, 2006), Rsb (Tang *et al.*, 2007)	Intra-population and Inter-population	< 1,200

Fixation index (F_{ST})

This statistic was initially introduced by Sewall Wright alongside two other statistics, F_{IT} and F_{IS} (Wright, 1951). Wright's F_{ST}, which is a measure of the degree of genetic differentiation between two populations, is defined as "the correlation between gametes chosen randomly from within the same sub-population relative to the entire population". This is equivalent to a reduction in heterozygosity within subpopulations relative to the entire total population (Holsinger and Weir, 2009). This statistic can be calculated for each locus using its allele frequency and, hence, heterozygosity, based on the following equation: $(H_T - H_S) / H_T$, where H_T is the heterozygosity of the total population and H_S is the average heterozygosity of the two subpopulations.

Under the condition of differential selective pressures, two populations may favour alternative alleles at the loci under selection. This, in turn, will reduce the heterozygosity in the two populations and, hence, increase their genetic differentiation at these loci (high F_{ST}).

Dn/Ds (ω ratio)

The Dn/Ds ratio (ω ratio) is defined as the ratio of the number of non-synonymous substitution per non-synonymous site (Dn) to the number of synonymous substitution per synonymous site (Ds). If non-synonymous substitutions in a gene are subjected to positive selection and, hence, are not eliminated, the ω ratio will be greater than 1, i.e., Dn > Ds. Conversely, under purifying selection on deleterious non-synonymous substitutions, Dn will be lower than Ds and, hence, the ω ratio is less than 1. In neutrality, Dn and Ds are expected to be the same and the ω ratio is equal to 1 (Nielsen and Yang, 1998, Biswas and Akey, 2006).

Extended haplotype homozygosity (EHH)-based approaches

The term Extended Haplotype Homozygosity (EHH) was defined by Sabeti *et al*. (2002) as “the probability that two randomly chosen chromosomes carrying the core haplotype of interest are identical by descent for the entire interval from the core loci to the point x”. This can be observed when positive selection causes a rapid increase in the frequency of a favourable allele in a short time period, thereby leading to an unusually long haplotype.

Subsequently, Voight *et al*. (2006) developed a statistic based on EHH called the integrated haplotype score *iHS*. This SNP-based statistic compares the observed decay of EHH with genomic distance for the ancestral allele to that of the derived allele for a core SNP within a specific population.

The *iHS* algorithm lacks the power to detect a selective sweep when the selected allele is at a high frequency or complete fixation (Biswas and Akey, 2006, Tang *et al*., 2007, Gautier and Naves, 2011). To address this, the *Rsb* statistic (Tang *et al.*, 2007) was developed. This statistic compares the observed decay of EHH for a specific SNP (EHHS), rather than an allele, with genomic distance between two populations. EHHS is calculated by averaging the EHH for the two alternative alleles weighted by their squared allele frequencies.

∆*DAF* and ∆*AF*

∆*DAF* was first described by Grossman *et al.* (2010) to indicate whether there is a difference in the derived allele frequency of each genotyped SNP between two populations (∆*DAF*). This test has been combined with four other statistical approaches in a single composite test to identify signals and causal variants of positive selection in the human genome. In contrast to F_{ST} and *Rsb*, ∆*DAF* is superior in distinguishing the causative, favourable, variants in the positively selected regions rather than narrowing down the candidate region intervals (Grossman *et al.*, 2010). ∆*AF* is an extension of this method, which investigates the absolute allele frequency difference between two populations (Carneiro *et al.*, 2014).

Selective sweep analysis (*Hp*)

This type of analysis has been used to detect genomic signatures of selection in chickens (Rubin *et al.*, 2010), pigs (Rubin *et al.*, 2012) and cattle (Liao *et al.*, 2013) using full genome sequence data. The purpose of this analysis is to assess pooled heterozygosity (*Hp*) in pre-specified windows along the genome to identify regions with low heterozygosity relative to the entire genome. As positive selection usually leads to a reduction in genomic diversity, i.e., high homozygosity, these regions are likely to be indicative of selective sweeps.

Identifying genomic signatures of positive selection in livestock

Detecting signatures of positive selection in the genomes of different livestock species is considered to be one of the main goals in population genetics. Several studies have been published that address this issue intensively. In chicken, analysing pooled full genome sequences of domestic chicken populations has identified sweep regions harbouring genes possibly selected during domestication, e.g., thyroid stimulating hormone receptor (TSHR) (Rubin *et al*., 2010). Another study on broiler chicken lines using genome-wide SNP genotyping has identified regions showing signatures of positive selection that carry genes associated with abdominal fat deposition, e.g., retinoblastoma 1 (*RB1*) and Bardet-Biedl syndrome 7 (*BBS7*) (Zhang *et al*., 2012). Sweep regions linked to fat deposition have also been characterized in thin- and fat-tailed sheep breeds using genome-wide SNP genotyping (Moradi *et al*., 2012). In domestic pig populations, regions with genes associated with coat colour (*KIT*), muscle development (*MAPK1*), brain development and neuronal function (*PPP1R1B*) and body length (*PLAG1*) have been shown to be under strong positive selection (Amaral *et al*., 2011, Rubin *et al*., 2012). In Rubin *et al*. (2012), non-synonymous variants within the identified sweep regions have been proposed to be the candidate loci subjected to selection. However, these putative causative mutations require further investigation.

Signatures of positive selection in cattle

The identification of candidate signatures of selection in cattle has been conducted in both artificially selected commercial breeds (beef and dairy) and in indigenous cattle populations. Several genes associated with the immune system, male reproduction, and skin and hair development have been shown to be under positive selection in tropical-adapted cattle from West Africa (Gautier *et al*., 2009), Caribbean (Creole) cattle (Gautier and Naves, 2011), a synthetic European taurine x Asian zebu crossbreed, the Senepol cattle (Flori *et al*., 2012), and in Gir cattle (Liao *et al*., 2013). Sweep regions harbouring, or in vicinity to, genes associated with milk and meat production and composition have been identified in commercial dairy and beef cattle breeds (Hayes *et al*., 2009, Stella *et al*., 2010, Qanbari *et al*., 2010, Qanbari *et al*., 2014). Some of these sweep regions have also been found to overlap with previously identified milk-production QTL (Larkin *et al*., 2012). These signatures of selection have been defined using either full genome sequence data or genome-wide SNP genotyping.

The identified candidate regions might also be caused by other demographic factors that yield genomic patterns that are similar to those resulting from selection, e.g., genetic drift. Validating these regions in different livestock populations and using different statistical tests will reduce the probability of false positives.

Why is identifying genomic signatures of positive selection in indigenous African cattle important?

Indigenous African cattle show a high degree of adaptability to the African environment in terms of disease challenge, forage quality and thermal stress (Western and Finch, 1986, Latif *et al*., 1991b, Gaughan *et al*., 1999, Rege *et al*., 2001, Hanotte *et al*., 2010). These genetic adaptations are threatened by the introgression of exotic cattle to Africa from commercial breeds with the aim of increasing the short-term productivity of indigenous African cattle (Hanotte *et al*., 2010). The introduction of the European milk-producing Holstein-Friesian and Jersey cattle, which do not possess any adaptation to the African environment, has been reported in many parts of Africa, including Kenya (Weir *et al*., 2009) and Ethiopia (Haile *et al*., 2011). Based on genome-wide SNP genotyping, some EASZ cattle from Kenya have been shown to exhibit European genomic introgression into their genome (Mbole-Kariuki *et al*., 2014). This exotic introgression has a negative impact on the adaptability of African cattle to various disease challenges (Murray *et al*., 2013).

Defining candidate genome regions, i.e., those showing signatures of positive selection, associated with indigenous African cattle adaptations is an important first step to conserve and utilise these unique genetic resources. This will help livestock breeders to select cattle for breeding purposes aiming to improve their productivity while conserving their valuable adaptations.

Objectives

The main objectives of this thesis are to characterize the genome of indigenous EASZ cattle from Kenya and to identify candidate regions showing signatures of positive selection at the autosomal, X chromosome and mtDNA levels. For this purpose, we have analysed genome-wide SNP genotyping using commercially available SNP arrays (Illumina BovineSNP50 Genotyping BeadChip and Illumina BovineHD Genotyping BeadChip), full genome sequencing using the next-generation SOLiD platform and exome data.

Chapter Two is a first attempt to identify signatures of positive selection in the genome of EASZ. This was addressed using a low-density SNP array (Illumina BovineSNP50 Genotyping BeadChip) to genotype 425 EASZ samples. Genome-wide SNP data from different reference cattle populations (Holstein-Friesian, Jersey, N'Dama and Nellore cattle) were also used for this purpose. Different statistical approaches were implemented; namely, two EHH-based approaches (*iHS* and *Rsb*) and an inter-population F_{ST} analysis.

Chapter Three extends the preceding chapter with the implementation of a higher density SNP array (the Illumina BovineHD Genotyping BeadChip) to improve the EASZ genome representation. In this chapter, more indigenous East and West African zebu-taurine admixed cattle populations were included to validate the identified candidate regions in other admixed populations. Full genome and exome sequence data of 10 EASZ cattle were used in our analyses to help define the most interesting candidate regions and the putative causative genomic variants under selection, e.g., SNPs, indels and CNV.

Chapter Four aims to identify candidate regions with signatures of positive selection on the EASZ sex chromosome (BTA X). In this chapter, high-density SNP genotyping (using the Illumina BovineHD Genotyping BeadChip), the full chromosome SOLiD sequences and the exome sequence data were analysed.

Chapter Five investigates the mtDNA of EASZ cattle. Having defined the mtDNA haplogroup(s) that is affiliated with this indigenous cattle population using full mtDNA sequencing data, it aims to provide an alternative explanation to the classical male-mediated zebu introgression hypothesis to explain the unique presence of taurine mtDNA in African cattle. It also attempts to identify signatures of selection in the mtDNA of African cattle.

Chapter Six is a final conclusion of the whole thesis that summarises the main key outcomes from the previous chapters and their possible implications in cattle breeding programmes. Finally, possible future directions are considered to improve the conducted analyses and refine the resulting outputs.

References

ACHILLI, A., OLIVIERI, A., PELLECCHIA, M., UBOLDI, C., COLLI, L., AL-ZAHERY, N., ACCETTURO, M., PALA, M., HOOSHIAR KASHANI, B., PEREGO, U. A., BATTAGLIA, V., FORNARINO, S., KALAMATI, J., HOUSHMAND, M., NEGRINI, R., SEMINO, O., RICHARDS, M., MACAULAY, V., FERRETTI, L., BANDELT, H. J., AJMONE-MARSAN, P. & TORRONI, A. 2008. Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Curr. Biol.,* 18**,** R157-8.

ADAMETZ, L. 1920. *Herkunft und wanderungen der Hamiten erschlossen aus ihren haustierrassen,* California, University of California Libraries.

AMARAL, A. J., FERRETTI, L., MEGENS, H. J., CROOIJMANS, R. P., NIE, H., RAMOS-ONSINS, S. E., PEREZ-ENCISO, M., SCHOOK, L. B. & GROENEN, M. A. 2011. Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PLoS One,* 6**,** e14782.

BAIG, M., BEJA-PEREIRA, A., KULKARNI, K., FARAH, S. & LUIKART, G. 2005. Phylogeography and origin of Indian domestic cattle. *current science* 89**,** 38-40.

BAKER, C. M. & MANWELL, C. 1980. Chemical classification of cattle. 1. Breed groups. *Anim. Blood Groups Biochem. Genet.,* 11**,** 127-50.

BICKHART, D. M., HOU, Y., SCHROEDER, S. G., ALKAN, C., CARDONE, M. F., MATUKUMALLI, L. K., SONG, J., SCHNABEL, R. D., VENTURA, M., TAYLOR, J. F., GARCIA, J. F., VAN TASSELL, C. P., SONSTEGARD, T. S., EICHLER, E. E. & LIU, G. E. 2012. Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res.,* 22**,** 778-90.

BISWAS, S. & AKEY, J. M. 2006. Genomic insights into positive selection. *Trends Genet.,* 22**,** 437-46.

BONFIGLIO, S., GINJA, C., DE GAETANO, A., ACHILLI, A., OLIVIERI, A., COLLI, L., TESFAYE, K., AGHA, S. H., GAMA, L. T., CATTONARO, F., PENEDO, M. C., AJMONE-MARSAN, P., TORRONI, A. & FERRETTI, L. 2012. Origin and spread of Bos taurus: new clues from mitochondrial genomes belonging to haplogroup T1. *PLoS One,* 7**,** e38601.

BRADLEY, D. G., LOFTUS, R. T., CUNNINGHAM, P. & MACHUGH, D. E. 1998. Genetics and domestic cattle origins. *Evolutionary Anthropology,* 6**,** 79-86.

BRADLEY, D. G., MACHUGH, D. E., CUNNINGHAM, P. & LOFTUS, R. T. 1996. Mitochondrial diversity and the origins of African and European cattle. *PNAS,* 93**,** 5131-5.

BRADLEY, D. G., MACHUGH, D. E., LOFTUS, R. T., SOW, R. S., HOSTE, C. H. & CUNNINGHAM, E. P. 1994. Zebu-taurine variation in Y chromosomal DNA: a sensitive assay for genetic introgression in west African trypanotolerant cattle populations. *Anim. Genet.,* 25**,** 7-12.

CHEN, S., LIN, B. Z., BAIG, M., MITRA, B., LOPES, R. J., SANTOS, A. M., MAGEE, D. A., AZEVEDO, M., TARROSO, P., SASAZAKI, S., OSTROWSKI, S., MAHGOUB, O., CHAUDHURI, T. K., ZHANG, Y. P., COSTA, V., ROYO, L. J., GOYACHE, F., LUIKART, G., BOIVIN, N., FULLER, D. Q., MANNEN, H., BRADLEY, D. G. & BEJA-PEREIRA, A. 2010. Zebu cattle are an exclusive legacy of the South Asia neolithic. *Mol. Biol. Evol.,* 27**,** 1-6.

COSART, T., BEJA-PEREIRA, A., CHEN, S., NG, S. B., SHENDURE, J. & LUIKART, G. 2011. Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics,* 12**,** 347.

DECKER, J. E., MCKAY, S. D., ROLF, M. M., KIM, J., MOLINA ALCALA, A., SONSTEGARD, T. S., HANOTTE, O., GOTHERSTROM, A., SEABURY, C. M., PRAHARANI, L., BABAR, M. E., CORREIA DE ALMEIDA REGITANO, L., YILDIZ, M. A., HEATON, M. P., LIU, W. S., LEI, C. Z., REECY, J. M., SAIF-UR-REHMAN, M., SCHNABEL, R. D. & TAYLOR, J. F. 2014. Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. *PLoS Genet.,* 10**,** e1004254.

EPSTEIN, H. 1971. *The origin of the domestic animals of Africa*, Africana publishing corporation. Vol. 1.

EPSTEIN, H. & MASON, I. L. 1984. *in Evolution of domesticated animals,* New York, Longman Inc.

FAN, B., DU, Z., GORBACH, D. & ROTHSCHILD, M. 2010. Development and Application of High-density SNP Arrays in Genomic Studies of Domestic Animals. *Asian-Aust. J. Anim. Sci.,* 23**,** 833-847.

FAO 2007a. *Gridded livestock of the world,*by G.R.W. Wint and T.P. Robinson. Rome.

FAO 2007b. *The State of the World's Animal Genetic Resources for food and Agriculture,*edited by Barbara Rischkowsky and Dafydd Pilling. Rome.

FLORI, L., GONZATTI, M. I., THEVENON, S., CHANTAL, I., PINTO, J., BERTHIER, D., ASO, P. M. & GAUTIER, M. 2012. A quasi-exclusive European ancestry in the Senepol tropical cattle breed highlights the importance of the slick locus in tropical adaptation. *PLoS One,* 7**,** e36133.

FRISCH, J. E., DRINKWATER, R., HARRISON, B. & JOHNSON, S. 1997. Classification of the southern African sanga and east African shorthorned zebu. *Anim. Genet.,* 28**,** 77-83.

GAUGHAN, J. B., MADER, T. L., HOLT, S. M., JOSEY, M. J. & ROWAN, K. J. 1999. Heat tolerance of Boran and Tuli crossbred steers. *J. Anim. Sci.,* 77**,** 2398-405.

GAUTIER, M., FLORI, L., RIEBLER, A., JAFFREZIC, F., LALOE, D., GUT, I., MOAZAMI-GOUDARZI, K. & FOULLEY, J. L. 2009. A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. *BMC Genomics,* 10**,** 550.

GAUTIER, M. & NAVES, M. 2011. Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Mol. Ecol.,* 20**,** 3128-43.

GIBBS, R. A., TAYLOR, J. F., VAN TASSELL, C. P., BARENDSE, W., EVERSOLE, K. A., GILL, C. A., GREEN, R. D., HAMERNIK, D. L., KAPPES, S. M., LIEN, S., MATUKUMALLI, L. K., MCEWAN, J. C., NAZARETH, L. V., SCHNABEL, R. D., WEINSTOCK, G. M., WHEELER, D. A., AJMONE-MARSAN, P., BOETTCHER, P. J., CAETANO, A. R., GARCIA, J. F., HANOTTE, O., MARIANI, P., SKOW, L. C., SONSTEGARD, T. S., WILLIAMS, J. L., DIALLO, B., HAILEMARIAM, L., MARTINEZ, M. L., MORRIS, C. A., SILVA, L. O., SPELMAN, R. J., MULATU, W., ZHAO, K., ABBEY, C. A., AGABA, M., ARAUJO, F. R., BUNCH, R. J., BURTON, J., GORNI, C., OLIVIER, H., HARRISON, B. E., LUFF, B., MACHADO, M. A., MWAKAYA, J., PLASTOW, G., SIM, W., SMITH, T., THOMAS, M. B., VALENTINI, A., WILLIAMS, P., WOMACK, J., WOOLLIAMS, J. A., LIU, Y., QIN, X., WORLEY, K. C., GAO, C., JIANG, H., MOORE, S. S., REN, Y., SONG, X. Z., BUSTAMANTE, C. D., HERNANDEZ, R. D., MUZNY, D. M., PATIL, S., SAN LUCAS, A., FU, Q., KENT, M. P., VEGA, R., MATUKUMALLI, A., MCWILLIAM, S., SCLEP, G., BRYC, K., CHOI, J., GAO, H., GREFENSTETTE, J. J., MURDOCH, B., STELLA, A., VILLA-ANGULO, R., WRIGHT, M., AERTS, J., JANN, O., NEGRINI, R., GODDARD, M. E., HAYES, B. J., BRADLEY, D. G., BARBOSA DA SILVA, M., LAU, L. P., LIU, G. E., LYNN, D. J., PANZITTA, F. & DODDS, K. G. 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science,* 324**,** 528-32.

GIFFORD-GONZALEZ, D. & HANOTTE, O. 2011. Domesticating Animals in Africa: Implications of Genetic and Archaeological Findings. *Journal of World Prehistory* 24**,** 1-23.

GRIGSON, C. 1991. An African origin for African cattle? — some archaeological evidence. *African Archaeological Review,* 9**,** 119-144.

GROENEN, M. A., ARCHIBALD, A. L., UENISHI, H., TUGGLE, C. K., TAKEUCHI, Y., ROTHSCHILD, M. F., ROGEL-GAILLARD, C., PARK, C., MILAN, D., MEGENS, H. J., LI, S., LARKIN, D. M., KIM, H., FRANTZ, L. A., CACCAMO, M., AHN, H., AKEN, B. L., ANSELMO, A., ANTHON, C., AUVIL, L., BADAOUI, B., BEATTIE, C. W., BENDIXEN, C., BERMAN, D., BLECHA, F., BLOMBERG, J., BOLUND, L., BOSSE, M., BOTTI, S., BUJIE, Z., BYSTROM, M., CAPITANU, B., CARVALHO-SILVA, D., CHARDON, P., CHEN, C., CHENG, R., CHOI, S. H., CHOW, W., CLARK, R. C., CLEE, C., CROOIJMANS, R. P., DAWSON, H. D., DEHAIS, P., DE SAPIO, F., DIBBITS, B., DROU, N., DU, Z. Q., EVERSOLE, K., FADISTA, J., FAIRLEY, S., FARAUT, T., FAULKNER, G. J., FOWLER, K. E., FREDHOLM, M., FRITZ, E., GILBERT, J. G., GIUFFRA, E., GORODKIN, J., GRIFFIN, D. K., HARROW, J. L., HAYWARD, A., HOWE, K., HU, Z. L., HUMPHRAY, S. J., HUNT, T., HORNSHOJ, H., JEON, J. T., JERN, P., JONES, M., JURKA, J., KANAMORI, H., KAPETANOVIC, R., KIM, J., KIM, J. H., KIM, K. W., KIM, T. H., LARSON, G., LEE, K., LEE, K. T., LEGGETT, R., LEWIN, H. A., LI, Y., LIU, W., LOVELAND, J. E., LU, Y., LUNNEY, J. K., MA, J., MADSEN, O., MANN, K., MATTHEWS, L., MCLAREN, S., MOROZUMI, T., MURTAUGH, M. P., NARAYAN, J., NGUYEN, D. T., NI, P., OH, S. J., ONTERU, S., PANITZ, F., PARK, E. W., *et al*. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature,* 491**,** 393-8.

GROSSMAN, S. R., SHLYAKHTER, I., KARLSSON, E. K., BYRNE, E. H., MORALES, S., FRIEDEN, G., HOSTETTER, E., ANGELINO, E., GARBER, M., ZUK, O., LANDER, E. S., SCHAFFNER, S. F. & SABETI, P. C. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science,* 327**,** 883-6.

HAILE, A., WORKNEH, A., NOAH, K., TADELLE, D. & AZAGE, T. 2011. *Breeding strategy to improve Ethiopian Boran cattle for meat and milk production. IPMS (Improving Productivity and Market Success) of Ethiopian Farmers Project Working Paper 26. ,* Nairobi, Kenya, ILRI.

HANOTTE, O., BRADLEY, D. G., OCHIENG, J. W., VERJEE, Y., HILL, E. W. & REGE, J. E. 2002. African pastoralism: genetic imprints of origins and migrations. *Science,* 296**,** 336-9.

HANOTTE, O., DESSIE, T. & KEMP, S. 2010. Ecology. Time to tap Africa's livestock genomes. *Science,* 328**,** 1640-1.

HANOTTE, O., RONIN, Y., AGABA, M., NILSSON, P., GELHAUS, A., HORSTMANN, R., SUGIMOTO, Y., KEMP, S., GIBSON, J., KOROL, A., SOLLER, M. & TEALE, A. 2003. Mapping of quantitative trait loci controlling trypanotolerance in a cross of tolerant West African N'Dama and susceptible East African Boran cattle. *PNAS,* 100**,** 7443-8.

HANOTTE, O., TAWAH, C. L., BRADLEY, D. G., OKOMO, M., VERJEE, Y., OCHIENG, J. & REGE, J. E. 2000. Geographic distribution and frequency of a taurine Bos taurus and an indicine Bos indicus Y specific allele amongst sub-saharan African cattle breeds. *Mol. Ecol.,* 9**,** 387-96.

HANSEN, P. J. 2004. Physiological and cellular adaptations of zebu cattle to thermal stress. *Anim. Reprod. Sci.,* 82-83**,** 349-60.

HAYES, B. J., CHAMBERLAIN, A. J., MACEACHERN, S., SAVIN, K., MCPARTLAN, H., MACLEOD, I., SETHURAMAN, L. & GODDARD, M. E. 2009. A genome map of divergent artificial selection between *Bos taurus* dairy cattle and *Bos taurus* beef cattle. *Anim. Genet.,* 40**,** 176-84.

HIRANO, T., KOBAYASHI, N., MATSUHASHI, T., WATANABE, D., WATANABE, T., TAKASUGA, A., SUGIMOTO, M. & SUGIMOTO, Y. 2013. Mapping and exome sequencing identifies a mutation in the IARS gene as the cause of hereditary perinatal weak calf syndrome. *PLoS One,* 8**,** e64036.

HOLSINGER, K. E. & WEIR, B. S. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nat. Rev. Genet.,* 10**,** 639-50.

KIEFFER, A. M. & CARTWRIGHT, T. C. 1968. Sex chromosome polymorphism in domestic cattle. *Journal of Heredity* 59**,** 35-36.

LAI, S. J., LIU, Y. P., LIU, Y. X., LI, X. W. & YAO, Y. G. 2006. Genetic diversity and origin of Chinese cattle revealed by mtDNA D-loop sequence variation. *Mol. Phylogenet. Evol.,* 38**,** 146-54.

LARKIN, D. M., DAETWYLER, H. D., HERNANDEZ, A. G., WRIGHT, C. L., HETRICK, L. A., BOUCEK, L., BACHMAN, S. L., BAND, M. R., AKRAIKO, T. V., COHEN-ZINDER, M., THIMMAPURAM, J., MACLEOD, I. M., HARKINS, T. T., MCCAGUE, J. E., GODDARD, M. E., HAYES, B. J. & LEWIN, H. A. 2012. Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. *PNAS,* 109**,** 7693-8.

LATIF, A. A., NOKOE, S., PUNYUA, D. K. & CAPSTICK, P. B. 1991a. Tick infestations on Zebu cattle in western Kenya: quantitative assessment of host resistance. *J. Med. Entomol.,* 28**,** 122-6.

LATIF, A. A. & PEGRAM, R. G. 1992. Naturally acquired host resistance in tick control in Africa. *International Journal of Tropical Insect Science,* 13**,** 505-513.

LATIF, A. A., PUNYUA, D. K., NOKOE, S. & CAPSTICK, P. B. 1991b. Tick infestations on Zebu cattle in western Kenya: individual host variation. *J. Med. Entomol.,* 28**,** 114-21.

LIAO, X., PENG, F., FORNI, S., MCLAREN, D., PLASTOW, G. & STOTHARD, P. 2013. Whole genome sequencing of Gir cattle for identifying polymorphisms and loci under selection. *Genome,* 56**,** 592-8.

LOFTUS, R. T., MACHUGH, D. E., BRADLEY, D. G., SHARP, P. M. & CUNNINGHAM, P. 1994. Evidence for 2 independent domestications of cattle. *PNAS,* 91**,** 2757-2761.

MACHUGH, D. E., SHRIVER, M. D., LOFTUS, R. T., CUNNINGHAM, P. & BRADLEY, D. G. 1997. Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics,* 146**,** 1071-86.

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., SONSTEGARD, T. S. & VAN TASSELL, C. P. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One,* 4**,** e5350.

MBOLE-KARIUKI, M. N., SONSTEGARD, T., ORTH, A., THUMBI, S. M., BRONSVOORT, B. M., KIARA, H., TOYE, P., CONRADIE, I., JENNINGS, A., COETZER, K., WOOLHOUSE, M. E., HANOTTE, O. & TAPIO, M. 2014. Genome-wide analysis reveals the ancient and recent admixture history of East African Shorthorn Zebu from Western Kenya. *Heredity (Edinb),* 113**,** 297-305.

MCCLURE, M. C., BICKHART, D., NULL, D., VANRADEN, P., XU, L., WIGGANS, G., LIU, G., SCHROEDER, S., GLASSCOCK, J., ARMSTRONG, J., COLE, J. B., VAN TASSELL, C. P. & SONSTEGARD, T. S. 2014. Bovine exome sequence analysis and targeted SNP genotyping of recessive fertility defects BH1, HH2, and HH3 reveal a putative causative mutation in SMC2 for HH3. *PLoS One,* 9**,** e92769.

MEADOW, R. H. 1993. *Animal domestication in the Middle East: a revised view from the Eastern Margin. In: Harappan Civilisation,* New Delhi, Oxford & IBH.

MEGHEN, C., MACHUGH, D. E. & BRADLEY, D. G. 1994. Genetic characterization and West African cattle. *World animal review,* 78**,** 59-66.

METZKER, M. L. 2010. Sequencing technologies - the next generation. *Nat. Rev. Genet.,* 11**,** 31-46.

MORADI, M. H., NEJATI-JAVAREMI, A., MORADI-SHAHRBABAK, M., DODDS, K. G. & MCEWAN, J. C. 2012. Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. *BMC Genet.,* 13**,** 10.

MURRAY, G. G., WOOLHOUSE, M. E., TAPIO, M., MBOLE-KARIUKI, M. N., SONSTEGARD, T. S., THUMBI, S. M., JENNINGS, A. E., VAN WYK, I. C., CHASE-TOPPING, M., KIARA, H., TOYE, P., COETZER, K., DEC BRONSVOORT, B. M. & HANOTTE, O. 2013. Genetic susceptibility to infectious disease in East African Shorthorn Zebu: a genome-wide analysis of the effect of heterozygosity and exotic introgression. *BMC Evol. Biol.,* 13**,** 246-253.

MURRAY, M., MORRISON, W. I. & WHITELAW, D. D. 1982. Host susceptibility to African trypanosomiasis: trypanotolerance. *Advance in parasitology,* 21**,** 1-68.

NIELSEN, R. & YANG, Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics,* 148**,** 929-36.

OLEKSYK, T. K., SMITH, M. W. & O'BRIEN, S. J. 2010. Genome-wide scans for footprints of natural selection. *Philos Trans R Soc Lond B Biol Sci,* 365**,** 185-205.

PAYNE, W. J. A. 1991. Cattle genetic resources. *In:* HICKMAN, C. G. (ed.). Amsterdam: Elsevier.

PEREZ O'BRIEN, A. M., UTSUNOMIYA, Y. T., MESZAROS, G., BICKHART, D. M., LIU, G. E., VAN TASSELL, C. P., SONSTEGARD, T. S., DA SILVA, M. V., GARCIA, J. F. & SOLKNER, J. 2014. Assessing signatures of selection through variation in linkage disequilibrium between taurine and indicine cattle. *Genet. Sel. Evol.,* 46**,** 19.

PERKINS, D., JR. 1969. Fauna of Catal Huyuk: evidence for early cattle domestication in Anatolia. *Science,* 164**,** 177-9.

PORTER, V. 2002. *Mason's World Dictionary of Livestock Breeds, Types and Varieties* CABI.

PORTO-NETO, L. R., SONSTEGARD, T. S., LIU, G. E., BICKHART, D. M., DA SILVA, M. V., MACHADO, M. A., UTSUNOMIYA, Y. T., GARCIA, J. F., GONDRO, C. & VAN TASSELL, C. P. 2013. Genomic divergence of zebu and taurine cattle identified through high-density SNP genotyping. *BMC Genomics,* 14**,** 876.

QANBARI, S., PAUSCH, H., JANSEN, S., SOMEL, M., STROM, T. M., FRIES, R., NIELSEN, R. & SIMIANER, H. 2014. Classic selective sweeps revealed by massive sequencing in cattle. *PLoS Genet.,* 10**,** e1004148.

QANBARI, S., PIMENTEL, E. C., TETENS, J., THALLER, G., LICHTNER, P., SHARIFI, A. R. & SIMIANER, H. 2010. A genome-wide scan for signatures of recent selection in Holstein cattle. *Anim. Genet.,* 41**,** 377-89.

RACA, G., JACKSON, C., WARMAN, B., BAIR, T. & SCHIMMENTI, L. A. 2010. Next generation sequencing in research and diagnostics of ocular birth defects. *Mol. Genet. Metab.,* 100**,** 184-92.

RAMEY, H. R., DECKER, J. E., MCKAY, S. D., ROLF, M. M., SCHNABEL, R. D. & TAYLOR, J. F. 2013. Detection of selective sweeps in cattle using genome-wide SNP data. *BMC Genomics,* 14**,** 382.

REGE, J. & TAWAH, C. 1999. The state of African cattle genetic resources II. Geographical distribution, characteristics and uses of present-day breeds and strains. *Animal Genetic Resources Information,* 26**,** 1-25.

REGE, J. E. O., KAHI, A., M., O.-A., MWACHARO, J. & HANOTTE, O. 2001. *Zebu cattle of Kenya: Uses, performance, farmer preferences and measures of genetic diversity*. Nairobi, Kenya, International Livestock Reaserch Institute.

RINCON, G., WEBER, K. L., EENENNAAM, A. L., GOLDEN, B. L. & MEDRANO, J. F. 2011. Hot topic: performance of bovine high-density genotyping platforms in Holsteins and Jerseys. *J. Dairy Sci.,* 94**,** 6116-21.

RUBIN, C. J., MEGENS, H. J., MARTINEZ BARRIO, A., MAQBOOL, K., SAYYAB, S., SCHWOCHOW, D., WANG, C., CARLBORG, O., JERN, P., JORGENSEN, C. B., ARCHIBALD, A. L., FREDHOLM, M., GROENEN, M. A. & ANDERSSON, L. 2012. Strong signatures of selection in the domestic pig genome. *PNAS,* 109**,** 19529-36.

RUBIN, C. J., ZODY, M. C., ERIKSSON, J., MEADOWS, J. R., SHERWOOD, E., WEBSTER, M. T., JIANG, L., INGMAN, M., SHARPE, T., KA, S., HALLBOOK, F., BESNIER, F., CARLBORG, O., BED'HOM, B., TIXIER-BOICHARD, M., JENSEN, P., SIEGEL, P., LINDBLAD-TOH, K. & ANDERSSON, L. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature,* 464**,** 587-91.

SABETI, P. C., REICH, D. E., HIGGINS, J. M., LEVINE, H. Z., RICHTER, D. J., SCHAFFNER, S. F., GABRIEL, S. B., PLATKO, J. V., PATTERSON, N. J., MCDONALD, G. J., ACKERMAN, H. C., CAMPBELL, S. J., ALTSHULER, D., COOPER, R., KWIATKOWSKI, D., WARD, R. & LANDER, E. S. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature,* 419**,** 832-7.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *PNAS,* 74**,** 5463-7.

SINGLETON, A. B. 2011. Exome sequencing: a transformative technology. *Lancet Neurology,* 10**,** 942-946.

STELLA, A., AJMONE-MARSAN, P., LAZZARI, B. & BOETTCHER, P. 2010. Identification of selection signatures in cattle breeds selected for dairy production. *Genetics,* 185**,** 1451-61.

STOCK, F. & GIFFORD-GONZALEZ, D. 2013. Genetics and African Cattle Domestication. *Afr. Archaeol. Rev.,* 30**,** 51-72.

STRANZINGER, G., ELMIGER, B. & HETZEL, D. T. S. 1987. Cytogenetic studies on different cattle breeds in Australia. *Journal of Animal Breeding and Genetics,* 104**,** 231-234.

TANG, K., THORNTON, K. R. & STONEKING, M. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.,* 5**,** e171.

TEER, J. K. & MULLIKIN, J. C. 2010. Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.,* 19**,** R145-51.

THEILMANN, J. L., SKOW, L. C., BAKER, J. F. & WOMACK, J. E. 1989. Restriction fragment length polymorphisms for growth hormone, prolactin, osteonectin, alpha crystallin, gamma crystallin, fibronectin and 21-steroid hydroxylase in cattle. *Anim. Genet.,* 20**,** 257-66.

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**,** 1088-91.

UTSUNOMIYA, Y. T., PEREZ O'BRIEN, A. M., SONSTEGARD, T. S., VAN TASSELL, C. P., DO CARMO, A. S., MESZAROS, G., SOLKNER, J. & GARCIA, J. F. 2013. Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. *PLoS One,* 8**,** e64280.

VOIGHT, B. F., KUDARAVALLI, S., WEN, X. & PRITCHARD, J. K. 2006. A map of recent positive selection in the human genome. *PLoS Biol.,* 4**,** e72.

WEIR, A., NOTLEY, M. & KATUI-KATUA, M. 2009. Finnish Aid in Western Kenya. Impact and Lessons Learned. Evaluation report 2009:5. Hakapaino Oy: Helsinki: Ministry for Foreign Affairs of Finland.

WENDORF, F. A. S., R. 1998. Nabta Playa and Its Role in Northeastern African Prehistory. *journal of anthropological archaeology,* 17**,** 97-123.

WESTERN, D. & FINCH, V. 1986. Cattle and pastoralism: survival and production in arid lands. *Human Ecology,* 14**,** 77-94.

WRIGHT, S. 1951. The genetical structure of populations. *Annals of Eugenics,* 15**,** 323-354.

ZHANG, H., WANG, S. Z., WANG, Z. P., DA, Y., WANG, N., HU, X. X., ZHANG, Y. D., WANG, Y. X., LENG, L., TANG, Z. Q. & LI, H. 2012. A genome-wide scan of selective sweeps in two broiler chicken lines divergently selected for abdominal fat content. *BMC Genomics,* 13**,** 704.

Chapter two

Signatures of positive selection in East African shorthorn zebu: A genome-wide single nucleotide polymorphism analysis[1]

[1] The results presented in this chapter have been accepted for publication: BAHBAHANI, H., CLIFFORD, H., WRAGG, D., MBOLE-KARIUKI, M. N., VAN TASSELL, C. P., SONSTEGARD, T., WOOLHOUSE, W. & HANOTTE, O. Signatures of positive selection in East African Shorthorn Zebu: A genome-wide single nucleotide polymorphism analysis. *Scientific Reports*.*In press*.

Abstract

The small EASZ is a type of cattle populating East Africa. Following the introduction of the Asian zebu to the African continent through the Horn of Africa, historical crossbreeding with the local African taurine resulted in the formation of EASZ populations. These cattle demonstrate adaptive characteristics to the tropical sub-humid environment of East Africa, and their crossbreed nature (Asian zebu x African taurine) has recently been revealed based on genome-wide single nucleotide polymorphisms (SNP) data. The main purpose of this chapter is to scan the whole genome of EASZ cattle from Western Kenya with the Illumina BovineSNP50 BeadChip v.1 to identify genomic signatures of positive selection. Following two Extended Haplotype Homozygosity (EHH)-based analyses (intra-population *iHS* and inter-population *Rsb*), and a single F_{ST}-based inter-population approach, 24 genome regions were considered as possible candidates for positive selection. Five of these regions showed substantial deficiencies in the Asian zebu ancestry, while a single region exhibited a substantial excess in this ancestry, suggesting possible origins of the selected haplotypes. A total of 409 genes, including candidate genes associated with tropical adaptation (e.g., male reproduction-related, immunological-related genes, and heat shock proteins), and 340 bovine quantitative trait loci (QTL), including trypanotolerant QTL, were identified within the candidate regions. Several biological pathways related to signalling, immunity, growth and development were enriched in these genes, suggesting the possibility that the admixed genomic background acts as an additional selective pressure.

Introduction

The humped *Bos taurus indicus* (zebu or indicine cattle) and humpless *Bos taurus taurus* (taurine cattle) are two recognized cattle subspecies. They were independently domesticated from the extinct wild aurochs, *Bos primigenius*, about 10,000 years BP (Loftus *et al*., 1994, Bradley *et al*., 1996, Chen *et al*., 2010). Although, an African centre of domestication for taurine cattle has been proposed (Bradley *et al*., 1996), it is now widely thought that these cattle might be an exclusive legacy of the Near East centre of livestock domestication (Loftus *et al*., 1999, Troy *et al*., 2001, Achilli *et al*., 2008, Bonfiglio *et al*., 2012). Loftus *et al*. (1994) have suggested, based on mtDNA analysis, two separate centers of domestication for taurine and zebu cattle. Later, a mtDNA analysis by Chen *et al*. (2010) indicated that the Indus valley, specifically, is the more likely domestication centre of zebu cattle, excluding the previously proposed domestication centres, the Ganges regions and South India (Allchin, 1963, Fuller, 2006).

The first cattle present on the African continent were of the taurine types, with undisputed domestic cattle becoming present from as early as ~ 5,000 BC (Gifford-Gonzalez and Hanotte, 2011). The first evidence of humped cattle, based on tomb paintings of the XIIth Dynasty, dates from 2,000 BC in Egypt, but the main entry point of Asian zebu cattle was the Horn of Africa (Hanotte *et al*., 2002). This introduction was followed by inland dispersion to the western and southern parts of the continent, as well as by cross-breeding with the local African taurine (Hanotte *et al*., 2002). Two waves of zebu introgression have been revealed, with the more recent one, ~ 700 AD, having likely been favoured by the rinderpest epidemics of the late 19th century (Hanotte *et al*., 2002).

Microsatellite diversity studies have revealed a background of zebu and taurine genetic admixture in modern East African indigenous zebu cattle (MacHugh *et al*., 1997, Rege *et al*., 2001). Although 100% taurine at the mtDNA level, these cattle are predominantly zebu at the autosomal level (Bradley *et al*., 1996, MacHugh *et al*., 1997, Rege *et al*., 2001), and they are nearly exclusively zebu at the *Y* chromosome (Hanotte *et al*., 2000). More recently, Decker *et al*. (2014) have revealed, using genome-wide single nucleotide polymorphism (SNP) data, on average, a 70% zebu genomic background in East African indigenous zebu cattle, with a gradual decline in indicine ancestry from east to west and central to south.

Indigenous EASZ cattle are classified into small and large types. The small EASZ are the most common type of East African cattle. They are found throughout Ethiopia, Uganda and Kenya across the humid and sub-humid agro-ecological zones, whereas the large EASZ are generally restricted to drier areas (Rege *et al*., 2001). Recently, the genetic admixture of Kenyan small EASZ has been studied in detail through genome-wide SNP genotyping, revealing an admixture level of around 84% zebu and 16% African taurine (Mbole-Kariuki *et al*., 2014). Interestingly, this study also shows that the small EASZ is a stabilized population with an even proportion of zebu and taurine genomic backgrounds across animals.

Artificial selection for productivity traits, such as cattle body size, milk yields and carcass traits, is common in livestock (Flori *et al*., 2009, Qanbari *et al*., 2010, Utsunomiya *et al*., 2013, Rothammer *et al*., 2013, Perez O'Brien *et al*., 2014, Fan *et al*., 2014). However, in indigenous livestock from the tropics and at the smallholder farmer level, natural environmental pressures (natural selection) continue to play a major role in shaping their genomes (e.g., thermotolerance and disease resistance) (Rege *et al*., 2001, Gautier *et al*., 2009, Gautier and Naves, 2011, Flori *et al*., 2012, Flori *et al*., 2014).

For example, zebu cattle are thought to be better adapted to dry and hot environments than taurine cattle, specifically European cattle (Turner, 1980, Western and Finch, 1986, Chan *et al*., 2010). Studies have indicated that zebu have a greater tolerance to *Rhipicephalus microplus* tick-burden, an ectoparasite typically found in dry and wet tropical environments across the world (Estrada-Pena *et al*., 2006), than taurine cattle (O'Kelly and Spiers, 1976, Utech *et al*., 1978). Zebu cattle have also been found to be better able to thermoregulate their body temperature in heat stress conditions than taurine cattle (Gaughan *et al*., 1999, Hansen, 2004). This thermotolerance might explain their superior fertility, semen quality and growth rate under heat stress when compared with taurine (Cartwright, 1955, Lampkin and Kennedy, 1965, Skinner and Louw, 1966, Hansen, 2004). The adaptation of African taurine cattle to the sub-humid and humid agro-ecological zones of the continent is also well documented, particularly their resistance/tolerance to parasitic diseases, e.g., trypanotolerance (Mattioli *et al*., 1998, Naessens *et al*., 2002, Bahbahani and Hanotte, 2015).

Several studies, using genome-wide genetic markers and full genome sequencing data, have explored the genomes of horses, sheep, pigs, dogs, chickens and cattle to identify signatures of selection following domestication (Rubin *et al*., 2010, Kijas *et al*., 2012, Rubin *et al*., 2012, Axelsson *et al*., 2013, Wilkinson *et al*., 2013, Petersen *et al*., 2013, Ramey *et al*., 2013, Yang *et al*., 2014, Flori *et al*., 2014). For example, the genomes of different horse breeds, including thoroughbred racing horses, have been investigated using microsatellite and SNP markers. Loci associated with enhanced exercise-related physiology, muscle strength, sexual reproduction, coat colour and style of locomotion have been revealed to be under strong positive selection (Gu *et al*., 2009, Petersen *et al*., 2013). Studies of sheep breeds have also revealed that genome regions linked to coat pigmentation, skeletal morphology, growth rate and fat deposition are under positive selection (Kijas *et al*., 2012, Moradi *et al*., 2012). Genomic loci related to reproduction, brain development and immunity have shown signatures of positive selection in dogs (Akey *et al*., 2010, Olsson *et al*., 2011, Axelsson *et al*., 2013) and pigs (Amaral *et al*., 2011, Rubin *et al*., 2012, Wilkinson *et al*., 2013) as well.

Studies involving cattle have analysed different breeds, ranging from the tropically adapted to dairy and beef breeds. Several genes associated with the immune system, male reproduction and skin and hair structure have been shown to be under positive selection in tropically adapted cattle from West Africa (Gautier *et al*., 2009, Flori *et al*., 2014, Xu *et al*., 2015), Caribbean (Creole) cattle (Gautier and Naves, 2011) and in a synthetic European taurine x Asian zebu crossbreed - Senepol cattle (Flori *et al*., 2012). Studies on artificially selected dairy and beef cattle have revealed signatures of selection on genome regions carrying genes associated with milk yield and composition (Hayes *et al*., 2009, Qanbari *et al*., 2010, Mancini *et al*., 2014, Xu *et al*., 2015) and carcass production and quality (Rothammer *et al*., 2013, Kemper *et al*., 2014).

We report here the identification of signatures of positive selection in the genome of EASZ, an indigenous zebu-taurine admixed cattle population from East Africa, through three genome-wide SNP analyses. Candidate genome regions for signatures of positive selection were identified through the analysis of genetic differentiation between EASZ and four reference populations (Holstein-Friesian, Jersey, N'Dama and Nellore), as well as through the identification of regions showing extended haplotype homozygosity within EASZ, as well as between EASZ and the combined reference populations. These regions include candidate genes associated with innate and acquired immunity, sexual reproduction and skin and hair development.

Materials and Methods

SNP genotyping and quality control

Non-European taurine introgressed EASZ (n = 425) from 20 different randomly selected sublocations that cover four distinct ecological zones in the Western and Nyanza provinces of Kenya (de Clare Bronsvoort *et al*., 2013, Mbole-Kariuki *et al*., 2014) were genotyped using the Illumina BovineSNP50 BeadChip v.1 that includes 55,777 SNPs, covering the 29 bovine autosomes, the sex chromosome (BTA X) and three unassigned linkage groups (Matukumalli *et al*., 2009). Genotyped SNP data from the same SNP array for four reference cattle populations, Holstein-Friesian (n = 64), Jersey (n = 28), N'Dama (n = 25) and Nellore (n = 21), were obtained from the Bovine HapMap consortium (Gibbs *et al*., 2009). Analyses were conducted on autosomes and BTA X separately to avoid any potential bias resulting from differences in effective population size. SNPs on unassigned linkage groups were excluded, as they are not annotated in the bovine reference genome UMD3.1 (Elsik *et al*., 2009), leaving a total of 55,675 markers. Quality control (QC) analysis for the 54,334 autosomal and 1,341 BTA X markers was implemented through the *check.marker* function of the GenABEL package (Aulchenko *et al*., 2007) for R software version 2.15.1 (R Development Core Team, 2012). The QC criteria included a minor allele frequency (MAF) threshold of 0.5% and a SNP call rate threshold of 95%. A total of 2,433 autosomal and 47 BTA X SNPs failed to pass the minor allele frequency criterion only, whilst 1,180 autosomal and 21 BTA X SNPs only failed the call rate criterion. An additional 5,471 autosomal and 352 BTA X SNPs failed to pass both the MAF and the SNP call rate criteria. In total, 45,250 autosomal (mean gap size = 55 kb, median gap size = 40 kb, standard deviation (SD) = 53 kb) and 921 BTA X SNPs (mean gap size = 161 kb, median gap size = 76 kb, SD = 276 kb) were retained for downstream analyses.

Additional QC criteria included a minimum sample call rate of 95% and a pairwise identity-by-state (IBS) of less than 95%, with the lower call rate animal eliminated. From the autosomal SNPs, one EASZ sample was excluded for having a low call rate, whilst one EASZ and one Holstein-Friesian sample were excluded following the IBS criterion. For the BTA X, only the criterion of low call rate was applied, which excluded two EASZ samples, as possible duplicate samples had already been removed based on autosomes quality control steps.

Inter-population genome-wide F_{ST} analysis

Inter-population F_{ST} analyses, based on Wright's fixation index (Wright, 1951), were conducted between the EASZ and each continental reference (European (Holstein-Friesian and Jersey), African (N'Dama) and Asian (Nellore)) population. F_{ST} values (weighted by population sample sizes) were calculated on sliding windows of 10 SNPs, overlapping by 5 SNPs. The upper 0.2% and 3% of the distribution of F_{ST} values were arbitrarily chosen as thresholds for the autosomes and BTA X analyses, respectively, taking into account the difference (9,032 *versus* 184) in the number of windows analysed between the two datasets. A genome region was defined as being a candidate for selection if at least two overlapping windows passed the upper distribution threshold. If this criterion was met, the window with the highest F_{ST} value was used as the candidate region interval.

Extended Haplotype Homozygosity-derived statistics

The method of calculating Extended Haplotype Homozygosity (EHH) was first introduced by Sabeti *et al*. (2002), and defined as the probability that two randomly chosen chromosomes, both carrying the same allele at a given focal SNP (the core allele), are identical by descent (homozygous at all SNPs) from the focal SNP to a distance *x*. A high value for EHH indicates the presence of an allele that has risen to fixation faster than expected under neutral evolution, allowing less opportunity for recombination at adjacent alleles. Two EHH-derived statistics, the Integrated Haplotype Score (*iHS)* (Voight *et al*., 2006) and *Rsb* (Tang *et al*., 2007b), have been applied in this study to identify signatures of positive selection in the genome of EASZ .

As a prerequisite for these two statistics, haplotypes were reconstructed by phasing the genotyped SNPs *via fastPHASE* software version 1.4 (Scheet and Stephens, 2006) using K10 and T10 as in Utsunomiya *et al*. (2013) as criteria to reduce the computation time. Population label information was used to estimate the phased haplotypes population background.

Another prerequisite, specifically for the EHH-based *iHS* analysis, is the need to determine the derived and ancestral allele of each SNP. These were inferred in two ways: i) the ancestral allele was inferred as the most common allele within a dataset of 13 Bovinae species genotyped with the same SNP chip used by Decker *et al*. (2009); ii) for SNPs with no genotyping data available in Decker *et al*. (2009), the ancestral allele was inferred as the most common allele within our dataset, consistent with the observation that, in humans, the SNP allele with the higher frequency was likely to represent the ancestral allele (Hacia *et al*., 1999).

The *iHS* analysis starts by computing the integral of the observed decay of EHH, which is the area under the EHH curve plotted against the physical genomic position (Figure S2.1a), as one moves away from a core SNP for both the ancestral and derived alleles until it reaches an arbitrary value of 0.05. These integrals, which are summed over both directions from the core SNP, are called $\mathrm{iHH_A}$ and $\mathrm{iHH_D}$ for the ancestral and derived alleles, respectively. The natural log of the ratio ($\mathrm{iHH_A}/\mathrm{iHH_D}$) is standardised to generate an *iHS* value for each SNP, as described by Voight *et al.* (2006) (Equation 1). The expectation (*Ep*) and standard deviation (*SDp*) of ln($\mathrm{iHH_A}/\mathrm{iHH_D}$) were estimated from the empirical distribution at SNPs whose allele frequency *p* matches the frequency at the core SNP.

$$iHS = \frac{\ln\left(\frac{iHH_A}{iHH_D}\right) - E_p\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]}{\mathrm{SD}_p\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]} \qquad \text{(Equation 1)}$$

Large negative values indicate unusually long haplotypes carrying the derived allele, while large positive values point to long haplotypes carrying the ancestral allele relative to the whole genome. The analysis was implemented using the R package *rehh* (Gautier and Vitalis, 2012) on SNPs showing a within-population MAF ≥ 0.5%. This additional MAF criterion was used because the *iHS* algorithm is not optimal for calculating statistics on fixed allele SNPs (Voight *et al.*, 2006). As the *iHS* values followed a normal distribution (Figure S2.2a), a two-tailed Z-test was applied to identify statistically significant SNPs under selection with either an extended haplotype of ancestral or derived alleles. With *iHS* values already standardized, the results were transformed directly to $–\log_{10}(1\text{-}2|\Phi(iHS)\text{-}0.5|)$, where $\Phi(iHS)$ represents the Gaussian cumulative distribution function. The resultant values of this transformation may also be interpreted as $–\log_{10}$ (*P*-value), where the *P*-value is the two-sided value. A threshold of 4, equivalent to a *P*-value of 0.0001, was used to define a significant *iHS* value.

The *Rsb* statistic was applied to provide a pairwise comparison of EHH measures for each SNP between EASZ and each continental reference (European (Holstein-Friesian and Jersey), African (N’Dama) and Asian (Nellore)) population, as well as all the reference populations combined, using the R *rehh* package. In these analyses, the EHH for the two alleles of a SNP was averaged and weighted by their squared allele frequencies, which provided the site-specific EHH (EHHS). As with EHH, the observed decay of EHHS for each core SNP was integrated and summed over both directions in both populations (iES) (Figure S2.1b). The natural log of the ratio (iES_{pop1}/iES_{pop2}) was standardized to generate an *Rsb* value for each core SNP, as described by Tang *et al*. (2007b) (Equation 2).

$$Rsb = \frac{\ln(Rsb) - median(\ln(Rsb))}{SD(\ln(Rsb))} \qquad \text{(Equation 2)}$$

Positive results were representative of a high EHHS in the EASZ compared with the other populations, and they were subsequently applied to a one-tailed Z-test due to the normal distribution of the *Rsb* values (Figure S2.2b). This involved the direct transformation of the *Rsb* values to $–\log_{10}(1- \Phi(Rsb))$, where once again, the resultant values could be interpreted as $-\log_{10}$ (*P*-value), where the *P*-value is the one-sided value. As before, a significance threshold of 4 was used. The Z-test was not applied to BTA X *Rsb* values due to their non-normal distribution (Shapiro–Wilk test; *P*-value $< 2.2 \times 10^{-16}$) (Figure S2.2b ii). In both the *iHS* and *Rsb* analyses, a genome region was considered to be a candidate for selection if two SNPs separated by ≤ 1 Mb passed the threshold.

Functional characterization of the candidate regions

The level of linkage disequilibrium (LD) between pairs of markers was calculated across each chromosome by applying the r^2 statistic (Ardlie *et al*., 2002) to determine the maximum distance from the most significant SNP within a candidate region, as identified by the *iHS* and *Rsb* analyses, from which to retrieve candidate genes' "candidate region interval". This was implemented in R through the *r2fast* function of the GenABEL package. The EASZ genotype data was divided into 11 bins of different sizes, and pairwise r^2 values were calculated and averaged for each bin size (Figure S2.3). Based on the rate of change in mean r^2 binned over distance, a distance of 0.5 Mb in both directions from the most significant SNP was used for gene retrieval. Indeed, at larger distances, we reached the r^2 plateau. This extent of LD has been confirmed in eight cattle breeds (taurine and zebu) in a previous study (McKay *et al*., 2007). For the interpopulation F_{ST} analysis, genes were retrieved within the most significant window of the candidate regions. The

Ensembl Genes 73 database (Flicek *et al*., 2013) was used for gene retrieval, utilizing the *Bos taurus taurus* genome assembly UMD 3.1. The derived gene list was processed using the functional annotation tool implemented in *DAVID* Bioinformatics resources 6.7 to define enriched functional annotation clusters (Huang da *et al*., 2009a, Huang da *et al*., 2009b). As recommended by the software, an enrichment score of 1.3, equivalent to a Fisher exact test *P*-value of 0.05, was used as the threshold for the identification of enriched clusters.

A list of all the bovine Quantitative Trait Loci (QTL) and their coordinates were downloaded from the cattle QTL database (http://www.animalgenome.org/cgi-bin/QTLdb/BT/index). Using the *intersectBed* function from the *BedTools* software (Quinlan and Hall, 2010), the overlapping QTL for the candidate regions were obtained.

Estimating zebu and African taurine ancestry proportions on BTA X

Admixture analysis *via* a Bayesian clustering method implemented in STRUCTURE software (Pritchard *et al*., 2000) was conducted on the BTA X for the whole dataset. An admixed model with independent allele frequencies was run with a burn-in period of 25,000 iterations and 50,000 Markov Chain Monte Carlo steps for K=3.

Estimation of excesses-deficiencies in zebu ancestry at candidate regions

LAMP software version 2.4 (Sankararaman *et al*., 2008) was used to estimate the Asian zebu and African taurine ancestry proportions of the genotyped SNPs. A genome-wide autosomal zebu ancestry proportion of 0.84 and an African taurine ancestry proportion of 0.16 were used as the averaged admixture proportions α (Mbole-Kariuki *et al*., 2014). For the BTA X, zebu and African taurine ancestry proportions of 0.89 and 0.11, respectively, were used, as estimated by the STRUCTURE analysis. An estimated number of 500 generations was set for the admixture in light of our current knowledge of zebu arrival on the continent, assuming a generation time of six years (Keightley and Eyre-Walker, 2000). A uniform recombination rate of 1 cM = 1 Mb was set as a pre-requisite of LAMP. The average excess/deficiency in Asian zebu ancestry at each SNP (ΔAZ) was calculated by subtracting the average estimated Asian zebu ancestry of the SNP from the average estimated Asian zebu ancestry of all SNPs. This calculation was conducted for autosomal and BTA X SNPs separately. Because of possible inter-marker variation in the Asian zebu ancestry proportion caused by genetic drift (Long, 1991, Tang *et al.*, 2007a), we decided to consider the median of ΔAZ for an arbitrary 5-SNP window, two SNPs on each side of the *iHS* and *Rsb* most significant candidate SNP under selection, to represent the ΔAZ of the *iHS* and *Rsb* candidate regions. For the identified F_{ST} candidate regions, the median ΔAZs for the SNPs within these regions were considered.

Results

Candidate genome region identification

The F_{ST}-based analysis performed between EASZ and the different continental-specific reference populations identified 13 regions that might be subjected to diversifying selective pressures between EASZ and these reference populations: one on BTA 2; two on BTA 4; one on BTA 7; two on BTA 13; one on BTA 14, BTA 19, BTA 22 and BTA 24, and three on BTA X (Figure 2.1, Figure 2.2, Table 2.1 and Table S2.1).

The *iHS* analysis on EASZ indicated that three candidate regions present on BTA 5, 23 and 29 demonstrated long EHH (Figure 2.3 and Table 2.1). These regions contained SNPs with significantly differentiated EHH between the two alleles relative to the whole genome.

The *Rsb* analysis revealed eight genome regions with differential EHH between EASZ and the combined reference populations: One on BTA 3, two on BTA 5, one on BTA 11, three on BTA 12 and one on BTA 19 (Figure 2.4 and Table 2.1). These are possible regions of positive selection on the EASZ genome. Six of these eight candidate regions showed significant SNPs in the European taurine and/or African taurine pairwise *Rsb* analyses (Figure 2.4 and

Table S2.1). The most remarkable region detected by the *Rsb* analysis was a region on BTA 12, with the most significant SNP at base pair 29,217,254.

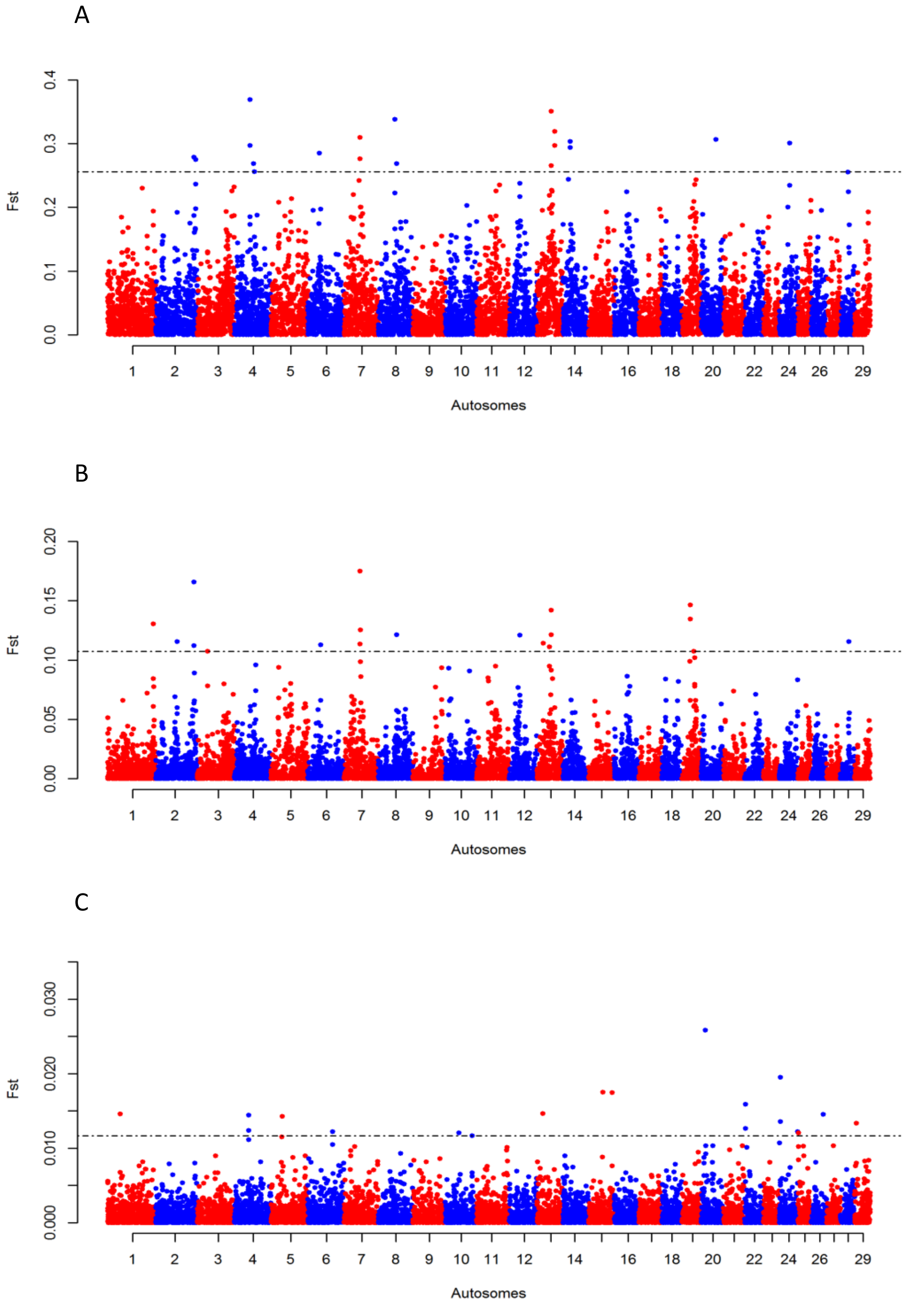

Figure 2.1: Manhattan plots of the pairwise genome-wide autosomal F_{ST} analyses. (A) EASZ with European taurine (Holstein-Friesian, Jersey), (B) EASZ with African taurine (N'Dama), and (C) EASZ with Asian zebu (Nellore). Significance thresholds (dashed lines) are set at the top 0.2% of the F_{ST} distribution.

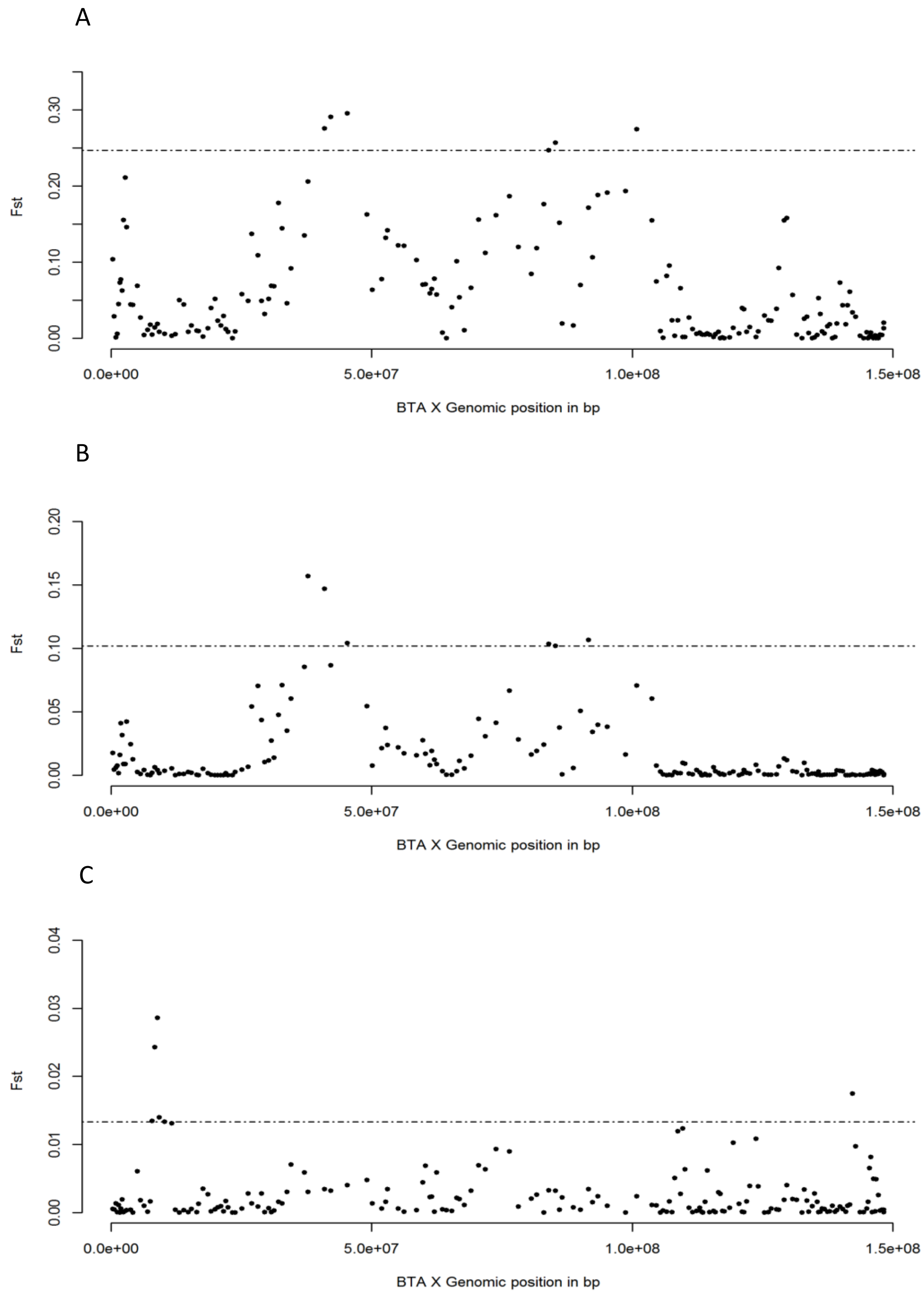

Figure 2.2: Manhattan plots of the pairwise BTA X F_{ST} analyses. (A) EASZ with European taurine (Holstein-Friesian, Jersey), (B) EASZ with African taurine (N'Dama), and (C) EASZ with Asian zebu (Nellore). Significance thresholds (dashed lines) are set at the top 3% of the F_{ST} distribution.

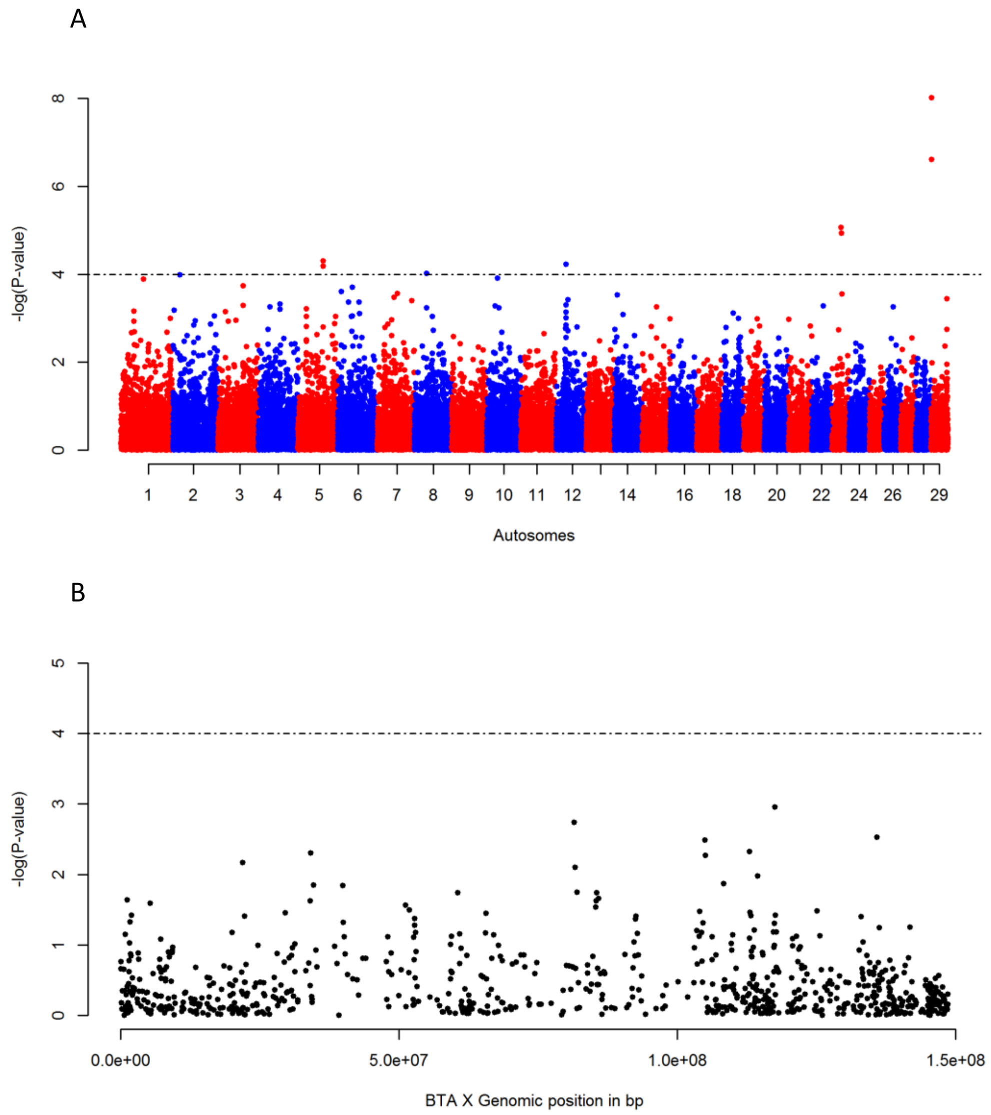

Figure 2.3: Manhattan plots of the genome-wide *iHS* analysis on EASZ, applied to a two-tailed Z-test. The plot in (A) shows the autosomal analysis, whilst (B) shows the BTA X analysis. Significance thresholds (dashed lines) are set at a $-\log_{10}$ (two-tailed *P*-value) of 4.

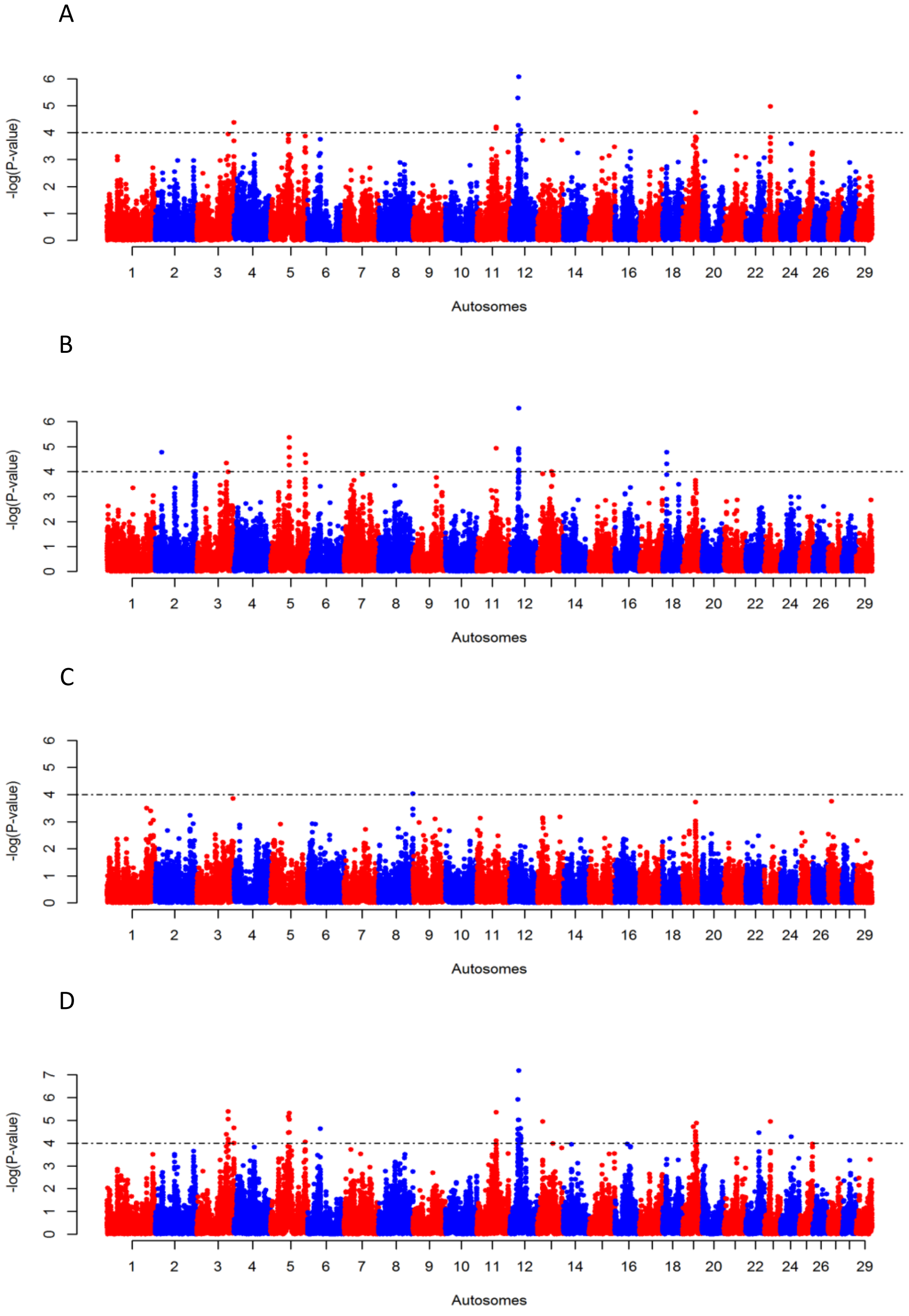

Figure 2.4: Manhattan plots of the genome-wide autosomal *Rsb* analyses. (A) EASZ with European taurine (Holstein-Friesian, Jersey), (B) EASZ with African taurine (N'Dama), (C) EASZ with Asian zebu (Nellore), and (D) EASZ with all reference populations (Holstein-Friesian, Jersey, N'Dama and Nellore) combined applied to one-tailed Z-tests. Significance thresholds (dashed lines) are set at a $-\log_{10}$ one-tailed *P*-value = 4.

Table 2.1: Candidate regions for signatures of positive selection in EASZ. Ref: previous studies on tropical adapted cattle and commercial breeds overlap with the identified candidate regions. * Commercial breeds studies. ** $-\log_{10}$(*P*-value) for *iHS* and *Rsb*; F_{ST} value of the window.

BTA	Position of the most significant SNPs/windows within candidate regions (bp)	Test	Value**	Ref
2	125,585,810 – 126,058,677	F_{ST}	0.17	Gautier and Naves (2011)
3	101,942,771	*Rsb*	5.06	Gautier *et al.* (2009) Gautier and Naves (2011)
4	47,195,467 - 47,539,595	F_{ST}	0.014	Gautier *et al.* (2009) Chan *et al.* (2010) Gautier and Naves (2011)
4	51,927,595 - 52,308,430	F_{ST}	0.37	Gautier *et al.* (2009) Chan *et al.* (2010) Larkin *et al*., (2012)* Qanbari *et al*., (2014)*
5	57,977,594	*Rsb*	5.15	Gautier *et al.* (2009) Flori *et al.* (2014) Kemper *et al.* (2014)*
5	60,556,520	*Rsb*	5.3	Gautier *et al.* (2009) Gautier and Naves (2011) Flori *et al.* (2014)
5	76,286,670	*iHS*	4.5	
7	52,224,595 - 52,720,797	F_{ST}	0.31	Gautier *et al.* (2009) Porto Neto *et al.* (2013)
11	62,629,106	*Rsb*	5.34	
12	27,181,474	*Rsb*	5.9	Gautier *et al.* (2009)
12	29,217,254	*Rsb*	7.17	Gautier *et al.* (2009) Gautier and Naves (2011) Porto Neto *et al.* (2013) Flori *et al.* (2014)
12	35,740,174	*Rsb*	4.64	
13	46,433,697-46,723,493	F_{ST}	0.35	Flori *et al.* (2014)
13	57,848,276-58,207,174	F_{ST}	0.32	Flori *et al.* (2014) Kemper *et al.* (2014)*
14	24,482,969-25,254,540	F_{ST}	0.29	Kemper *et al.* (2014)* Porto Neto *et al.* (2013)
19	27,369,763 - 27,763,447	F_{ST}	0.15	Gautier *et al.* (2009)
19	42,696,815	*Rsb*	4.51	Chan *et al.* (2010)
22	2,314,019 - 2,788,566	F_{ST}	0.02	
23	28,281,915	*iHS*	5.13	Gautier *et al.* (2009) Flori *et al.* (2014)
24	4,118,163 - 4,474,760	F_{ST}	0.02	
29	1,898,171	*iHS*	7.74	Flori *et al.* (2014)
X	8,582,093 - 9,248,137	F_{ST}	0.03	
X	39,942,044 - 43,999,854	F_{ST}	0.29	Porto Neto *et al.* (2013)
X	84,566,018 - 85,993,719	F_{ST}	0.26	Porto Neto *et al.* (2013)

Identification of candidate genes

Within the total 24 candidate region intervals obtained from the inter-population F_{ST} analysis and the two EHH-based analyses *(iHS* and *Rsb*), 192, 72 and 145 genes were identified, respectively (Table S2.2). Following DAVID analysis, these genes grouped into 53 functional term clusters (Table S2.3). Five of these clusters were significantly enriched relative to the whole bovine genome (Table 2.2). A total of 340 bovine QTL, e.g., tick resistance QTL and sperm motility QTL, were also found within these intervals (Table S2.4). Three of these QTL are linked to trypanotolerance in African cattle (Hanotte *et al.*, 2003) (Table S2.5). Other production traits QTL were also found, e.g., milk fat percentage and milk yield QTL.

Table 2.2: Significantly enriched functional term clusters of genes mapped within the candidate region intervals.

Functional term cluster	Enrichment score
Intermediate protein filaments and keratin	2.11
Immune response and antigen processing and presenting	1.88
Ribosome structure	1.62
Regulation of cell adhesion and mammary gland development	1.55
Regulation of steroid and growth hormone signalling pathways	1.53

Estimation of excesses-deficiencies of Asian zebu ancestry at the candidate regions – SNPs

LAMP software 2.4 (Sankararaman *et al.*, 2008) was used to investigate any excesses/deficiencies of Asian zebu ancestry (ΔAZ) in the candidate regions identified. This analysis, in addition to the F_{ST} analyses between EASZ and the putative parental populations, African taurine (N'Dama) and Asian zebu (Nellore), might help to define the ancestral origin of the selected haplotypes.

The average Asian zebu ancestries of all the SNPs over the EASZ samples were estimated to be 0.93 (SD = 0.07) and 0.95 (SD = 0.05) for the autosomes and BTA X, respectively. These estimations were higher than equivalent estimations (0.84 for autosomes calculated by Mbole-Kariuki *et al*. (2014) and 0.89 for BTA X) obtained using STRUCTURE software. This discrepancy might be due to the difference in the algorithms each software uses to estimate the admixture proportion. The mean ∆AZs for all the SNPs in EASZ were 0 for autosomes (SD = 0.07) and 0 for BTA X (SD = 0.05).

Candidate regions identified in this study exhibited a degree of balance in the number of regions with excesses and deficiencies in Asian zebu ancestry (10 regions showed excesses and fourteen regions showed deficiencies) (Table 2.3). Some of these regions showed substantial, ≥ +/- 1 SD from the mean, ∆AZ (five regions with substantial deficiencies and one region with a substantial excess). One of the five "zebu deficient" regions (BTA X 8.58–9.25 Mb) was also genetically differentiated from Nellore, supporting a possible taurine origin of the selected haplotype. These five regions contained genes involved in different biological pathways, such as the bovine acquired immune response (*IL-17D* and *IRAK1*) (O'Neill and Greene, 1998, Chang and Dong, 2011), mRNA processing regulation (*U5* and *U6*) (Lamond, 1991) and cell cycle regulation (*HECTD3*) (Yu *et al*., 2008).

Although the ancestral origins of most of these candidates were not clearly defined based on the F_{ST} analyses, seven regions were suggested to be of zebu origin, while four were of taurine origin (Table 2.3). The candidate region in BTA 7 (52.2–52.7 Mb), which showed genetic differentiation when EASZ was compared with N'Dama cattle but not to Nellore, also revealed a substantial excess of zebu ancestry, thereby supporting the possible zebu origin of the selected haplotype. Interestingly, an overlapping region (BTA 7: 50.95–53.75 Mb) was also found to be highly divergent between zebu and taurine cattle in a previous study (Porto-Neto *et al*., 2013). This region contains genes associated with critical biological pathways suggested to be under selection in tropical-adapted cattle (Gautier and Naves, 2011), such as protein folding and the heat shock response (*DNAJC7*) (Kampinga and Craig, 2010) and male reproduction and fertility (*SPATA24*) (Brancorsini *et al*., 2008).

Table 2.3: Defining the possible ancestral origin of the EASZ candidate regions, and their median excess/deficiency of Asian zebu ancestry (ΔAZ).

	EASZ most significant SNPs/windows with candidate regions	**EASZ *vs.* N'Dama**	**EASZ *vs.* Nellore**	**ancestral origin***	**median ΔAZ**
BTA	**Chromosomal position (bp)**				
2	125,585,810 – 126,058,677	125,585,810 - 126,058,677	_	zebu	-0.003
3	**101,942,771**	_	_	**unclear**	**-0.132**
4	47,195,467 - 47,539,595	_	47,195,467 - 47,539,595	taurine	0.016
4	51,927,595 - 52,308,430	_	_	unclear	-0.051
5	57,977,594	_	_	unclear	-0.003
5	60,556,520	_	_	unclear	0.049
5	76,286,670	_	_	unclear	0.043
7	**52,224,595 - 52,720,797**	**52,224,595 - 52,720,797**	**_**	**zebu**	**0.07**
11	62,629,106	_	_	unclear	0.008
12	**27,181,474**	**_**	**_**	**unclear**	**-0.188**
12	29,217,254	_	_	unclear	-0.038
12	**35,740,174**	**_**	**_**	**unclear**	**-0.084**
13	46,433,697 - 46,723,493	46,433,697 - 46,723,493	_	zebu	0.022
13	57,848,276 - 58,207,174	_	_	unclear	0.051
14	24,482,969- 25,254,540	_	_	unclear	-0.042
19	27,369,763 - 27,763,447	27,369,763 - 27,763,447	_	zebu	-0.012
19	42,696,815	42,660,383 - 43,068,079*	_	zebu	-0.004
22	2,314,019 - 2,788,566	_	2,314,019 - 2,788,566	taurine	0.035
23	28,281,915	_	_	unclear	-0.004
24	4,118,163 - 4,474,760	_	4,118,163 - 4,474,760	taurine	0.006
29	1,898,171	_	_	unclear	0.022
X	**8,582,093 - 9,248,137**	**_**	**8,582,093 - 9,248,137**	**taurine**	**-0.113**
X	**39,942,044 - 43,999,854**	**39,942,044 - 42,024,368**	**_**	**zebu**	**-0.050**
X	84,566,018 - 85,993,719	84,566,018 - 85,993,721	_	zebu	-0.034

Bold (deviation by more than +/- 1 SD from the genome-wide mean ΔAZ).
*Based on F_{ST} analyses

Discussion

Overlap between candidate selected regions

In this study, we used three types of analysis that can detect different patterns of signatures of selection. In contrast with the *iHS* statistic, which is optimal for intermediate frequency haplotypes within a population (Voight *et al*., 2006), *Rsb* and F_{ST} can identify fixed haplotypes under selection (Akey *et al*., 2002, Tang *et al*., 2007b). The choice of the EASZ population is particularly pertinent given the genetics of this population, which has been shaped by centuries of natural and human selection in East Africa (Rege *et al*., 2001). Additionally, the zebu-taurine admixture of EASZ offers an opportunity to address biological questions (e.g., the impact on fertility and development) related to the admixture of two cattle lineages that shared a common ancestor perhaps as long as 0.5 million years ago (Bradley *et al*., 1996, MacHugh *et al*., 1997).

Because we are interested in detecting signatures of selection on the admixed EASZ genome, we pooled all the non-admixed "pure" cattle breeds into a single reference population and compared them against EASZ in the *Rsb* analysis. As shown in Figure 2.4, the pooling approach increased the signals of selection in the *Rsb*-specific candidate regions compared with their signals in the non-pooled analyses. Such an empirical haplotype pooling approach has been suggested by Gautier and Naves (2011) to "smooth out" population-specific LD that probably results from genetic drift.

The lack of any overlap in the results between the *iHS* and *Rsb* analyses can be accounted for by (i) the reduced power of *iHS* to detect regions in which alleles have almost reached fixation, and (ii) that candidate genome regions identified by *iHS* within EASZ might also be subjected to selection in the reference population. The latter will affect the ability of the *Rsb* analysis to detect these regions. Although both the *Rsb* and F_{ST} statistics are optimal to detect signatures of selection reaching fixation, the absence of overlaps between them is likely a consequence of the selection time-scale, with *Rsb* being more suitable for detecting recent selection (Oleksyk *et al*., 2010).

A total of 18 candidate genome regions/SNPs were identified in this study that overlap with genome regions previously found to be under positive selection in other tropical-adapted cattle populations (Table 2.1). These studies focused on West African cattle (Gautier *et al*., 2009, Flori *et al*., 2014), admixed Creole cattle (Gautier and Naves, 2011) or zebu cattle (Chan *et al*., 2010, Porto-Neto *et al*., 2013). Additionally, four of the identified candidates overlap with regions under positive selection in beef and dairy cattle breeds, e.g., BTA 5: 52.8-64.75 Mb in Charolais cattle (Kemper *et al*., 2014), BTA 5: 40.65-61.8 Mb in Murrey Grey cattle (Kemper *et al*., 2014), BTA 13: 57.45-66.45 Mb in Shorthorn cattle (Kemper *et al*., 2014), BTA 4: 48.06-58.35 Mb in Holstein (Larkin *et al*., 2012) and BTA 4: 52.2-52.28 Mb in Fleckvieh cattle (Qanbari *et al*., 2014). This overlapping, in parallel with the intersected production trait QTL, might suggest prior selection for production traits in EASZ and/or LD between loci involved in different metabolic pathway.

Candidate regions: signature of selection or spurious effect?

It may be expected that the pattern of diversity in the genome of EASZ has been influenced by demographic events in a way similar to natural selection, i.e., a reduction in genetic diversity and the persistence of introgressed LD blocks, particularly in the context of our understanding of the history of African zebu cattle that likely has exhibited founding events and introgression. Distinguishing between genomic signatures of selection and/or demographics is difficult (Akey *et al*., 2002, Qanbari and Simianer, 2014).

It is also worth noting that the SNP chip used in this analysis includes mainly polymorphic variants in European taurine breeds (Matukumalli *et al*., 2009). This may have led to lower SNP diversity and, hence, increased haplotype homozygosity in zebu cattle in comparison with European taurine breeds. However, the overlap identified between our candidate selected genome regions and those identified in previous studies (Table 2.1) provides some support to the fact that these regions have been selected by natural selection, as opposed to resulting from genetic drift events.

Interestingly, we have successfully detected substantial excesses-deficiencies of Asian zebu ancestry at some of the identified candidate regions (Table 2.3), which allows one to infer the origin of the selected haplotypes. This, by itself, supports the role that natural selection played on these regions (Tang *et al*., 2007a). Given that the overall African taurine ancestry proportion in the EASZ genome is only ~ 0.07 (autosomes) and 0.04 (BTA X), as estimated by LAMP 2.4, the presence of "zebu deficient" regions might be the consequence of selection of taurine-specific alleles in the EASZ.

Both of the two approaches used in tackling the ancestral origin issue of the identified candidate regions, LAMP and the inter-population F_{ST} analysis, are based on analysing the allele frequencies of the genotyped SNPs. This may be a major issue for the Illumina Bovine SNP50 BeadChip v.1. The nature of ascertainment bias towards European taurine cattle in the SNP chip used may result in false inferences when attempting to address questions relating to ancestry outside of these breeds. Specifically, lower diversity in zebu than in taurine cattle might be expected, which can lead to lower genetic differentiation between EASZ and Nellore cattle than between EASZ and African taurine cattle (Matukumalli *et al*., 2009). Inferring the origin of the selected haplotypes in EASZ will be easier and more accurate once the indicine reference genome becomes available.

Potential selection pressures on the candidate regions

We stated in Table 2.2 that various functional term clusters "biological pathways" are enriched in the genes mapped within the candidate region intervals. Selection pressures on these biological pathways would be expected to contribute towards maintaining proper development and growth, as well as an effective immune system in EASZ, in accordance with African environmental constraints. However, it is also important to understand that the admixed genetic background of these cattle might have been another important selective pressure on the genome of this crossbred population.

A literature survey of the individual genes indicates that several biological pathways/functions could have been targeted by selection pressures associated with the African tropical environment and/or the admixed genetic background. Specific examples of these pathways include: innate and adaptive immunity, male reproduction and fertility, and heat stress tolerance (Table 2.4).

Table 2.4: Candidate genes considered in this chapter. (1) F_{ST}. (2) *iHS*. (3) *Rsb*.

Biological role	Candidate region interval	Gene ID	Gene name
Immunological-related	BTA 5: 60,056,520 - 61,056,520	*LTA4H*(3)	Leukotriene a-4 hydrolase
	BTA 5: 75,786,670 - 76,786,670	*RAC2*(2)	Ras-related c3 botulinum toxin substrate 2
		IL2RB(2)	Interleukin 2 receptor, beta
	BTA 11: 62,129,106 - 63,129,106	*PELI1*(3)	Pellino homolog 1
	BTA 12: 35,240,174 - 36,240,174	*IL17D*(3)	Interleukin 17D
	BTA 23: 27,781,915 - 28,781,915	*BOLA*(2)	MHC class I heavy chain isoform 1
		JSP.1(2)	MHC class I jsp.1
		BOLA-NC(2)	MHC, class I
	BTA X: 39,942,044 - 43,999,854	*IRAK1*(1)	Interleukin-1 receptor-associated kinase 1
	BTA X: 84,566,018 - 85,993,719	*IL2RG*(1)	Interleukin 2 receptor, gamma
Muscle function and structure-related	BTA 5: 57,477,594 - 58,477,594	*MYL6*(3)	Myosin light polypeptide 6
		MYL6B(3)	Myosin light chain 6b
	BTA 5: 75,786,670 - 76,786,670	*SYT10*(2)	Synaptotagmin 10
Hair structure	BTA 19: 42,196,815 - 43,196,815	*KRT*(3)	Members of keratin gene family
Male reproduction-related	BTA 5: 57,477,594 - 58,477,594	*OR*(3)	Members of olfactory receptor family
	BTA 7:52,224,595 - 52,720,797	*SPATA24*(1)	Spermatogenesis-associated protein 24
	BTA 12: 28,717,254 - 29,717,254	*RXFP2*(3)	Relaxin/insulin-like family peptide receptor 2
	BTA 19: 27,369,763 - 27,763,447	*SPEM1*(1)	Spermatid maturation 1
	BTA X: 84,566,018-85,993,719	*TEX11* (1)	Testis expressed 11
Heat stress response	BTA 2: 125,585,810 – 126,058,677	*BNAJC8*(1)	DnaJ (Hsp40) homolog, subfamily c, member 8
	BTA 5: 57,477,594 - 58,477,594	*DNAJC14*(3)	DnaJ (Hsp40) homolog, subfamily c, member 14
	BTA 7:52,224,595-52,720,797	*DNAJC18* (1)	DnaJ (Hsp40) homolog, subfamily c, member 18
	BTA 19: 42,196,815 - 43,196,815	*DNAJC7*(3)	DnaJ (Hsp40) homolog, subfamily c, member 7
		HSPB9(3)	Heat shock protein beta-9
	BTA 23: 27,781,915 - 28,781,915	*PPP1R10*(2)	Serine/threonine-protein phosphatase 1 regulatory subunit 10
Coat colour	BTA 5: 57,477,594 - 58,477,594	*PMEL17*(3)	Premelanosome protein

Indigenous cattle from the tropical environment are affected by several infectious diseases, e.g., babesiosis, tropical theileriosis, East Coast Fever (ECF) (uniquely in Africa) and anaplasmosis (Glass, 2001, Di Giulio *et al*., 2009, de Clare Bronsvoort *et al*., 2013, Thumbi *et al*., 2014). It may be expected that these cattle have developed tolerance or resistance to these diseases and their vectors, as shown by the improved resistance to tick burden observed in zebu cattle compared with European taurine cattle (O'Kelly and Spiers, 1976, Porto Neto *et al*., 2010, Bahbahani and Hanotte, 2015). For these reasons, genes with immunological-related functions are potential hotspots for positive selection in the genome of EASZ.

Several candidate immunological-related genes have been mapped on the candidate genome regions identified, including, for instance, the *Rac2* (Ras-related C3 botulinum toxin substrate 2) gene. *Rac2* has been found to be involved in the differentiation of myeloid precursor cells to a type of innate immune cell (neutrophils) (Didsbury *et al*., 1989). *IL2RB* (interleukin 2 receptor beta) and *IL2RG* (interleukin 2 receptor gamma) also fall into this category. The IL-2 receptor is a heterotrimeric molecule composed of alpha, beta and gamma subunits. This receptor is expressed on the membrane of T-lymphocytes, and interacts with the IL-2 cytokine to mediate its function in activating antigen-activated T-cells (Malek and Castro, 2010, Wuest *et al*., 2011). Several Major Histocompatibility Complex (MHC) class I genes have been identified in an *iHS*-specific candidate region on BTA 23. MHC class I molecules play a crucial role in initiating the immune response upon infection. These molecules are responsible for presenting antigen peptides to cytotoxic T-cells to induce their immunological response (Raghavan *et al*., 2008). A member of the interleukin 17 cytokine superfamily (*IL17D*) was identified on BTA 12, and although the function of IL17D is poorly understood (Chang and Dong, 2011), other members of this cytokine family were shown to play a role in inducing the secretion of proinflammatory cytokines, e.g., IL-6 and IL-8 (Fossiez *et al*., 1996), and in promoting a Th2-type response (Pappu *et al*., 2008). A gene identified in the candidate region on BTA 11, *PELI1*, expresses a scaffold protein associated with regulating the immune response induced by Toll-like receptor (TLR) and interleukin 1 receptor (IL-1R) signalling. This protein is specifically involved in transmitting the TLR/IL-1R intracellular signal to activate the NF-κB transcription factor and, hence, regulate the expression of proinflammatory genes (Schauvliege *et al.*, 2007). *IRAK1* (Interleukin-1 receptor-associated kinase 1) is another gene identified in one of the candidate regions on BTA X, and it is also involved in the IL-1R signalling pathway (O'Neill and Greene, 1998). Finally, the candidate gene leukotriene A-4 hydrolase (*LTA4H*) has been found to be associated with immune response regulation and inflammation in mammals (Thunnissen *et al.*, 2001).

Another category of genes mapped within the identified candidate genome regions relates to skeletal muscle function and structure. Two of the major purposes of zebu cattle in Kenya are ploughing and transportation (Rege *et al.*, 2001). Consequently, genes improving these activities might have been subjected to positive selection.

The synaptotagmin 10 gene (*SYT10*) has a role in controlling skeletal muscle contraction. The expressed protein of this gene is a synaptic vesicle-specific protein that has been found to be involved in regulating Ca^+-dependent neurotransmitter release (Littleton *et al.*, 1993). Littleton *et al*. (1993) reported, following a *Drosophila* knockout experiment, that this gene is important in coordinating muscle contraction by regulating the release of neurotransmitters in synapses. Members of the myosin light chain are also potential candidate genes due to their associated role in skeletal muscle function, specifically actin filament sliding velocity (Lowey *et al.*, 1993).

Compared with the relatively low temperature in temperate environments, the tropical conditions of the EASZ environment are characterized by a warm climate (20–23^{o}C) (Schmidt *et al.*, 1979) and high humidity (60–80%) (http://www.weather-and-climate.com/average-monthly-Rainfall-Temperature-Sunshine-in-Kenya). Previous studies in zebu and taurine cattle have confirmed the expected tolerance (thermotolerance) to heat stress of zebu cattle, which may be characterized by better growth and reproduction rates in these harsh conditions compared with European taurine cattle (Cartwright, 1955, Lampkin and Kennedy, 1965, Skinner and Louw, 1966, Gaughan *et al*., 1999, Hansen, 2004).

Members of the heat shock protein family were identified in both inter-population F_{ST} and EHH-based *Rsb* approaches. This gene family plays a critical role in maintaining protein folding and structure under stress (Parsell and Lindquist, 1994, Coleman *et al*., 1995). Members of the DnaJ family (*DNAJC7*, *DNAJC 8*, *DNAJC 14* and *DNAJC18*), which act as cofactors for the heat shock protein 70 (Hsp70) (Kampinga and Craig, 2010), were also found within the candidate regions identified in this study.

Serine/threonine-protein phosphatase 1 (PPP1) regulatory subunit 10 (*PPP1R10*) is another example of a stress response protein. A previous *in vitro* study by Shi and Manley (2007) has indicated a role of PPP1 in the heat shock response. This study deduced that PPP1 dephosphorylates and, hence, activates a splicing repressor (SRp38) in response to heat shock. Gautier and Naves (2011) detected another PPP1 regulatory subunit (PPP1R8) in a positively-selected genome region in Creole cattle.

Previous studies on zebu cows and bulls have demonstrated superior fertility and semen quality compared with European taurine breeds under similar heat stress conditions (Lampkin and Kennedy, 1965, Skinner and Louw, 1966, Hansen, 2004). *SPATA24* and *SPEM1*, which are mapped within the candidate regions in BTA 7 and BTA 19, play critical roles during spermatogenesis by mediating chromatin remodelling and cytoplasm removal in developing sperm (Zheng *et al*., 2007, Brancorsini *et al*., 2008).

An important gene located in the most significant *Rsb* candidate genome region in BTA 12 is the relaxin/insulin-like family peptide receptor 2 (*RXFP2*). The role of this gene in testicular descent development has been demonstrated to impact male fertility (Gorlov *et al*., 2002, Agoulnik, 2007, Feng *et al*., 2009). Because testicular descent is an important adaptation to maintaining proper spermatogenesis when the core body temperature reaches 34–35^{o}C (Park *et al*., 2008), selection on *RXFP2* may, therefore, be of importance for the adaptation of EASZ to warm conditions, thereby allowing EASZ to maintain normal male reproductive function. Interestingly, the genome region harbouring *RXFP2* has also been shown to be under positive selection in tropically adapted Creole cattle (Gautier and Naves, 2011) and West African admixed Borgou cattle (Flori *et al.*, 2014), as well as under diversifying selection between zebu and taurine cattle (Porto-Neto *et al.*, 2013). Recently, this gene has been linked to reproductive success and survival rate in a Soay sheep population (Johnston *et al.*, 2013).

The olfactory receptor family genes identified in the candidate region (BTA 5: 57,977,594 bp) can be classified as reproduction-related genes. Members of this gene family have been found to be expressed in human and dog testes (Parmentier *et al.*, 1992), and more specifically in mature sperm cells (Vanderhaeghen *et al.*, 1993, Spehr *et al.*, 2003). It has been suggested that this receptor family plays a role during the guidance of sperm to the oocyte during fertilization *via* an interaction with various chemoattractants secreted by the oocyte-cumulus cells complex in a phenomenon called "chemotaxis" (Spehr *et al.*, 2003, Fukuda *et al.*, 2004, Guidobaldi *et al.*, 2012).

Several members of the keratin gene family have been found in a candidate *Rsb* genome region interval on BTA 19. Keratin is the key protein involved in the process of regulating skin pigmentation and hair follicle growth (Gu and Coulombe, 2007). Moreover, *PMEL17* was identified in a candidate region on BTA 5 and it has a role in regulating coat colour *via* eumelanin formation (Theos *et al.*, 2005, McGlinchey *et al.*, 2009). Variants in this gene have been linked to the silver coat colour phenotype in horses (Brunberg *et al.*, 2006) and hypopigmentation in chickens (Kerje *et al.*, 2004). This gene has also been considered to be a candidate for positive selection in West African admixed Borgou cattle (Flori *et al.*, 2014). In a recent study by Porto-Neto *et al.* (2014), a large overlapping interval on BTA 5 (20–60 Mb) in Brahman and tropical composite cattle has been associated with several traits, such as coat colour, penile sheath and parasite resistance.

Hair colour and structure are associated with tick-resistance and thermotolerance in cattle. Martinez *et al.* (2006) stated that short-straight hair and light coat colour conferred more tick-burden resistance in Gir x Holstein cattle than long hair and dark coats. In another study, short-smooth hair was also linked to thermoregulation in Holstein cattle under heat stress conditions (Dikmen *et al*., 2008). Tropical-adapted cattle, e.g., Senepol, are characterized by the short sleek hair coat “Slick phenotype” to maintain low body temperature under heat stress (Hammond *et al*., 1998, Olson *et al*., 2003). Given that EASZ cattle are mainly have brown, short hair coats (Mbole-Kariuki, 2012), these genes might be considered to be targets of selection to confer some level of resistance to tick infestation and thermotolerance.

Trypanotolerance QTL

Interestingly, the tolerant alleles of the trypanotolerance QTL identified on BTA 7 are of taurine origin, whilst the parasite detection rate QTL on BTA 13 was found to be of zebu origin. The remaining trypanotolerant QTL on BTA 13 and BTA 29 demonstrate an overdominance inheritance mode (Hanotte *et al*., 2003). This explains the adaptive role of the zebu-taurine admixture in EASZ. It is likely that a degree of trypanotolerance in EASZ is conferred by these QTL, given that other East African cattle have also been shown to have the same adaptation (e.g., Orma Boran and Mursi cattle) (Dolan, 1987, Mwangi *et al*., 1993, Bahbahani and Hanotte, 2015). It has been shown that the introgression of the trypanotolerance QTL from African N’Dama cattle into Kenyan Boran cattle increased the level of trypanotolerance in this zebu cattle population (Orenge *et al*., 2012).

Conclusion

In this chapter, the whole genome of an EASZ cattle population has been explored to define regions harbouring signatures of positive selection using the low-density Illumina BovineSNP50 BeadChip v.1. After conducting three different signature selection analyses using more than 40,000 genome-wide

SNP genotypes, 24 genome regions were defined to be candidates for positive selection on the genome of EASZ.

Because of the confounding effect of natural demographic events in the EASZ genome diversity pattern, it is important to validate the role of natural selection at these candidate regions in other admixed cattle populations living under similar environmental conditions. Moreover, the ascertainment bias and low genome coverage issues associated with this SNP chip are additional directions that require improvement to fine map these selection signatures. All of these issues have been considered more carefully in the next two chapters of this thesis.

References

ACHILLI, A., OLIVIERI, A., PELLECCHIA, M., UBOLDI, C., COLLI, L., AL-ZAHERY, N., ACCETTURO, M., PALA, M., HOOSHIAR KASHANI, B., PEREGO, U. A., BATTAGLIA, V., FORNARINO, S., KALAMATI, J., HOUSHMAND, M., NEGRINI, R., SEMINO, O., RICHARDS, M., MACAULAY, V., FERRETTI, L., BANDELT, H. J., AJMONE-MARSAN, P. & TORRONI, A. 2008. Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Curr. Biol.,* 18**,** R157-8.

AGOULNIK, A. I. 2007. Relaxin and related peptides in male reproduction. *Adv. Exp. Med. Biol.,* 612**,** 49-64.

AKEY, J. M., RUHE, A. L., AKEY, D. T., WONG, A. K., CONNELLY, C. F., MADEOY, J., NICHOLAS, T. J. & NEFF, M. W. 2010. Tracking footprints of artificial selection in the dog genome. *PNAS,* 107**,** 1160-5.

AKEY, J. M., ZHANG, G., ZHANG, K., JIN, L. & SHRIVER, M. D. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.,* 12**,** 1805-14.

ALLCHIN, B. 1963. *Neolithic cattle-keepers of South India,* Cambridge, Cambridge University Press.

AMARAL, A. J., FERRETTI, L., MEGENS, H. J., CROOIJMANS, R. P., NIE, H., RAMOS-ONSINS, S. E., PEREZ-ENCISO, M., SCHOOK, L. B. & GROENEN, M. A. 2011. Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PLoS One,* 6**,** e14782.

ARDLIE, K. G., KRUGLYAK, L. & SEIELSTAD, M. 2002. Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.,* 3**,** 299-309.

AULCHENKO, Y. S., RIPKE, S., ISAACS, A. & VAN DUIJN, C. M. 2007. GenABEL: an R library for genome-wide association analysis. *Bioinformatics,* 23**,** 1294-6.

AXELSSON, E., RATNAKUMAR, A., ARENDT, M. L., MAQBOOL, K., WEBSTER, M. T., PERLOSKI, M., LIBERG, O., ARNEMO, J. M., HEDHAMMAR, A. & LINDBLAD-TOH, K. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature,* 459**,** 360-4.

BAHBAHANI, H. & HANOTTE, O. 2015. Genetic resistance: tolerance to vector-borne diseases, prospect and challenges of genomics. *OIE Scientific and Technical Review,* 34**,** 185-97.

BONFIGLIO, S., GINJA, C., DE GAETANO, A., ACHILLI, A., OLIVIERI, A., COLLI, L., TESFAYE, K., AGHA, S. H., GAMA, L. T., CATTONARO, F., PENEDO, M. C., AJMONE-MARSAN, P., TORRONI, A. & FERRETTI, L. 2012. Origin and spread of *Bos taurus*: new clues from mitochondrial genomes belonging to haplogroup T1. *PLoS One,* 7**,** e38601.

BRADLEY, D. G., MACHUGH, D. E., CUNNINGHAM, P. & LOFTUS, R. T. 1996. Mitochondrial diversity and the origins of African and European cattle. *PNAS,* 93**,** 5131-5.

BRANCORSINI, S., DAVIDSON, I. & SASSONE-CORSI, P. 2008. TIPT, a male germ cell-specific partner of TRF2, is chromatin-associated and interacts with HP1. *Cell Cycle,* 7**,** 1415-22.

BRUNBERG, E., ANDERSSON, L., COTHRAN, G., SANDBERG, K., MIKKO, S. & LINDGREN, G. 2006. A missense mutation in PMEL17 is associated with the Silver coat color in the horse. *BMC Genet.,* 7**,** 46-56.

CARTWRIGHT, T. C. 1955. Responses of beef cattle to high ambient temperatures. *J. Anim. Sci.,* 14**,** 350-362.

CHAN, E. K., NAGARAJ, S. H. & REVERTER, A. 2010. The evolution of tropical adaptation: comparing taurine and zebu cattle. *Anim. Genet.,* 41**,** 467-77.

CHANG, S. H. & DONG, C. 2011. Signaling of interleukin-17 family cytokines in immunity and inflammation. *Cell Signal.,* 23**,** 1069-75.

CHEN, S., LIN, B. Z., BAIG, M., MITRA, B., LOPES, R. J., SANTOS, A. M., MAGEE, D. A., AZEVEDO, M., TARROSO, P., SASAZAKI, S., OSTROWSKI, S., MAHGOUB, O., CHAUDHURI, T. K., ZHANG, Y. P., COSTA, V., ROYO, L. J., GOYACHE, F., LUIKART, G., BOIVIN, N., FULLER, D. Q., MANNEN, H., BRADLEY, D. G. & BEJA-PEREIRA, A. 2010. Zebu cattle are an exclusive legacy of the South Asia neolithic. *Mol. Biol. Evol.,* 27**,** 1-6.

COLEMAN, J. S., HECKATHORN, S. A. & HALLBERG, R. L. 1995. Heat-shock proteins and thermotolerance: linking molecular and ecological perspectives. *Trends Ecol. Evol.,* 10**,** 305-6.

DE CLARE BRONSVOORT, B. M., THUMBI, S. M., POOLE, E. J., KIARA, H., AUGUET, O. T., HANDEL, I. G., JENNINGS, A., CONRADIE, I., MBOLE-KARIUKI, M. N., TOYE, P. G., HANOTTE, O., COETZER, J. A. & WOOLHOUSE, M. E. 2013. Design and descriptive epidemiology of the Infectious Diseases of East African Livestock (IDEAL) project, a longitudinal calf cohort study in western Kenya. *BMC Vet. Res.,* 9**,** 171-192.

DECKER, J. E., MCKAY, S. D., ROLF, M. M., KIM, J., MOLINA ALCALA, A., SONSTEGARD, T. S., HANOTTE, O., GOTHERSTROM, A., SEABURY, C. M., PRAHARANI, L., BABAR, M. E., CORREIA DE ALMEIDA REGITANO, L., YILDIZ, M. A., HEATON, M. P., LIU, W. S., LEI, C. Z., REECY, J. M., SAIF-UR-REHMAN, M., SCHNABEL, R. D. & TAYLOR, J. F. 2014. Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. *PLoS Genet.,* 10**,** e1004254.

DECKER, J. E., PIRES, J. C., CONANT, G. C., MCKAY, S. D., HEATON, M. P., CHEN, K., COOPER, A., VILKKI, J., SEABURY, C. M., CAETANO, A. R., JOHNSON, G. S., BRENNEMAN, R. A., HANOTTE, O., EGGERT, L. S., WIENER, P., KIM, J. J., KIM, K. S., SONSTEGARD, T. S., VAN TASSELL, C. P., NEIBERGS, H. L., MCEWAN, J. C., BRAUNING, R., COUTINHO, L. L., BABAR, M. E., WILSON, G. A., MCCLURE, M. C., ROLF, M. M., KIM, J., SCHNABEL, R. D. & TAYLOR, J. F. 2009. Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *PNAS,* 106**,** 18644-9.

DI GIULIO, G., LYNEN, G., MORZARIA, S., OURA, C. & BISHOP, R. 2009. Live immunization against East Coast fever--current status. *Trends Parasitol.,* 25**,** 85-92.

DIDSBURY, J., WEBER, R. F., BOKOCH, G. M., EVANS, T. & SNYDERMAN, R. 1989. rac, a novel ras-related family of proteins that are botulinum toxin substrates. *J. Biol. Chem.,* 264**,** 16378-82.

DIKMEN, S., ALAVA, E., PONTES, E., FEAR, J. M., DIKMEN, B. Y., OLSON, T. A. & HANSEN, P. J. 2008. Differences in thermoregulatory ability between slick-haired and wild-type lactating Holstein cows in response to acute heat stress. *J. Dairy Sci.,* 91**,** 3395-402.

DOLAN, R. B. 1987. Genetics and trypanotolerance. *Parasitol. Today,* 3**,** 137-43.

ESTRADA-PENA, A., BOUATTOUR, A., CAMICAS, J. L., GUGLIELMONE, A., HORAK, I., JONGEJAN, F., LATIF, A., PEGRAM, R. & WALKER, A. R. 2006. The known distribution and ecological preferences of the tick subgenus Boophilus (Acari: Ixodidae) in Africa and Latin America. *Exp. Appl. Acarol.,* 38**,** 219-35.

FAN, H., WU, Y., QI, X., ZHANG, J., LI, J., GAO, X., ZHANG, L. & GAO, H. 2014. Genome-wide detection of selective signatures in Simmental cattle. *J. Appl. Genet.*, 55**,** 343-51.

FENG, S., FERLIN, A., TRUONG, A., BATHGATE, R., WADE, J. D., CORBETT, S., HAN, S., TANNOUR-LOUET, M., LAMB, D. J., FORESTA, C. & AGOULNIK, A. I. 2009. INSL3/RXFP2 signaling in testicular descent. *Ann. N. Y. Acad. Sci.,* 1160**,** 197-204.

FLICEK, P., AHMED, I., AMODE, M. R., BARRELL, D., BEAL, K., BRENT, S., CARVALHO-SILVA, D., CLAPHAM, P., COATES, G., FAIRLEY, S., FITZGERALD, S., GIL, L., GARCIA-GIRON, C., GORDON, L., HOURLIER, T., HUNT, S., JUETTEMANN, T., KAHARI, A. K., KEENAN, S., KOMOROWSKA, M., KULESHA, E., LONGDEN, I., MAUREL, T., MCLAREN, W. M., MUFFATO, M., NAG, R., OVERDUIN, B., PIGNATELLI, M., PRITCHARD, B., PRITCHARD, E., RIAT, H. S., RITCHIE, G. R., RUFFIER, M., SCHUSTER, M., SHEPPARD, D., SOBRAL, D., TAYLOR, K., THORMANN, A., TREVANION, S., WHITE, S., WILDER, S. P., AKEN, B. L., BIRNEY, E., CUNNINGHAM, F., DUNHAM, I., HARROW, J., HERRERO, J., HUBBARD, T. J., JOHNSON, N., KINSELLA, R., PARKER, A., SPUDICH, G., YATES, A., ZADISSA, A. & SEARLE, S. M. 2013. Ensembl 2013. *Nucleic Acids Res.,* 41**,** D48-55.

FLORI, L., FRITZ, S., JAFFREZIC, F., BOUSSAHA, M., GUT, I., HEATH, S., FOULLEY, J. L. & GAUTIER, M. 2009. The genome response to artificial selection: a case study in dairy cattle. *PLoS One,* 4**,** e6595.

FLORI, L., GONZATTI, M. I., THEVENON, S., CHANTAL, I., PINTO, J., BERTHIER, D., ASO, P. M. & GAUTIER, M. 2012. A quasi-exclusive European ancestry in the Senepol tropical cattle breed highlights the importance of the slick locus in tropical adaptation. *PLoS One,* 7**,** e36133.

FLORI, L., THEVENON, S., DAYO, G. K., SENOU, M., SYLLA, S., BERTHIER, D., MOAZAMI-GOUDARZI, K. & GAUTIER, M. 2014. Adaptive admixture in the West African bovine hybrid zone: insight from the Borgou population. *Mol. Ecol.,* 23**,** 3241-57.

FOSSIEZ, F., DJOSSOU, O., CHOMARAT, P., FLORES-ROMO, L., AIT-YAHIA, S., MAAT, C., PIN, J. J., GARRONE, P., GARCIA, E., SAELAND, S., BLANCHARD, D., GAILLARD, C., DAS MAHAPATRA, B., ROUVIER, E., GOLSTEIN, P., BANCHEREAU, J. & LEBECQUE, S. 1996. T cell interleukin-17 induces stromal cells to produce proinflammatory and hematopoietic cytokines. *J. Exp. Med.,* 183**,** 2593-603.

FUKUDA, N., YOMOGIDA, K., OKABE, M. & TOUHARA, K. 2004. Functional characterization of a mouse testicular olfactory receptor and its role in chemosensing and in regulation of sperm motility. *J Cell Sci,* 117**,** 5835-45.

FULLER, D. 2006. Agricultural Origins and Frontiers in South Asia: A Working Synthesis. *Journal of World Prehistory,* 20**,** 1-86.

GAUGHAN, J. B., MADER, T. L., HOLT, S. M., JOSEY, M. J. & ROWAN, K. J. 1999. Heat tolerance of Boran and Tuli crossbred steers. *J. Anim. Sci.,* 77**,** 2398-405.

GAUTIER, M., FLORI, L., RIEBLER, A., JAFFREZIC, F., LALOE, D., GUT, I., MOAZAMI-GOUDARZI, K. & FOULLEY, J. L. 2009. A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. *BMC Genomics,* 10**,** 550.

GAUTIER, M. & NAVES, M. 2011. Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Mol. Ecol.,* 20**,** 3128-43.

GAUTIER, M. & VITALIS, R. 2012. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics,* 28**,** 1176-7.

GIBBS, R. A., TAYLOR, J. F., VAN TASSELL, C. P., BARENDSE, W., EVERSOLE, K. A., GILL, C. A., GREEN, R. D., HAMERNIK, D. L., KAPPES, S. M., LIEN, S., MATUKUMALLI, L. K., MCEWAN, J. C., NAZARETH, L. V., SCHNABEL, R. D., WEINSTOCK, G. M., WHEELER, D. A., AJMONE-MARSAN, P., BOETTCHER, P. J., CAETANO, A. R., GARCIA, J. F., HANOTTE, O., MARIANI, P., SKOW, L. C., SONSTEGARD, T. S., WILLIAMS, J. L., DIALLO, B., HAILEMARIAM, L., MARTINEZ, M. L., MORRIS, C. A., SILVA, L. O., SPELMAN, R. J., MULATU, W., ZHAO, K., ABBEY, C. A., AGABA, M., ARAUJO, F. R., BUNCH, R. J., BURTON, J., GORNI, C., OLIVIER, H., HARRISON, B. E., LUFF, B., MACHADO, M. A., MWAKAYA, J., PLASTOW, G., SIM, W., SMITH, T., THOMAS, M. B., VALENTINI, A., WILLIAMS, P., WOMACK, J., WOOLLIAMS, J. A., LIU, Y., QIN, X., WORLEY, K. C., GAO, C., JIANG, H., MOORE, S. S., REN, Y., SONG, X. Z., BUSTAMANTE, C. D., HERNANDEZ, R. D., MUZNY, D. M., PATIL, S., SAN LUCAS, A., FU, Q., KENT, M. P., VEGA, R., MATUKUMALLI, A., MCWILLIAM, S., SCLEP, G., BRYC, K., CHOI, J., GAO, H., GREFENSTETTE, J. J., MURDOCH, B., STELLA, A., VILLA-ANGULO, R., WRIGHT, M., AERTS, J., JANN, O., NEGRINI, R., GODDARD, M. E., HAYES, B. J., BRADLEY, D. G., BARBOSA DA SILVA, M., LAU, L. P., LIU, G. E., LYNN, D. J., PANZITTA, F. & DODDS, K. G. 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science,* 324**,** 528-32.

GIFFORD-GONZALEZ, D. & HANOTTE, O. 2011. Domesticating Animals in Africa: Implications of Genetic and Archaeological Findings. *Journal of World Prehistory* 24**,** 1-23.

GLASS, E. J. 2001. The balance between protective immunity and pathogenesis in tropical theileriosis: what we need to know to design effective vaccines for the future. *Res. Vet. Sci.,* 70**,** 71-75.

GORLOV, I. P., KAMAT, A., BOGATCHEVA, N. V., JONES, E., LAMB, D. J., TRUONG, A., BISHOP, C. E., MCELREAVEY, K. & AGOULNIK, A. I. 2002. Mutations of the GREAT gene cause cryptorchidism. *Hum. Mol. Genet.,* 11**,** 2309-18.

GU, J., ORR, N., PARK, S. D., KATZ, L. M., SULIMOVA, G., MACHUGH, D. E. & HILL, E. W. 2009. A genome scan for positive selection in thoroughbred horses. *PLoS One,* 4**,** e5767.

GU, L. H. & COULOMBE, P. A. 2007. Keratin function in skin epithelia: a broadening palette with surprising shades. *Curr. Opin. Cell Biol.,* 19**,** 13-23.

GUIDOBALDI, H. A., TEVES, M. E., UNATES, D. R. & GIOJALAS, L. C. 2012. Sperm transport and retention at the fertilization site is orchestrated by a chemical guidance and oviduct movement. *Reproduction,* 143**,** 587-96.

HACIA, J. G., FAN, J. B., RYDER, O., JIN, L., EDGEMON, K., GHANDOUR, G., MAYER, R. A., SUN, B., HSIE, L., ROBBINS, C. M., BRODY, L. C., WANG, D., LANDER, E. S., LIPSHUTZ, R., FODOR, S. P. & COLLINS, F. S. 1999. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat. Genet.,* 22**,** 164-7.

HAMMOND, A. C., CHASE, C. C., JR., BOWERS, E. J., OLSON, T. A. & RANDEL, R. D. 1998. Heat tolerance in Tuli-, Senepol-, and Brahman-sired F1 Angus heifers in Florida. *J. Anim. Sci.,* 76**,** 1568-77.

HANOTTE, O., BRADLEY, D. G., OCHIENG, J. W., VERJEE, Y., HILL, E. W. & REGE, J. E. 2002. African pastoralism: genetic imprints of origins and migrations. *Science,* 296**,** 336-9.

HANOTTE, O., RONIN, Y., AGABA, M., NILSSON, P., GELHAUS, A., HORSTMANN, R., SUGIMOTO, Y., KEMP, S., GIBSON, J., KOROL, A., SOLLER, M. & TEALE, A. 2003. Mapping of quantitative trait loci controlling trypanotolerance in a cross of tolerant West African N'Dama and susceptible East African Boran cattle. *PNAS,* 100**,** 7443-8.

HANOTTE, O., TAWAH, C. L., BRADLEY, D. G., OKOMO, M., VERJEE, Y., OCHIENG, J. & REGE, J. E. 2000. Geographic distribution and frequency of a taurine *Bos taurus* and an indicine *Bos indicus* Y specific allele amongst sub-saharan African cattle breeds. *Mol Ecol,* 9**,** 387-96.

HANSEN, P. J. 2004. Physiological and cellular adaptations of zebu cattle to thermal stress. *Anim. Reprod. Sci.,* 82-83**,** 349-60.

HAYES, B. J., CHAMBERLAIN, A. J., MACEACHERN, S., SAVIN, K., MCPARTLAN, H., MACLEOD, I., SETHURAMAN, L. & GODDARD, M. E. 2009. A genome map of divergent artificial selection between *Bos taurus* dairy cattle and *Bos taurus* beef cattle. *Anim. Genet.,* 40**,** 176-84.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.,* 37**,** 1-13.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.,* 4**,** 44-57.

JOHNSTON, S. E., GRATTEN, J., BERENOS, C., PILKINGTON, J. G., CLUTTON-BROCK, T. H., PEMBERTON, J. M. & SLATE, J. 2013. Life history trade-offs at a single locus maintain sexually selected genetic variation. *Nature,* 502**,** 93-5.

KAMPINGA, H. H. & CRAIG, E. A. 2010. The HSP70 chaperone machinery: J proteins as drivers of functional specificity. *Nat. Rev. Mol. Cell Biol.,* 11**,** 579-92.

KEIGHTLEY, P. D. & EYRE-WALKER, A. 2000. Deleterious mutations and the evolution of sex. *Science,* 290**,** 331-3.

KEMPER, K. E., SAXTON, S. J., BOLORMAA, S., HAYES, B. J. & GODDARD, M. E. 2014. Selection for complex traits leaves little or no classic signatures of selection. *BMC Genomics,* 15**,** 246.

KERJE, S., SHARMA, P., GUNNARSSON, U., KIM, H., BAGCHI, S., FREDRIKSSON, R., SCHUTZ, K., JENSEN, P., VON HEIJNE, G., OKIMOTO, R. & ANDERSSON, L. 2004. The Dominant white, Dun and Smoky color variants in chicken are associated with insertion/deletion polymorphisms in the PMEL17 gene. *Genetics,* 168**,** 1507-18.

KIJAS, J. W., LENSTRA, J. A., HAYES, B., BOITARD, S., PORTO NETO, L. R., SAN CRISTOBAL, M., SERVIN, B., MCCULLOCH, R., WHAN, V., GIETZEN, K., PAIVA, S., BARENDSE, W., CIANI, E., RAADSMA, H., MCEWAN, J. & DALRYMPLE, B. 2012. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol.,* 10**,** e1001258.

LAMOND, A. I. 1991. Nuclear RNA processing. *Curr. Opin. Cell Biol.,* 3**,** 493-501.

LAMPKIN, G. H. & KENNEDY, J. F. 1965. Some observations on reproduction, weight change under lactation stress and mothering ability of British and crossbred-zebu cattle in tropics. *The Journal of Agricultural Science,* 64**,** 407-412.

LARKIN, D. M., DAETWYLER, H. D., HERNANDEZ, A. G., WRIGHT, C. L., HETRICK, L. A., BOUCEK, L., BACHMAN, S. L., BAND, M. R., AKRAIKO, T. V., COHEN-ZINDER, M., THIMMAPURAM, J., MACLEOD, I. M., HARKINS, T. T., MCCAGUE, J. E., GODDARD, M. E., HAYES, B. J. & LEWIN, H. A. 2012. Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. *PNAS,* 109**,** 7693-8.

LITTLETON, J. T., STERN, M., SCHULZE, K., PERIN, M. & BELLEN, H. J. 1993. Mutational analysis of Drosophila synaptotagmin demonstrates its essential role in Ca(2+)-activated neurotransmitter release. *Cell,* 74**,** 1125-34.

LOFTUS, R. T., ERTUGRUL, O., HARBA, A. H., EL-BARODY, M. A., MACHUGH, D. E., PARK, S. D. & BRADLEY, D. G. 1999. A microsatellite survey of cattle from a centre of origin: the Near East. *Mol. Ecol.,* 8**,** 2015-22.

LOFTUS, R. T., MACHUGH, D. E., BRADLEY, D. G., SHARP, P. M. & CUNNINGHAM, P. 1994. Evidence for 2 independent domestications of cattle. *PNAS,* 91**,** 2757-2761.

LONG, J. C. 1991. The genetic structure of admixed populations. *Genetics,* 127**,** 417-28.

LOWEY, S., WALLER, G. S. & TRYBUS, K. M. 1993. Skeletal muscle myosin light chains are essential for physiological speeds of shortening. *Nature,* 365**,** 454-6.

MACHUGH, D. E., SHRIVER, M. D., LOFTUS, R. T., CUNNINGHAM, P. & BRADLEY, D. G. 1997. Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics,* 146**,** 1071-86.

MALEK, T. R. & CASTRO, I. 2010. Interleukin-2 receptor signaling: at the interface between tolerance and immunity. *Immunity,* 33**,** 153-65.

MANCINI, G., GARGANI, M., CHILLEMI, G., NICOLAZZI, E. L., MARSAN, P. A., VALENTINI, A. & PARISET, L. 2014. Signatures of selection in five Italian cattle breeds detected by a 54K SNP panel. *Mol. Biol. Rep.,* 41**,** 957-65.

MARTINEZ, M. L., MACHADO, M. A., NASCIMENTO, C. S., SILVA, M. V., TEODORO, R. L., FURLONG, J., PRATA, M. C., CAMPOS, A. L., GUIMARAES, M. F., AZEVEDO, A. L., PIRES, M. F. & VERNEQUE, R. S. 2006. Association of BoLA-DRB3.2 alleles with tick (*Boophilus microplus*) resistance in cattle. *Genet. Mol. Res.,* 5**,** 513-24.

MATTIOLI, R. C., JAITNER, J., CLIFFORD, D. J., PANDEY, V. S. & VERHULST, A. 1998. Trypanosome infections and tick infestations: susceptibility in N'Dama, Gobra zebu and Gobra x N'Dama crossbred cattle exposed to natural challenge and maintained under high and low surveillance of trypanosome infections. *Acta Trop.,* 71**,** 57-71.

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., SONSTEGARD, T. S. & VAN TASSELL, C. P. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One,* 4**,** e5350.

MBOLE-KARIUKI, M. N. 2012. *Genomic diversity of East African shorthorn Zebu of western Kenya.* PhD, University of Nottingham.

MBOLE-KARIUKI, M. N., SONSTEGARD, T., ORTH, A., THUMBI, S. M., BRONSVOORT, B. M., KIARA, H., TOYE, P., CONRADIE, I., JENNINGS, A., COETZER, K., WOOLHOUSE, M. E., HANOTTE, O. & TAPIO, M. 2014. Genome-wide analysis reveals the ancient and recent admixture history of East African Shorthorn Zebu from Western Kenya. *Heredity (Edinb),* 113**,** 297-305.

MCGLINCHEY, R. P., SHEWMAKER, F., MCPHIE, P., MONTERROSO, B., THURBER, K. & WICKNER, R. B. 2009. The repeat domain of the melanosome fibril protein Pmel17 forms the amyloid core promoting melanin synthesis. *PNAS,* 106**,** 13731-6.

MCKAY, S. D., SCHNABEL, R. D., MURDOCH, B. M., MATUKUMALLI, L. K., AERTS, J., COPPIETERS, W., CREWS, D., DIAS NETO, E., GILL, C. A., GAO, C., MANNEN, H., STOTHARD, P., WANG, Z., VAN TASSELL, C. P., WILLIAMS, J. L., TAYLOR, J. F. & MOORE, S. S. 2007. Whole genome linkage disequilibrium maps in cattle. *BMC Genet.,* 8**,** 74.

MORADI, M. H., NEJATI-JAVAREMI, A., MORADI-SHAHRBABAK, M., DODDS, K. G. & MCEWAN, J. C. 2012. Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. *BMC Genet.,* 13**,** 10.

MWANGI, E. K., STEVENSON, P., GETTINBY, G. & MURRAY, M. Variation in susceptibility to tsetse-borne trypanosomiasis among Bos indicus cattle breeds in East Africa. *In:* ROWLANDS, G. J. & TEALE, A. J., eds. Towards increased use of trypanotolerance: current research and future directions, 1993 Nairobi, Kenya. 81-86.

NAESSENS, J., TEALE, A. J. & SILEGHEM, M. 2002. Identification of mechanisms of natural resistance to African trypanosomiasis in cattle. *Vet. Immunol. Immunopathol.,* 87**,** 187-94.

O'KELLY, J. C. & SPIERS, W. G. 1976. Resistance to Boophilus microplus (Canestrini) in genetically different types of calves in early life. *J. Parasitol.,* 62**,** 312-7.

O'NEILL, L. A. & GREENE, C. 1998. Signal transduction pathways activated by the IL-1 receptor family: ancient signaling machinery in mammals, insects, and plants. *J. Leukoc. Biol.,* 63**,** 650-7.

OLEKSYK, T. K., SMITH, M. W. & O'BRIEN, S. J. 2010. Genome-wide scans for footprints of natural selection. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.,* 365**,** 185-205.

OLSON, T. A., LUCENA, C., CHASE, C. C., JR. & HAMMOND, A. C. 2003. Evidence of a major gene influencing hair length and heat tolerance in *Bos taurus* cattle. *J. Anim. Sci.,* 81**,** 80-90.

OLSSON, M., MEADOWS, J. R., TRUVE, K., ROSENGREN PIELBERG, G., PUPPO, F., MAUCELI, E., QUILEZ, J., TONOMURA, N., ZANNA, G., DOCAMPO, M. J., BASSOLS, A., AVERY, A. C., KARLSSON, E. K., THOMAS, A., KASTNER, D. L., BONGCAM-RUDLOFF, E., WEBSTER, M. T., SANCHEZ, A., HEDHAMMAR, A., REMMERS, E. F., ANDERSSON, L., FERRER, L., TINTLE, L. & LINDBLAD-TOH, K. 2011. A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs. *PLoS Genet.,* 7**,** e1001332.

ORENGE, C. O., MUNGA, L., KIMWELE, C. N., KEMP, S., KOROL, A., GIBSON, J. P., HANOTTE, O. & SOLLER, M. 2012. Trypanotolerance in N'Dama x Boran crosses under natural trypanosome challenge: effect of test-year environment, gender, and breed composition. *BMC Genet.,* 13**,** 87-102.

PAPPU, B. P., ANGKASEKWINAI, P. & DONG, C. 2008. Regulatory mechanisms of helper T cell differentiation: new lessons learned from interleukin 17 family cytokines. *Pharmacol. Ther.,* 117**,** 374-84.

PARK, J. I., SEMYONOV, J., CHANG, C. L., YI, W., WARREN, W. & HSU, S. Y. 2008. Origin of INSL3-mediated testicular descent in therian mammals. *Genome Res.,* 18**,** 974-85.

PARMENTIER, M., LIBERT, F., SCHURMANS, S., SCHIFFMANN, S., LEFORT, A., EGGERICKX, D., LEDENT, C., MOLLEREAU, C., GERARD, C., PERRET, J. & *ET AL*. 1992. Expression of members of the putative olfactory receptor gene family in mammalian germ cells. *Nature,* 355**,** 453-5.

PARSELL, D. A. & LINDQUIST, S. 1994. *In The Biology of Heat Schock Proteins and Molecular Chaperones*, Cold Spring Harbor Laboratory Press.

PEREZ O'BRIEN, A. M., UTSUNOMIYA, Y. T., MESZAROS, G., BICKHART, D. M., LIU, G. E., VAN TASSELL, C. P., SONSTEGARD, T. S., DA SILVA, M. V., GARCIA, J. F. & SOLKNER, J. 2014. Assessing signatures of selection through variation in linkage disequilibrium between taurine and indicine cattle. *Genet. Sel. Evol.,* 46**,** 19.

PETERSEN, J. L., MICKELSON, J. R., RENDAHL, A. K., VALBERG, S. J., ANDERSSON, L. S., AXELSSON, J., BAILEY, E., BANNASCH, D., BINNS, M. M., BORGES, A. S., BRAMA, P., DA CAMARA MACHADO, A., CAPOMACCIO, S., CAPPELLI, K., COTHRAN, E. G., DISTL, O., FOX-CLIPSHAM, L., GRAVES, K. T., GUERIN, G., HAASE, B., HASEGAWA, T., HEMMANN, K., HILL, E. W., LEEB, T., LINDGREN, G., LOHI, H., LOPES, M. S., MCGIVNEY, B. A., MIKKO, S., ORR, N., PENEDO, M. C., PIERCY, R. J., RAEKALLIO, M., RIEDER, S., ROED, K. H., SWINBURNE, J., TOZAKI, T., VAUDIN, M., WADE, C. M. & MCCUE, M. E. 2013. Genome-wide analysis reveals selection for important traits in domestic horse breeds. *PLoS Genet.,* 9**,** e1003211.

PORTO-NETO, L. R., REVERTER, A., PRAYAGA, K. C., CHAN, E. K., JOHNSTON, D. J., HAWKEN, R. J., FORDYCE, G., GARCIA, J. F., SONSTEGARD, T. S., BOLORMAA, S., GODDARD, M. E., BURROW, H. M., HENSHALL, J. M., LEHNERT, S. A. & BARENDSE, W. 2014. The genetic architecture of climatic adaptation of tropical cattle. *PLoS One,* 9**,** e113284.

PORTO-NETO, L. R., SONSTEGARD, T. S., LIU, G. E., BICKHART, D. M., DA SILVA, M. V., MACHADO, M. A., UTSUNOMIYA, Y. T., GARCIA, J. F., GONDRO, C. & VAN TASSELL, C. P. 2013. Genomic divergence of zebu and taurine cattle identified through high-density SNP genotyping. *BMC Genomics,* 14**,** 876.

PORTO NETO, L. R., BUNCH, R. J., HARRISON, B. E., PRAYAGA, K. C. & BARENDSE, W. 2010. Haplotypes that include the integrin alpha 11 gene are associated with tick burden in cattle. *BMC Genet.,* 11**,** 55.

PRITCHARD, J. K., STEPHENS, M. & DONNELLY, P. 2000. Inference of population structure using multilocus genotype data. *Genetics,* 155**,** 945-59.

QANBARI, S., PAUSCH, H., JANSEN, S., SOMEL, M., STROM, T. M., FRIES, R., NIELSEN, R. & SIMIANER, H. 2014. Classic selective sweeps revealed by massive sequencing in cattle. *PLoS Genet.,* 10**,** e1004148.

QANBARI, S., PIMENTEL, E. C., TETENS, J., THALLER, G., LICHTNER, P., SHARIFI, A. R. & SIMIANER, H. 2010. A genome-wide scan for signatures of recent selection in Holstein cattle. *Anim. Genet.,* 41**,** 377-89.

QANBARI, S. & SIMIANER, H. 2014. Mapping signatures of positive selection in the genome of livestock. *Livestock Science,* 166**,** 133-143.

QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-2.

RAGHAVAN, M., DEL CID, N., RIZVI, S. M. & PETERS, L. R. 2008. MHC class I assembly: out and about. *Trends Immunol.,* 29**,** 436-43.

RAMEY, H. R., DECKER, J. E., MCKAY, S. D., ROLF, M. M., SCHNABEL, R. D. & TAYLOR, J. F. 2013. Detection of selective sweeps in cattle using genome-wide SNP data. *BMC Genomics,* 14**,** 382.

R Development Core Team 2012. R: A language and environment for statistical computing. Vienna, Austria.

REGE, J. E. O., KAHI, A., M., O.-A., MWACHARO, J. & HANOTTE, O. 2001. *Zebu cattle of Kenya: Uses, performance, farmer preferences and measures of genetic diversity*. Nairobi, Kenya, International Livestock Reaserch Institute.

ROTHAMMER, S., SEICHTER, D., FORSTER, M. & MEDUGORAC, I. 2013. A genome-wide scan for signatures of differential artificial selection in ten cattle breeds. *BMC Genomics,* 14**,** 908.

RUBIN, C. J., MEGENS, H. J., MARTINEZ BARRIO, A., MAQBOOL, K., SAYYAB, S., SCHWOCHOW, D., WANG, C., CARLBORG, O., JERN, P., JORGENSEN, C. B., ARCHIBALD, A. L., FREDHOLM, M., GROENEN, M. A. & ANDERSSON, L.

2012. Strong signatures of selection in the domestic pig genome. *PNAS,* 109**,** 19529-36.

RUBIN, C. J., ZODY, M. C., ERIKSSON, J., MEADOWS, J. R., SHERWOOD, E., WEBSTER, M. T., JIANG, L., INGMAN, M., SHARPE, T., KA, S., HALLBOOK, F., BESNIER, F., CARLBORG, O., BED'HOM, B., TIXIER-BOICHARD, M., JENSEN, P., SIEGEL, P., LINDBLAD-TOH, K. & ANDERSSON, L. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature,* 464**,** 587-91.

SABETI, P. C., REICH, D. E., HIGGINS, J. M., LEVINE, H. Z., RICHTER, D. J., SCHAFFNER, S. F., GABRIEL, S. B., PLATKO, J. V., PATTERSON, N. J., MCDONALD, G. J., ACKERMAN, H. C., CAMPBELL, S. J., ALTSHULER, D., COOPER, R., KWIATKOWSKI, D., WARD, R. & LANDER, E. S. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature,* 419**,** 832-7.

SANKARARAMAN, S., SRIDHAR, S., KIMMEL, G. & HALPERIN, E. 2008. Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.,* 82**,** 290-303.

SCHAUVLIEGE, R., JANSSENS, S. & BEYAERT, R. 2007. Pellino proteins: novel players in TLR and IL-1R signalling. *J. Cell Mol. Med.,* 11**,** 453-61.

SCHEET, P. & STEPHENS, M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.,* 78**,** 629-44.

SCHMIDT, H., SWOBODA, R., JÄTZOLD, R. & SECTION, K. F. M. 1979. *Farm management handbook of Kenya*, Ministry of Agriculture, Farm Management Section.

SHI, Y. & MANLEY, J. L. 2007. A complex signaling pathway regulates SRp38 phosphorylation and pre-mRNA splicing in response to heat shock. *Mol. Cell,* 28**,** 79-90.

SKINNER, J. D. & LOUW, G. N. 1966. Heat stress and spermatogenesis in Bos indicus and *Bos taurus* cattle. *J. Appl. Physiol.,* 21**,** 1784-90.

SPEHR, M., GISSELMANN, G., POPLAWSKI, A., RIFFELL, J. A., WETZEL, C. H., ZIMMER, R. K. & HATT, H. 2003. Identification of a testicular odorant receptor mediating human sperm chemotaxis. *Science,* 299**,** 2054-8.

TANG, H., CHOUDHRY, S., MEI, R., MORGAN, M., RODRIGUEZ-CINTRON, W., BURCHARD, E. G. & RISCH, N. J. 2007a. Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am. J. Hum. Genet.,* 81**,** 626-33.

TANG, K., THORNTON, K. R. & STONEKING, M. 2007b. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.,* 5**,** e171.

THEOS, A. C., TRUSCHEL, S. T., RAPOSO, G. & MARKS, M. S. 2005. The Silver locus product Pmel17/gp100/Silv/ME20: controversial in name and in function. *Pigment Cell Res.,* 18**,** 322-36.

THUMBI, S. M., BRONSVOORT, B. M., POOLE, E. J., KIARA, H., TOYE, P. G., MBOLE-KARIUKI, M. N., CONRADIE, I., JENNINGS, A., HANDEL, I. G., COETZER, J. A., STEYL, J. C., HANOTTE, O. & WOOLHOUSE, M. E. 2014. Parasite co-infections and their impact on survival of indigenous cattle. *PLoS One,* 9**,** e76324.

THUNNISSEN, M. M., NORDLUND, P. & HAEGGSTROM, J. Z. 2001. Crystal structure of human leukotriene A(4) hydrolase, a bifunctional enzyme in inflammation. *Nat. Struct. Biol.,* 8**,** 131-5.

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**,** 1088-91.

TURNER, J. W. 1980. Genetic and biological aspects of Zebu adaptability. *J. Anim. Sci.,* 50**,** 1201-5.

UTECH, K. B. W., WHARTON, R. H. & KERR, J. D. 1978. Resistance to Boophilus microplus (Canestrini) in different breeds of cattle. *Australian Journal of Agricultural Research,* 29**,** 885-895.

UTSUNOMIYA, Y. T., PEREZ O'BRIEN, A. M., SONSTEGARD, T. S., VAN TASSELL, C. P., DO CARMO, A. S., MESZAROS, G., SOLKNER, J. & GARCIA, J. F. 2013.

Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. *PLoS One,* 8**,** e64280.

VANDERHAEGHEN, P., SCHURMANS, S., VASSART, G. & PARMENTIER, M. 1993. Olfactory receptors are displayed on dog mature sperm cells. *J. Cell Biol.,* 123**,** 1441-52.

VOIGHT, B. F., KUDARAVALLI, S., WEN, X. & PRITCHARD, J. K. 2006. A map of recent positive selection in the human genome. *PLoS Biol.,* 4**,** e72.

WESTERN, D. & FINCH, V. 1986. Cattle and pastoralism: survival and production in arid lands. *Human Ecology,* 14**,** 77-94.

WILKINSON, S., LU, Z. H., MEGENS, H. J., ARCHIBALD, A. L., HALEY, C., JACKSON, I. J., GROENEN, M. A., CROOIJMANS, R. P., OGDEN, R. & WIENER, P. 2013. Signatures of diversifying selection in European pig breeds. *PLoS Genet.,* 9**,** e1003453.

WRIGHT, S. 1951. The genetical structure of populations. *Annals of Eugenics,* 15**,** 323-354.

WUEST, S. C., EDWAN, J. H., MARTIN, J. F., HAN, S., PERRY, J. S., CARTAGENA, C. M., MATSUURA, E., MARIC, D., WALDMANN, T. A. & BIELEKOVA, B. 2011. A role for interleukin-2 trans-presentation in dendritic cell-mediated T cell activation in humans, as revealed by daclizumab therapy. *Nat. Med.,* 17**,** 604-9.

XU, L., BICKHART, D. M., COLE, J. B., SCHROEDER, S. G., SONG, J., VAN TASSELL, C. P., SONSTEGARD, T. S. & LIU, G. E. 2015. Genomic signatures reveal new evidences for selection of important traits in domestic cattle. *Mol. Biol. Evol.,* 32**,** 711-25.

YANG, S., LI, X., LI, K., FAN, B. & TANG, Z. 2014. A genome-wide scan for signatures of selection in Chinese indigenous and commercial pig breeds. *BMC Genet.,* 15**,** 7.

YU, J., LAN, J., ZHU, Y., LI, X., LAI, X., XUE, Y., JIN, C. & HUANG, H. 2008. The E3 ubiquitin ligase HECTD3 regulates ubiquitination and degradation of Tara. *Biochem. Biophys. Res. Commun.,* 367**,** 805-12.

ZHENG, H., STRATTON, C. J., MOROZUMI, K., JIN, J., YANAGIMACHI, R. & YAN, W. 2007. Lack of Spem1 causes aberrant cytoplasm removal, sperm deformation, and male infertility. *PNAS,* 104**,** 6852-7.

Chapter three

Signatures of positive selection in East African shorthorn zebu: Autosomal high-density genome-wide single nucleotide polymorphism and full genome sequence analyses

Abstract

Genome-wide identification of signatures of positive selection requires high genome coverage with informative markers and linkage disequilibrium over short distances. These requirements were only partly met by the Illumina BovineSNP50 BeadChip (Chapter 2) given the relatively low number and the ascertainment bias of the markers. Here, we genotyped East African Shorthorn Zebu (EASZ) with the Illumina BovineHD BeadChip, which includes more than 770,000 SNP markers to identify candidate regions of signatures of positive selection across all EASZ autosomes. To increase power, we combined SNP genotyping information from two Extended Haplotype Homozygosity (EHH)-based (*Rsb* and *iHs*) tests and one inter-population allele frequency (ΔAF) test in a single composite analysis. We identified 101 candidate regions of positive selection. The full genome of 10 pooled EASZ was also sequenced, and pooled heterozygosities (*Hp*) of 100 kb sliding windows were calculated, leading to the identification of 165 candidate sweep regions for positive selection. Thirty-five candidate regions were common to both analyses. Fourteen showed substantial zebu ancestral deviations from the mean of the genome, suggesting a possible origin for the selected haplotypes. We assessed the presence of these 35 candidate regions in zebu-taurine admixed cattle populations from Uganda (East Africa) and Nigeria (West Africa). Fifteen regions are present in East African zebu cattle, and seven are shared between East and West African zebu cattle. Furthermore, the identification of candidate regions common to East and West African zebu cattle populations allows us to fine map these regions up to ~ 94 kb. Genes and QTL associated with adaptive traits, e.g., reproduction, immunity and heat stress, were found within the regions, indicating possible selection for adaptation to the African environment, and/or the intrinsic zebu-taurine admixed genome background. Four non-synonymous variants were found in three candidate genes as possible causative mutations under selection. The results of this chapter are paving the way for the identification of the functional autosomal causative mutations linked to the adaptation of the EASZ.

Introduction

In Chapter 2, we used the Illumina BovineSNP50 BeadChip to successfully identify the candidate signatures of positive selection in East African Shorthorn Zebu. However, our approach had several methodological limitations, including insufficient genome coverage and SNP ascertainment bias towards European taurine breeds (Matukumalli *et al.*, 2009). The subsequent development of the high-density Illumina BovineHD BeadChip addresses to a large extent these issues (Rincon *et al.*, 2011). It has been estimated that at least 300,000 evenly distributed SNPs are required to adequately represent the entire cattle genome and demonstrate a consistent LD phase between markers across breeds (Goddard and Hayes, 2009). The Illumina BovineHD genotyping BeadChip contains more than 700,000 SNPs with an average inter-marker space of 3.43 kb. Also, a substantial number of zebu cattle (a total of 104 samples from Brahman, Gir and Nellore) were used in the process of SNP validation, thereby addressing the European taurine SNP ascertainment bias typically observed in the Illumina BovineSNP50 BeadChip(http://res.illumina.com/documents/products/datasheets/datasheet_bovinehd.pdf). Thus, the use of this new tool may prove to be particularly informative for our comprehensive search of the signatures of positive selection in EASZ.

Already, a few studies have demonstrated its usefulness in zebu and taurine cattle. Santana *et al.* (2014) and Utsunomiya *et al.* (2013a) performed association studies between SNP markers and feed efficiency and birth weight, respectively. The same tool has also been used in Holstein and Red Dairy cattle for the prediction of genomic breeding values for protein yield, fertility and udder health traits (Su *et al.*, 2012). However, the utilization of this high-density SNP chip in the detection of signatures of positive selection has only been reported hitherto in two studies, dealing with improved dairy and/or beef cattle breeds (Utsunomiya *et al.*, 2013b, Perez O'Brien *et al.*, 2014). Utsunomiya *et al.* (2013b) have reported for the first time in cattle the use of a composite mapping index “Meta-analysis of Selection Signals” to define footprints of positive selection related to beef and milk production traits, with the anticipation that “*If each signature provides distinct information about selective sweeps, combining the signals should have greater power for localizing the source of selection than any single test*” (Grossman *et al*., 2010). More specifically, in this study the SNP-specific *P*-values of four genome-wide analyses, *iHS*, *Rsb*, Δ*DAF* (difference in derived allele frequency between two populations) and a SNP-specific local heterozygosity depression analysis (*SHp*), were combined into a single value. The approach uses the Z-transformation method “Stouffer’s method” (Stouffer *et al*., 1949), which can be used to combine the SNP-specific *P*-values from multiple genome-wide tests (Whitlock, 2005).

In addition to high-density SNP chips, the full genome sequence data have also been analysed in livestock for the detection of candidate signatures of positive selection. Indeed, several next-generation sequencing platform manufacturers have been established in recent years, including Roche/454, Illumina, SOLiD, etc. (Metzker, 2010). They are now routinely in use, while new platforms are being developed. The continuous improvement of these platforms in terms of read length, sequencing speed and accuracy, and reduction in sequencing cost means that full genome sequences may be increasingly investigated for the detection of signatures of positive selection in livestock.

Already, by assessing pooled heterozygosity in sliding windows within the chicken genome, several selective sweep regions in broiler and layer chicken breeds have been found to be associated with growth rates and reproduction (Rubin *et al*., 2010). Based on the same approach, positively selected genome regions carrying genes linked to coat colour, muscle development, growth and body composition have been identified in domestic pig breeds (Amaral *et al*., 2011, Rubin *et al*., 2012). In cattle, the detection of signatures of selection through full genome sequencing has been reported in the commercial Holstein-Friesian breed, where Larkin *et al*. (2012) detected positive selection signals in genome regions overlapping with previously identified milk-production QTL and genes related to milk production, e.g., *Plasminogen*. Recently, the genome of the *Bos taurus indicus* cattle (Gir) has also been fully sequenced, and characterised for selective sweeps (Liao *et al*., 2013). Genes associated with heat tolerance, innate immunity and brain function have been identified. These may have contributed to the Gir adaptability and resistance to parasites in tropical climates (Berman, 2011, Liao *et al*., 2013).

Following the detection of candidate regions and genes under positive selection, identification of the putative causative variants has also been attempted in livestock. These variants can be classified into SNPs and copy number variants (CNV), e.g., small (≤ 10 bp) insertions/ deletion (indels) and large-scale (> 1 kb) CNV. For example, a single glycine to arginine non-synonymous mutation at residue 558 on *TSHR* has been considered as a candidate causal variant for the *TSHR* sweep during chicken domestication (Rubin *et al*., 2010). Moreover, in commercial European domestic pig breeds, a non-synonymous substitution at residue 192 (proline to leucine) in the NR6A1 protein has been proposed as the causative mutation for the QTL associated with increased number of vertebrae in the comparison of domestic pigs to the wild boar ancestor (Mikawa *et al*., 2007).

Beside non-synonymous mutations, several CNV have been shown to be associated with different phenotypes in livestock. For example, a deletion in all but the first exon of the *SH3RF2* is highly associated with increased growth rate in domestic chicken (Rubin *et al*., 2010). In pigs, four duplication events (one encompasses the entire *KIT* gene, one upstream and two downstream of the same gene) are responsible for the white coat colour (Rubin *et al*., 2012). Similarly, the same phenotype in sheep has been attributed to a duplication of the entire *ASIP* gene (Norris and Whan, 2008).

Here, we report for the first time an autosomal-wide analysis for signatures of positive selection in the EASZ using high-density genome-wide SNP genotyping and full genome (autosomal) sequencing. Regions revealed by these two tools were considered candidate regions for positive selection, and they were functionally characterized to define biological pathways. These regions were further validated by comparing them with candidate regions identified in different indigenous zebu-taurine admixed cattle from East (Uganda) and West (Nigeria) Africa through high-density genome-wide SNP analysis. Putative causative mutations (SNPs and CNV) were investigated within candidate genes located in the EASZ full genome and genome-wide SNP analyses' overlapping regions.

Materials and Methods

SNP genotyping and quality control

A total of 92 non-European introgressed EASZ from 20 different randomly selected sublocations covering four distinct ecological zones in the Western and Nyanza provinces of Kenya (de Clare Bronsvoort *et al*., 2013, Mbole-Kariuki *et al*., 2014), which were genotyped for 777,962 SNPs mapped to the UMD3.1 bovine reference genome (Elsik *et al*., 2009) using the Illumina BovineHD Genotyping BeadChip (Rincon *et al*., 2011). Genotypes from 105 Nigerian (NGR) zebu (25 Adamawa Gudali (AG), 2 Azawak (AZ), 22 Bunaji (BJ), 22 Red bororo (OR), 19 Sokoto Gudali (SO), 3 Wadara (WD) and 12 Yakanaji (YK)), 8 Muturu (MT) (NGR taurine), 77 Ugandan (UGN) zebu (25 Ankole (AO), 16 Karamojong zebu (KR), 23 Nanda (NG), 13 Serere zebu (ZS)), 24 N'Dama (NDM) from Guinea (African taurine), 63 Holstein-Friesian (HOL) (European taurine), 36 Jersey (JER) (European taurine), 35 Nellore (NEL) (Asian zebu) and 30 Gir (GIR) (Asian zebu) were provided by Dr Tad Sonstegard and Dr Heather Huson (USDA-ARS, Maryland), Dr Oyekanmi Nash (NABDA, Nigeria) and Dr Christopher Mukasa (Ahmadu Bello University, Nigeria).

The quality control (QC) was conducted on the autosomal SNPs with known mapping coordinates (735,297 SNPs) in all the cattle samples using *check.marker* function implemented in the GenABEL package (Aulchenko *et al*., 2007) for the R software (R Development Core Team, 2012). SNPs with a minor allele frequency (MAF) of less than 5% (n = 68,731) or call rate of less than 95% (n = 18,667) were filtered out from the dataset. These included a small number (n = 1,712) of SNPs found to fail both criteria. In total, 649,611 SNPs were therefore retained. The ancestral allele for 373,005 SNPs of these SNPs has been reported previously by Utsunomiya *et al*. (2013b) following the genotyping of three non-cattle *Bovinae* species: two *Bos gaurus* (gaur), six *Bubalus bubalis* (water buffalo) and two *B. grunniens* (yak), with fixed allele in the three species considered as ancestral. These SNPs only were included in the signatures of selection analyses (mean genome gap size = 6.7 kb, median genome gap size = 4.4 kb, SD = 12.1 kb).

QC also excluded samples with a genotyping call rate below 95%, or in which pairwise identity-by-state (IBS) was greater than 95%, with the lower call rate animal eliminated. One ZS sample did not pass the genotyping call rate criterion, while 15 samples (2 GIR, 1 NEL, 4 HOL, 4 JER, 2 AG, 1 AZ and 1 WD) were excluded due to the IBS criterion.

Extended Haplotype Homozygosity (EHH)-based statistics (*Rsb* and *iHS*)

Rsb (Tang *et al*., 2007) analyses were performed between each of the African stable zebu-taurine admixed cattle populations (Tijjani, 2013, Mbole-Kariuki *et al*., 2014) (EASZ, combined UGN zebu cattle populations (AO, KR, NG, ZS) and combined NGR zebu cattle populations (AG, AZ, BJ, OR, SO, WD, YK)) and the combined reference cattle populations (NEL, GIR, NDM, MT, HOL, JER) using the *rehh* package (Gautier and Vitalis, 2012) for the R software (R Development Core Team, 2012). Additional *Rsb* analyses were also performed between EASZ and the continental-specific cattle reference populations, namely European taurine (HOL and JER), African taurine (NDM and MT), and Asian zebu (NEL and GIR). The standardized *Rsb* values were normally distributed (Figure S3.1), so a one-tailed Z-test was applied to identify statistically significant SNPs under selection on the African stable admixed cattle genome. One-sided upper-tail *P*-values were derived as *1-Φ(Rsb)* from the Gaussian cumulative density function *Φ*. Candidate regions were defined as having five adjacent SNPs not separated by more than 500 kb (the average extent of LD in cattle (McKay *et al*., 2007)) passing the threshold of -$\log_{10}$ *P*-value = 4, which corresponds to *P*-values equal to 0.0001.

iHS (Voight *et al*., 2006) analyses were conducted on EASZ, combined UGN zebu cattle and combined NGR zebu cattle populations using the *rehh* package

(Gautier and Vitalis, 2012) for the R software (R Development Core Team, 2012). This statistic was calculated for SNPs that passed the QC criteria and exhibit a within-population MAF of at least 0.05, since the algorithm of *iHS* is not optimal for the calculation of the statistic on fixed allele SNPs. As with *Rsb*, the standardized *iHS* values followed a normal distribution (Figure S3.1), so a two-tailed Z-test was applied to identify statistically significant SNPs under selection with either an unusual extended haplotype of ancestral or derived alleles relative to the autosomes. Two-sided *P*-values were derived as $1\text{-}2|\Phi(iHS)\text{-}0.5|$ from the Gaussian cumulative density function Φ. Candidate regions were defined as in *Rsb*.

As a prerequisite to the *Rsb* and *iHS* analyses, *fastPHASE* 1.4 (Scheet and Stephens, 2006) was used to phase the genotyped SNPs into the corresponding haplotypes using K10 and T10 criteria (Utsunomiya *et al*., 2013b). Population label information was used to estimate the phased haplotype background.

Inter-population change in SNP allele frequency (∆*AF*)

Grossman *et al*. (2010) described a method called ∆*DAF* to assess if there is a difference in derived allele frequency between two populations at a SNP. ∆*AF* is a derivation of ∆*DAF*, which investigates absolute allele frequency difference between two populations (Carneiro *et al*., 2014). In this analysis, the mean frequency of allele 1 was estimated for EASZ, combined UGN zebu cattle and combined NGR zebu cattle populations, separately (AF_{pop1}). Likewise, the mean frequency of allele 1 for each SNP was calculated for the reference cattle populations, combined and separated (AF_{pop2}). The values of ∆*AF* ($AF_{pop1} – AF_{pop2}$) were normally distributed (Figure S3.1). These values were standardized using the distribution mean and standard deviation (SD) (standardized ∆*AF* = (∆*AF* – mean ∆*AF*) / SD ∆*AF*) and a two-tailed Z-test was applied to identify statistically significant SNPs showing higher allele frequency in population 1. Two-sided *P*-values were derived as $1\text{-}2|\Phi(\text{standardized } \Delta\text{AF})\text{-}0.5|$ from the Gaussian cumulative density function Φ. Candidate regions were defined as having five adjacent SNPs not separated by more than 500 kb passing the threshold of -log_{10} *P*-value = 4.

Meta-analysis of Selection Signals (*meta-SS*)

The Stouffer method was used to combine the *P*-values obtained from the three tests, *Rsb*, *iHS* and Δ*AF*, for each analysed set of populations, i.e., EASZ, UGN and NGR, in *meta-SS* analyses as in Utsunomiya *et al*. (2013b). Each value, for every SNP in each test, was transformed to a Z-score *via* the quantile function *qnorm* in R software. Then, the SNP-specific Z-scores were combined together according to the following equation: $Z_i = (Z_{Rsb} + Z_{iHS} + Z_{\Delta AF}) / \sqrt{k}$, where i and *k* are numbers of SNPs and tests, respectively. The resulting Z-scores were referred back to the standard normal distribution to obtain combined *P*-values. Candidate regions were defined as having five adjacent SNPs not separated by more than 500 kb passing the threshold of $-\log_{10}$ *P*-value = 4.

EASZ whole genome sequencing

Two pools of five unrelated EASZ DNA samples were sequenced using an ABI SOLiD 4 genetic analyser. One pool included five animals that survived one year on farm from their date of birth, while the other pool included five animals that died within one year following infection episode(s) (Table S3.1). SOLiD 2 x 50 bp mate-paired libraries (1.5 kb average insertion) were constructed and sequenced at the Deep Seq facility at the University of Nottingham according to the manufacturer's instructions. Reads were mapped to the UMD3.1 bovine reference genome (Elsik *et al*., 2009) using the LifeScope Genomic Analysis software 2.5.1 re-sequencing mapping pipeline (http://www.lifetechnologies.com/lifescope). The sequences of the two pools were combined into a single pool of 10 EASZ samples. SNPs and indels were called using the diBayes package implemented in LifeScope. A minimum coverage of two uniquely mapped reads and two non-reference allele counts were required. Additionally, a minimum read mapping quality of 20 (MAPQ ≥ 20) and base quality of 20 were also implemented.

Selective sweep analysis (pooled heterozygosity *Hp*)

Pooled heterozygosity *Hp* of SNPs detected in the pooled 10 EASZ full genome SOLiD sequences were calculated on 100 kb sliding windows with 10 kb incremental steps. Window sizes were extended by the number of uncovered bases to improve the accuracy of the calculation and consistency across windows. For each SNP in the window, the number of reads for the most and least frequent allele was counted (n_{MAJ} and n_{MIN}, respectively). *Hp* values were calculated using the following formula: $Hp = 2 \sum n_{MAJ} \sum n_{MIN} / (\sum n_{MAJ} + \sum n_{MIN})^2$ (Rubin *et al*., 2010). The *Hp* values were Z-transformed ($ZHp = Hp – \text{mean } Hp / \text{SD } Hp$). A $ZHp \leq - 4$ was applied as a threshold to specify windows carrying a selective sweep as in Liao *et al*. (2013). Overlapping candidate windows were merged into a single region as described by Rubin *et al*. (2010) and Liao *et al*. (2013).

Candidate regions characterization

Protein-coding and RNA genes mapped within the candidate regions were processed using the functional annotation tool implemented in *DAVID* Bioinformatics resources 6.7 to determine the over-represented (enriched) functional terms (Huang da *et al*., 2009a, Huang da *et al*., 2009b). An enrichment score of 1.3, which is equivalent to the Fisher exact test *P*-value = 0.05, was used as a threshold to define the significantly enriched functional terms in comparison to the whole bovine reference genome background. The list of genes mapped on the UMD3.1 reference bovine genome was obtained from the *Ensembl Genes 73* database (Flicek *et al*., 2013) using the *BioMart* tool (Kinsella *et al*., 2011). The coordinates of the genes were intersected with the candidate sweep regions using *intersectBed* function from the *BedTools* software (Quinlan and Hall, 2010).

Variants (SNPs and indels) in the genes were annotated using the variant effect predictor tool on the Ensembl website (Flicek *et al*., 2013). Comparisons with the previously discovered bovine variants listed in the dbSNP database (Sherry *et al*., 2001) (http://www.ncbi.nlm.nih.gov/SNP/) classified these variants into

EASZ-specific and general bovine variants. The biological effects of the candidate non-synonymous variants were predicted by PolyPhen-2 online tool (Adzhubei *et al*., 2010).

The bovine Quantitative Trait Loci (QTL) and their genome coordinates (UMD 3.1) were downloaded from the cattle QTL database (http://www.animalgenome.org/cgi-bin/QTLdb/BT/index). The *intersectBed* function from the *BedTools* software (Quinlan and Hall, 2010) was used to overlap these QTL with the identified candidate regions for positive selection.

EASZ exome enrichment and sequencing

The Agilent SureSelect^{XT} target enrichment kit (cat no. G7530-90004) was used for bovine exome sequence enrichment. It covers the coding regions in the UMD3.1 reference genome (coding regions from Refseq and Ensembl, no UTR "untranslated regions") and microRNA, or a total of ~ 45 Mb of the bovine sequence. The exomes of 10 EASZ samples (Table S3.1) were sequenced at the Deep Seq facility at the University of Nottingham using an ABI SOLiD 5500 genetic analyser.

Using the LifeScope Enrichment Sequencing Pipeline, the generated 75 bp reads were mapped to the UMD3.1 bovine reference genome (Elsik *et al*., 2009). Only those reads with MAPQ ≥ 30 were considered for depth of coverage analysis. Table S3.2 summarizes the number of aligned reads, average depth of coverage and percentage of reference exome for each sample. The same SNP calling criteria used in the EASZ whole genome sequence were implemented for the exome analysis except that reads with MAPQ ≥ 30 and base quality of 28 were used.

Copy Number Variation (CNV) analysis

To identify putative CNV (multiple copies), we identified regions in the 10 sequenced EASZ exomes that have different sequence depth coverage compared to the average depth coverage for the whole exome. The depth of coverage (DOC) was calculated using the Genome Analysis Toolkit (GATK) (McKenna *et al.*, 2010). For each EASZ sequence, these values were normalized against the total number of reads divided by 1,000,000 for comparison between samples. As the distributions of these normalized values were positively skewed (Figure S3.2), they were standardised based on their median and standard deviation (SD) using the following equation: SDOC = (DOC-median DOC)/SD-DOC. SDOC value of three was arbitrarily chosen to define targeted regions with multiple copies. The GC content of the 10 EASZ exome sequences was calculated using GATK (McKenna *et al.*, 2010).

Estimating Asian zebu ancestry proportion in EASZ (admixture analysis)

Admixture analysis *via* a Bayesian clustering method implemented in STRUCTURE software version 2.3 (Pritchard *et al.*, 2000) was conducted on the autosomes for the EASZ, NDM, MT, NEL, GIR. Three independent replicates of an admixed model with independent allele frequencies were run for a burn-in period of 25,000 iterations and 50,000 Markov Chain Monte Carlo steps for K = 2. The mean output file was generated using *CLUMPP* software version 1.1.2 (Jakobsson and Rosenberg, 2007) and graphically displayed by *Distruct* software version 1.1 (Rosenberg, 2004).

Estimation of excesses-deficiencies in Asian zebu ancestry at candidate regions

LAMP software version 2.4 (Sankararaman *et al.*, 2008) was used to estimate the Asian zebu and African taurine ancestry proportions of the genotyped SNPs in EASZ samples. The genome-wide autosomal zebu ancestry proportion (71%) and the African taurine ancestry proportion (29%) in EASZ were obtained from the admixture proportions α of the STRUCTURE analysis (Pritchard *et al.*, 2000). An estimated number of 500 generations was set for the beginning of the zebu-taurine admixture in light of our current knowledge of zebu arrival on the continent, assuming a generation time of six years (Keightley and Eyre-Walker, 2000). A uniform recombination rate of 1 cM = 1 Mb was assumed. The average excess/deficiency in Asian zebu ancestry at each SNP (ΔAZ) was calculated by subtracting the average estimated Asian zebu ancestry of the SNP from the average estimated Asian zebu ancestry of all SNPs across autosomes. The median ΔAZ for the significant SNPs of EASZ *meta-SS* analysis within candidate regions were considered.

Results

Genome-wide SNP analyses

Results of *iHS*, *Rsb* and Δ*AF* analyses between EASZ and the combined reference populations are presented in Figure 3.1 with the genome position of the significant regions presented in Table S3.3. A single candidate region in BTA 7 was identified by the *iHS* analysis, while 19 and six regions were considered as candidates by the *Rsb* and Δ*AF* analyses, respectively.

As the results of the three tests followed a normal distribution (Figure S3.1), and their genome-wide average *P*-values were weakly correlated (Pearson correlation coefficient $r \leq 0.228$, Table S3.4), the *P*-values of each SNP for the three tests were combined in a *meta-SS* analysis. A total of 98 autosomal candidate regions were defined by this analysis (Figure 3.1 and Table S3.3). All candidate regions revealed by each individual test were identified as candidate regions in the *meta-SS* analysis, with the exception of three regions on BTA 13 identified by the Δ*AF* analysis only. Seventy-seven new candidate regions were detected after combining the *P*-values of the three tests.

Of the total 98 candidate regions, 58 and 60 were identified when the EASZ population was analysed against European taurine (HOL and JER) and African taurine (NDM and MT) populations, respectively, in separate *meta-SS* analyses (Figure S3.6 and Table S3.5). However, only 15 regions were identified upon comparison against the Asian zebu (Figure S3.3 and Table S3.5).

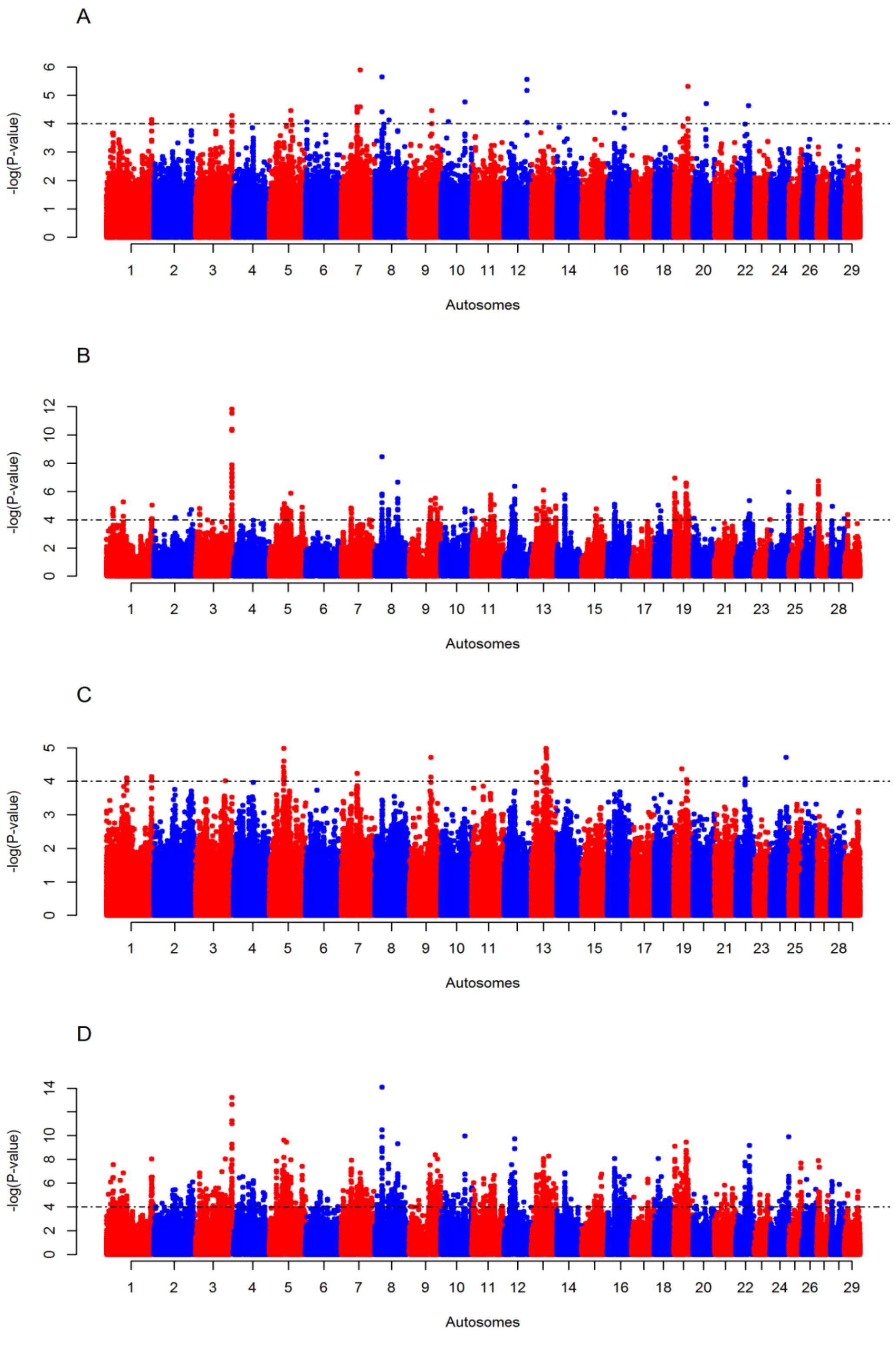

Figure 3.1: Manhattan plots of the genome-wide autosomal (A) EASZ *iHS*, (B) *Rsb*, (C) Δ*AF* and (D) *meta-SS* analyses between EASZ and combined reference populations (Holstein-Friesian, Jersey, N'Dama, Muturu, Nellore and Gir). Threshold set as – $\log_{10}$ *P*-value = 4.

Results of analyses between the zebu cattle populations of Uganda or Nigeria with the combined reference populations are presented in Figures 3.2 and 3.3 with the genome position of the significant regions presented in Tables S3.6 and S3.7. In zebu cattle populations from Uganda (UGN), the three tests (*iHS*, *Rsb* and Δ*AF*) revealed 6, 25 and 3 autosomal candidate regions, respectively. The analyses with zebu cattle from Nigeria (NGR) showed 4, 22 and 4 autosomal candidate regions for the *iHS, Rsb* and Δ*AF* tests, respectively. After combining the tests *P*-values, 86 and 97 autosomal regions were considered as candidate regions for positive selection in UGN and NGR zebu cattle populations, respectively (Figure 3.2, Figure 3.3, Table S3.6, and Table S3.7).

Of the 101 identified autosomal candidate regions from EASZ comparisons, 32 candidate regions were identified in the zebu cattle populations from Uganda (East African zebu-sharing candidate regions). Fourteen regions were found to overlap between the EASZ and the NGR zebu cattle candidate regions (Table S3.8). Whilst, 22 regions were shared across the three comparisons (East and West African zebu-sharing candidate regions) (Figure 3.4).

The genome coordinates of East African zebu-sharing and East and West African zebu-sharing candidate regions are indicated in Table 3.1. Comparisons between EASZ, UGN and NGR allowed us to narrow down the size of 18 candidate regions. As indicated in Table 3.2, these regions now range in size from ~ 94 kb to ~ 894 kb with the largest reduction in size (~ 1.9 Mb) observed for the candidate region on BTA 12.

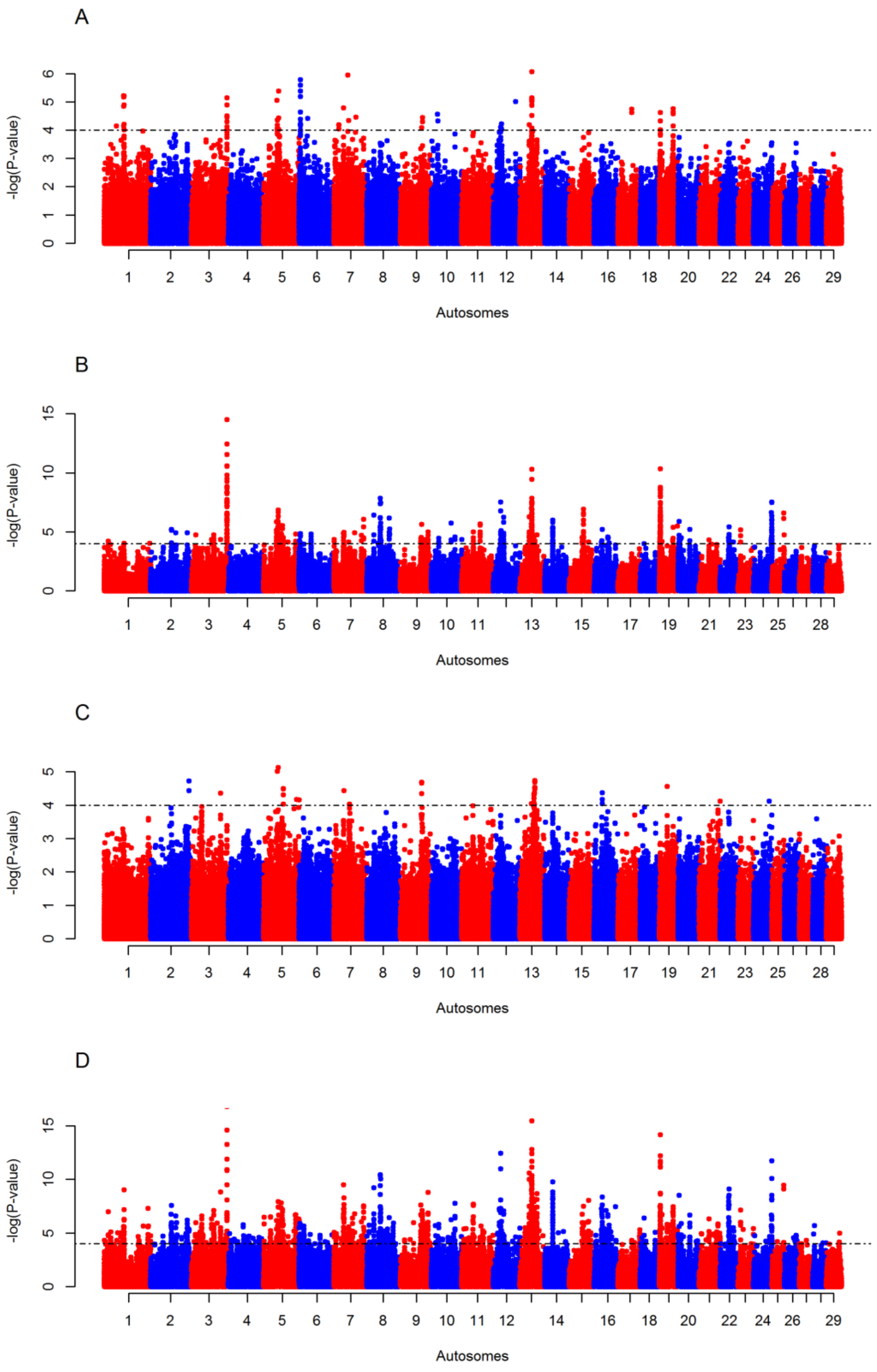

Figure 3.2: Manhattan plots of the genome-wide autosomal (A) *iHS* on zebu cattle populations from Uganda (UGN), (B) *Rsb*, (C) Δ*AF* and (D) *meta-SS* analyses between UGN and combined reference populations (Holstein-Friesian, Jersey, N'Dama, Muturu, Nellore and Gir). Threshold set as – $\log_{10}$ *P*-value = 4.

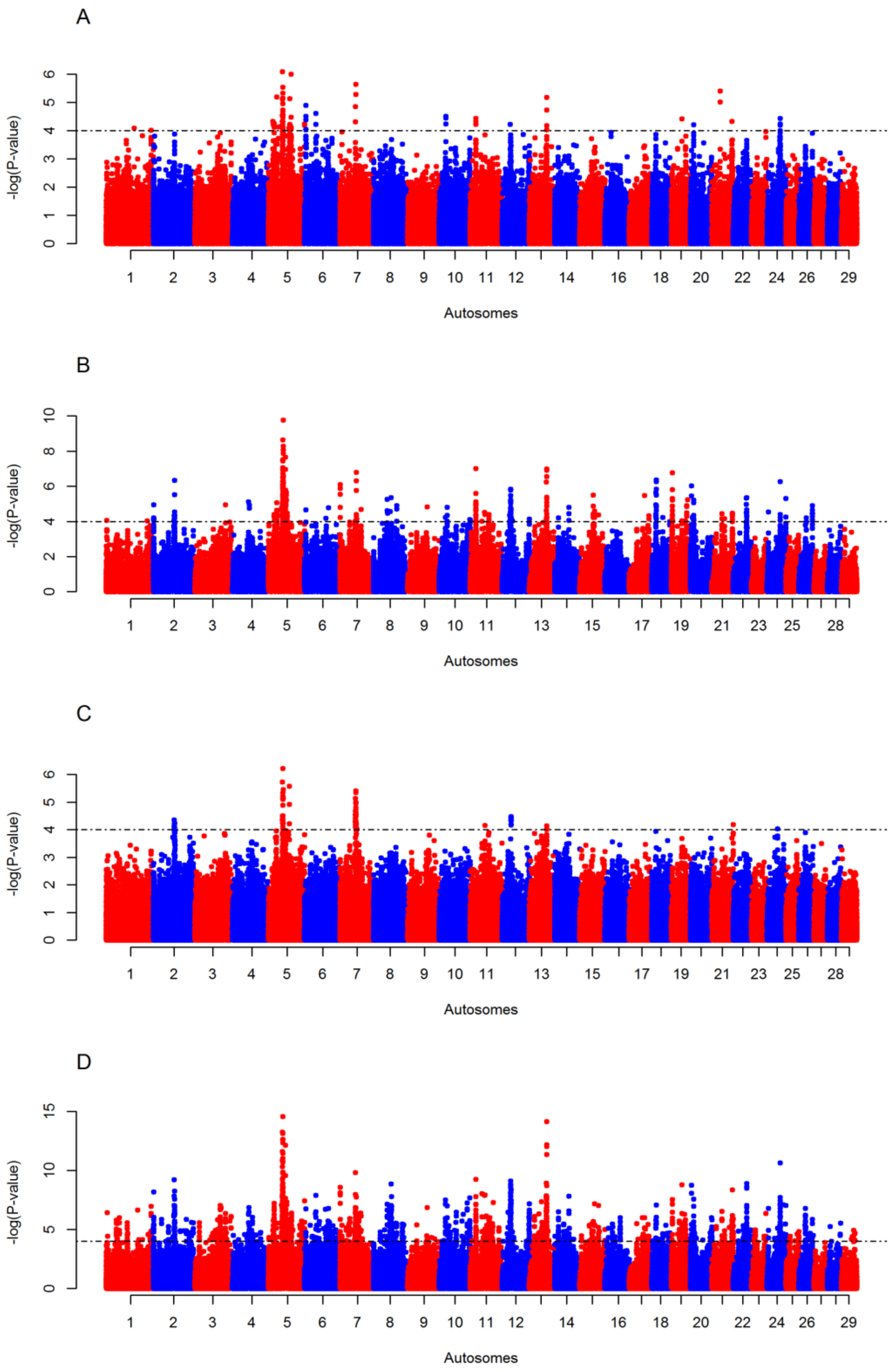

Figure 3.3: Manhattan plots of the genome-wide autosomal (A) *iHS* on zebu cattle populations from Nigeria (NGR), (B) *Rsb*, (C) ΔAF and (D) *meta-SS* analyses between NGR and combined reference populations (Holstein-Friesian, Jersey, N'Dama, Muturu, Nellore and Gir). Threshold set as $-\log_{10}$ *P*-value = 4.

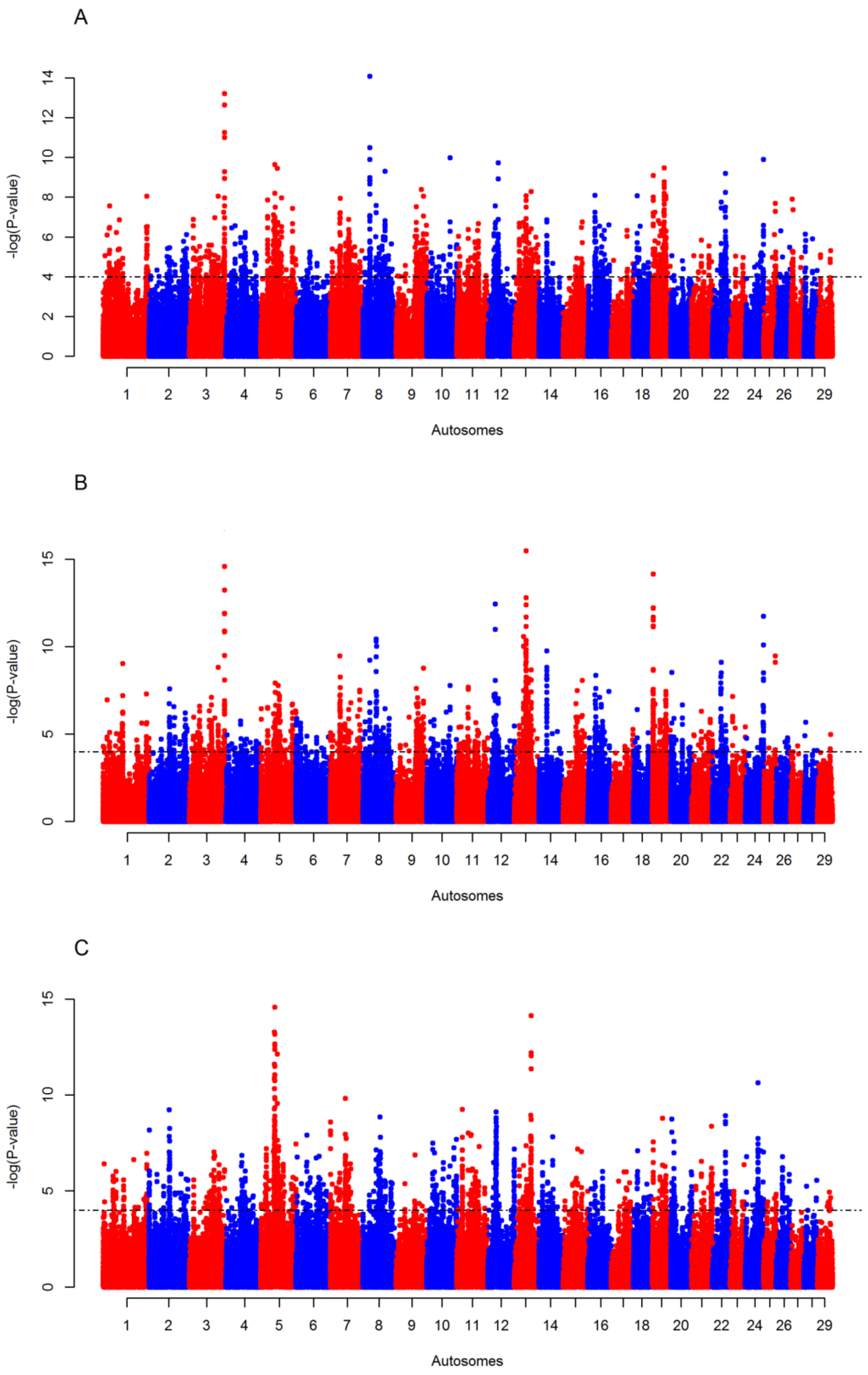

Figure 3.4: Manhattan plots of the genome-wide autosomal *meta-SS* analyses between (A) EASZ, (B) Uganda, (C) Nigerian zebu cattle populations and combined reference populations (Holstein-Friesian, Jersey, N'Dama, Muturu, Nellore and Gir). Threshold set as – $\log_{10}$ *P*-value = 4.

Table 3.1: Shared candidate regions obtained by the autosomal genome-wide SNP analyses. (A) East African zebu-sharing candidate regions (EASZ and Uganda (UGN) zebu cattle) (B) East and West African zebu-sharing candidate regions (EASZ, Uganda (UGN) and Nigeria (NGR) zebu cattle populations)

A	**EASZ and UGN**				**Other Studies[2]**	**B**	**EASZ, UGN and NGR**				**Other Studies[2]**
	Position (UMD 3.1)		**value [3]**	**ΔAZ[1]**			**Position (UMD 3.1)**		**value [3]**	**ΔAZ[1]**	
BTA	**start**	**stop**				**BTA**	**start**	**stop**			
2	125,159,084	125,994,861	5.17	0.07390136	Gautier and Navas, 2011	**1**	**149,547,998**	**149,960,460**	6.50	**0.1771624**	
3	34,254,043	34,727,876	5.52	0.09020536		2	70,314,631	71,161,113	5.45	0.1119444	Gautier *et al*., 2009; Liao *et al*., 2013
3	120,601,191	121,238,836	13.22	0.07390136		**3**	**76,084,701**	**76,413,468**	6.6	**-0.1652296**	Kemper *et al*., 2014*
5	23,652,016	24,338,695	7.86	0.09020536	Gautier *et al*., 2009	3	98,862,402	99,283,161	8.05	0.1200969	
5	109,303,999	109,688,098	7.43	0.07933636		5	43,834,751	44,574,214	7.12	0.1065094	
7	61,232,987	61,396,966	5.90	-0.1217506	Gautier *et al*., 2009	5	48,477,903	49,212,943	9.64	0.1282494	Gautier *et al*., 2009; Ramey *et al*.,2013; Perez O'Brien *et al*., 2014; Xu *et al*., 2015
8	23,344,221	23,663,852	14.08	-0.0402296		5	62,272,683	62,587,423	5.69	-0.0837076	Kemper *et al*., 2014*
8	65,373,897	65,634,601	5.55	0.05487936		7	32,640,500	33,093,884	7.17	0.0956404	
9	69,198,185	69,406,467	7.53	0.07933636		7	50,281,923	50,670,070	4.49	0.1336834	
9	73,280,867	74,185,868	5.73	0.10107536		7	62,551,178	62,782,874	6.89	0.0793364	
9	**76,289,561**	**76,853,587**	5.06	**0.13911836**		**11**	**62,343,547**	**62,548,419**	5.76	**0.1662924**	
9	94,121,197	94,242,831	8.06	0.11194436	Larkin *et al*., 2012*	**12**	**28,949,354**	**29,151,436**	4.23	**-0.2141426**	Gautier *et al*., 2009 & Gautier and Navas, 2011 & Porto Neto *et al*., 2013 & Flori *et al*., 2014; Liao *et al*., 2013

10	80,515,703	80,796,559	9.99	0.07118386		**13**	**18,132,557**	**18,320,265**	6.7	**0.1662924**	
11	38,402,190	39,743,107	6.37	-0.0076206	Gautier *et al.*, 2009 Kemper *et al.*, 2014*	16	25,389,029	25,540,339	8.09	0.1173794	
11	71,387,248	72,221,099	6.33	-0.1299031		16	50,610,769	50,762,363	4.69	0.0412924	Gautier and Navas, 2011
12	21,086,969	21,254,061	6.17	-0.0565336		19	2,568,979	2,765,065	9.1	-0.0619686	
12	24,843,013	25,658,768	7.56	0.11194486		19	27,004,483	27,143,239	6.66	0.1228144	
12	35,689,908	36,746,504	9.74	0.05759636		19	44,788,419	44,924,467	6.3	0.1336834	
13	39,579,929	41,356,847	8.07	0.08748786	Perez O'Brien *et al.*, 2014	**19**	**46,580,102**	**46,673,984**	8.05	**0.1391184**	
13	**47,532,424**	**48,142,997**	**4.22**	**0.14455335**		21	33,590,777	33,696,403	5.03	-0.0021856	
13	49,433,476	49,762,965	4.99	0.12281436	Porto Neto *et al.*, 2013	22	45,231,901	46,126,149	9.20	0.1119444	Gautier *et al.*, 2009 & Flori *et al.*, 2014
13	**50,616,630**	**50,837,529**	5.68	**0.18259636**		24	61,972,128	62,530,799	9.90	0.0032489	
13	58,273,562	58,599,491	8.29	0.06846636	Flori *et al.*, 2014 Kemper *et al.*, 2014*						
14	28,186,226	28,430,215	6.86	0.08477036							
16	**26,979,772**	**27,160,301**	5.77	**0.18259636**							
16	**46,869,577**	**47,614,377**	6.08	**0.18803136**	Liao *et al.*, 2013 Kemper *et al.*, 2014*						
19	**3,337,282**	**3,823,638**	5.19	**-0.1815336**							
19	9,515,063	9,780,078	7.24	-0.0945776							

19	39,330,233	39,519,992	4.68	0.10107536							
19	40,045,779	40,808,559	9.48	0.03585736	Gautier *et al.*, 2009						
21	**60,026,698**	**60,449,172**	5.55	**0.17716236**							
22	**29,533,544**	**30,366,810**	7.76	**0.15542236**							

[1]ΔAZ = estimated excess/deficiency of the Asian zebu ancestry proportion.

[2]The candidate regions were cross-referenced with the ones obtained previously on tropical-adapted cattle and commercial breeds.

[3]- log(*P*-value) of highest EASZ SNP within the region.

Bold (deviation by more than +/- 1 standard deviation from the autosomal mean ΔAZ).

*Commercial breeds studies

Table 3.2: Comparing genomic coordinates of EASZ candidate regions with East and West African zebu-sharing candidate regions (EASZ, Uganda and Nigeria).

EASZ candidate regions (UMD3.1)				African zebu candidate regions (UMD3.1)				reduction in size (bp)
BTA	start	stop	size	BTA	start	stop	size	
1	149,241,884	149,992,523	750,639	1	149,547,998	149,960,460	412,462	338,177
2	70,314,631	71,161,113	846,482	2	70,314,631	71,161,113	846,482	0
3	76,084,701	76,781,970	697,269	3	76,084,701	76,413,468	328,767	368,502
3	98,862,402	99,422,213	559,811	3	98,862,402	99,283,161	420,759	139,052
5	43,230,619	44,574,214	1,343,595	5	43,834,751	44,574,214	739,463	604,132
5	48,477,903	49,268,610	790,707	5	48,477,903	49,212,943	735,040	55,667
5	62,272,683	62,659,987	387,304	5	62,272,683	62,587,423	314,740	72,564
7	31,748,136	33,875,610	2,127,474	7	32,640,500	33,093,884	453,384	1,674,090
7	50,281,923	50,809,190	527,267	7	50,281,923	50,670,070	388,147	139,120
7	62,415,406	63,117,931	702,525	7	62,551,178	62,782,874	231,696	470,829
11	61,877,437	62,548,419	670,982	11	62,343,547	62,548,419	204,872	466,110
12	27,050,192	29,151,436	2,101,244	12	28,949,354	29,151,436	202,082	1,899,162
13	18,130,223	18,421,481	291,258	13	18,132,557	18,320,265	187,708	103,550
16	24,517,859	25,540,339	1,022,480	16	25,389,029	25,540,339	151,310	871,170
16	50,610,769	50,762,363	151,594	16	50,610,769	50,762,363	151,594	0
19	2,568,979	2,765,065	196,086	19	2,568,979	2,765,065	196,086	0
19	26,909,816	27,143,239	233,423	19	27,004,483	27,143,239	138,756	94,667
19	44,788,419	45,414,418	625,999	19	44,788,419	44,924,467	136,048	489,951
19	46,031,543	46,786,391	754,848	19	46,580,102	46,673,984	93,882	660,966
21	33,590,777	33,696,403	105,626	21	33,590,777	33,696,403	105,626	0
22	45,102,551	46,400,273	1,297,722	22	45,231,901	46,126,149	894,248	403,474
24	61,008,938	62,530,799	1,521,861	24	61,972,128	62,530,799	558,671	963,190

Candidate regions characterization

A total of 94 EASZ candidate regions contain genes based on UMD3.1 reference genome annotation, while the remaining seven were considered as gene desert regions (Table S3.9). The gene regions include 1024 genes (Table S3.10). Based on DAVID functional cluster analysis, these genes were clustered into 110 functional clusters (Table S3.11). Six of these clusters were significantly enriched relative to the bovine genome as indicated in Table 3.3. A total of 309 genes are within the candidate regions common to EASZ and cattle populations from Uganda (Table S3.10). These genes were grouped into 33 functional term clusters (Table S3.11), in which three were significantly enriched (Table 3.3). The regions shared between EASZ and Nigerian zebu cattle populations include 217 genes. A total of 19 functional term clusters were identified (Table S3.11), in which two were significantly enriched (Table 3.3). For candidate regions shared across all populations (EASZ, UGN and NGR), 87 genes were identified (Table S3.10). They were grouped into 10 functional term clusters (Table S3.11), in which a single cluster, associated with immune response to bacterial infection, was significantly enriched (Table 3.3).

Table 3.3: Significantly enriched functional term clusters in EASZ, East African zebu (EASZ and zebu cattle populations from Uganda (UGN)), EASZ and Nigerian zebu cattle populations (NGR), and African zebu (EASZ, UGN and NGR zebu cattle populations) candidate regions.

EASZ		**East African zebu**		**EASZ and NGR**		**African zebu**	
Functional term cluster	**Score***	**Functional term cluster**	**Score***	**Functional term cluster**	**Score***	**Functional term cluster**	**Score***
Intermediate protein filaments and keratin	4.2	nucleoplasm and nuclear lumen	1.8	Lysozyme activity and defence response to bacterial infection	1.6	Lysozyme activity and defence response to bacterial infection	1.8
Cytoskeleton	2.4	cell-cell junction	1.6	oxidation reduction process and glucose dehydrogenase	1.4		
Enzyme inhibitor activity	2.2	Lysozyme activity and defence response to bacterial infection	1.5				
cell-cell junction	1.7						
cell-substrate (e.g., extracellular matrix) junction	1.5						
Immune response and antigen processing and presenting	1.3						

*Enrichment score following DAVID analysis. A score = 1.3, equivalent to Fisher exact test *P*-value = 0.05, was used as a significant threshold.

Identification of sweep regions using EASZ full genome sequence

Pooling the 10 EASZ full autosomal sequences generated a total of 615,413,240 reads with MAPQ ≥ 20 (0.1% probability of incorrect alignment) mapped on the UMD3.1 bovine reference genome. These reads covered ~ 97% of the reference autosomes with, on average, 11 times depth coverage. SNP calling using LifeScope diBayes package identified a total of 10,466,699 autosomal SNPs (8,114,664 heterozygotes and 2,352,035 homozygotes). In addition, 288,099 autosomal indels (154,002 deletions, 133,897 insertions, and a further 200 representing a combination of these two features) were identified.

Regions with signatures of selective sweep were defined by assessing the pooled SNP heterozygosity *Hp* of 100 kb windows incremented by 10 kb. To determine the most appropriate window size to calculate *Hp* values, two window sizes were initially tested (50 kb and 100 kb). The 100 kb window size was chosen due to the low fraction of windows with ≤ 20 SNPs, 48 windows out of total 250,931 autosomal windows (~ 0.01%) compared to those of 50 kb, which yielded 1,705 windows (~ 0.3%) with SNP number ≤ 20 SNPs. Windows with low SNP number can lead to spurious fixation (Rubin *et al*., 2010). The 100 kb window size has also been used to detect sweep regions in the Gir cattle genome using the same approach (Liao *et al*., 2013).

Figure S3.4a shows the distribution of SNPs in the 100 kb autosomal windows with a mean of 297 SNPs per window. The resulting mean autosomal *Hp* value was 0.42 with (SD = 0.025). The *Hp* value of each window was Z-transformed to quantify its degree of deviation from the mean autosomal *Hp* value. Of the total 250,930 windows, 1,825 (~ 0.73%) had a ZHp score of ≤ - 4 (Figure 3.5, Figure S3.4b and Table S3.12), resulting in 165 candidate sweep regions (Table S3.13). The largest sweep region, ~ 2 Mb in size, was on BTA 7 (51.4 – 53.4 Mb). This region contained windows with the lowest ZHp value (ZHp = -16.6).

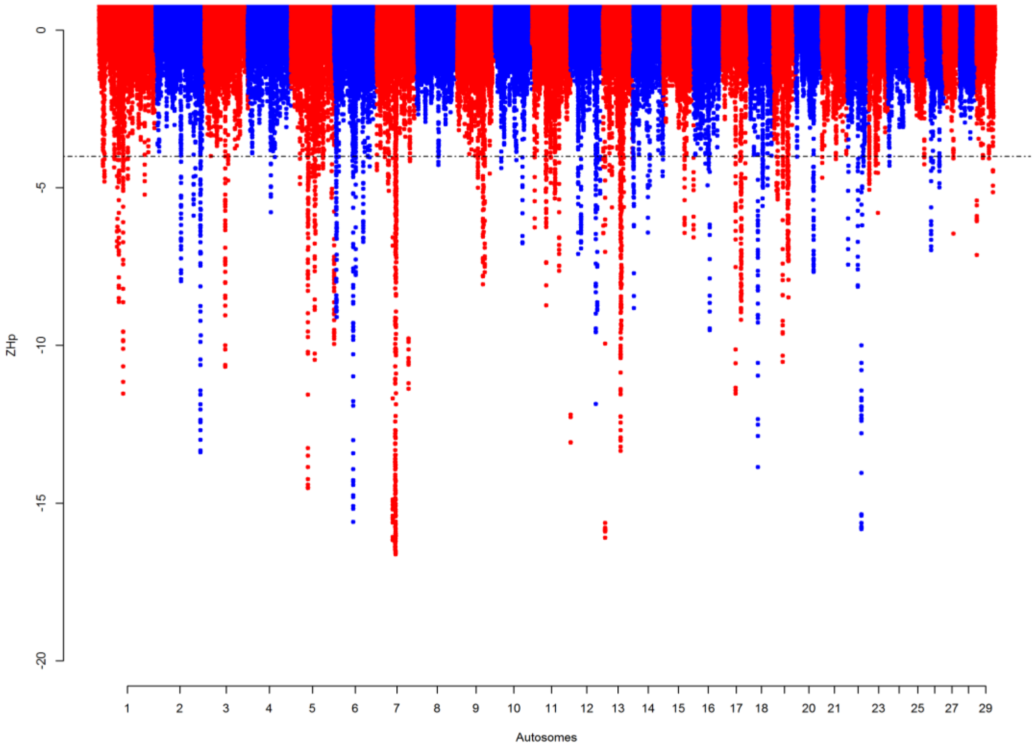

Figure 3.5: Manhattan plot of the genome-wide autosomal *Hp* analysis on EASZ. Each point represents a 100 kb window. The significant threshold is ZHp = - 4.

Following the Ensembl-annotated UMD3.1 bovine reference genome database, 518 genes were found within 133 of these sweep regions (Table S3.14). The remaining 32 regions are gene desert islands with no coding regions identified so far (Table S3.9). DAVID analyses were conducted in two levels including; i) genes within EASZ *Hp* candidate regions and ii) genes within all EASZ candidate regions from SNPs and *Hp* analyses. The first DAVID analysis identified 57 functional term clusters (Table S3.11) with seven significantly enriched clusters (Table 3.4), whilst the second one defined 148 clusters (Table S3.11) with seven significantly enriched clusters (Table 3.4).

Table 3.4: Significantly enriched functional term clusters of the genes mapped within (A) *Hp* candidate sweep regions (B) all EASZ candidate regions (SNPs and *Hp* analyses).

A		B	
Functional term cluster	**Score***	**Functional term cluster**	**Score***
Cell-cell adhesion	4.52	Intermediate protein filaments and keratin	3.23
Response to hormones stimuli (e.g., growth hormones)	1.63	Enzyme inhibitor activity	2.34
Regulation of cell cycle and differentiation	1.42	Protein transport and localization	1.98
Regulation of growth and development	1.34	Cell-cell adhesion	1.98
Chemotaxis and locomotory behaviour	1.34	Cytoskeleton	1.45
Regulation of T and B cells proliferation and activation	1.32	Cell-cell junction and connexin	1.42
Regulation of myeloid leukocyte differentiation	1.31	Nuclear lumen and nucleoplasm	1.34

*Enrichment score following DAVID analysis (a score equals to 1.3, equivalent to Fisher exact test *P*-value = 0.05, was used as a significant threshold).

Overlapping candidate sweep regions between genome-wide SNPs and *Hp* sequence analyses

Among the 165 autosomal candidate sweep regions, 35 regions overlapped with the genome-wide SNP analyses candidate regions identified in EASZ. These include 22 regions also revealed in the SNP analyses in cattle populations from Uganda, in which seven regions shared between the East and West (Nigeria) African zebu populations (Table 3.5).

Within the 35 overlapping candidate regions, 185 genes were identified (Table S3.15). DAVID analysis revealed 23 functional clusters (Table S3.11) with two significantly enriched functional clusters: response to hormone stimulus and signalling pathway (enrichment score = 2.07), and transcription regulation (enrichment score = 1.3). Also worth mentioning is a functional cluster associated with the immune system development and regulation, although it does not reach the 1.3 threshold (enrichment score = 1.24). For the 22 East African overlapping sweep regions, 98 genes were identified (Table S3.15); these were grouped into four functional clusters: GTPase regulator activity (enrichment score = 1.17), protein complexes assembly (enrichment score =

0.76), regulation of transcription (enrichment score = 0.45), and nucleotides and ribonucleotides binding (enrichment score = 0.13), but none are significant. A total of 24 genes are within the seven overlapping regions across the East and West African populations (Table S3.15). These genes were grouped into a single cluster associated with ion binding (enrichment score = 0.14).

Table 3.5: The overlapping candidate sweep regions between EASZ *Hp* and genome-wide SNP analyses.

BTA	Start	Stop	Mean ZHp	ΔAZ[1]	Other studies[2]	BTA	Start	Stop	Mean ZHp	ΔAZ[1]	Other studies[2]
1	**54,880,001**	**55,141,728**	-6.56	**0.139118**		**12****	**29,110,001**	**29,438,417**	-5.83	**-0.21414**	Gautier *et al*., 2009; Gautier and Navas, 2011; Liao *et al*., 2013; Porto Neto *et al*., 2013; Flori *et al*., 2014
1	55,150,001	55,253,859	-4.12	NA		**13***	**47,980,001**	**48,164,495**	-4.99	**0.204336**	
2**	70,570,001	70,811,366	-6.68	0.114662	Gautier *et al*., 2009; Liao *et al*., 2013; Kemper *et al*., 2014¥	**13**	**48,650,001**	**49,056,444**	-10.07	**0.20977**	Porto Neto *et al*., 2013
2**	70,990,001	71,191,313	-5.27	0.111944	Gautier *et al*., 2009; Liao *et al*., 2013; Kemper *et al*., 2014¥	13*	49,340,001	49,551,378	-5.80	-0.00762	Porto Neto *et al*., 2013
2*	**125,300,001**	**125,620,820**	-8.12	**0.144553**	Gautier *et al*., 2009	**13***	**49,590,001**	**49,844,283**	-6.78	**0.182596**	Porto Neto *et al*., 2013
2*	125,640,001	126,083,262	-7.66	0.073901	Gautier *et al*., 2009	**13***	**50,240,001**	**50,852,056**	-7.32	**0.182596**	
5**	48,610,001	49,021,113	-5.01	0.125531	Liao *et al*., 2013; Kemper *et al*., 2014¥; Xu *et al*., 2015; Perez O'Brien *et al*., 2014	13	55,510,001	55,623,671	-4.48	0.128249	
5**	49,120,001	49,241,076	-4.47	0.128249	Liao *et al*., 2013; Kemper *et al*., 2014¥; Perez O'Brien *et al*., 2014	13	82,010,001	82,111,606	-4.06	0.09564	
7	31,740,001	31,897,059	-4.54	0.08477	Flori *et al*., 2014	19*	9,500,001	9,631,079	-4.72	-0.09458	
7	33,100,001	33,293,306	-4.43	-0.00219		19**	26,890,001	27,154,002	-7.60	0.122814	Gautier *et al*., 2009
7	**51,360,001**	**53,362,761**	-10.79	**0.198901**	Gautier *et al*., 2009; Liao *et al*., 2013; Porto Neto *et al*., 2013; Qanbari *et al*., 2014¥	19*	39,270,001	39,422,844	-4.42	0.014119	

9*	**73,890,001**	**74,081,863**	-5.45	**0.188031**		19*	40,490,001	40,714,976	-4.60	0.035857	
9*	76,600,001	76,876,188	-6.12	0.09564		**19**	**40,960,001**	**41,450,870**	-5.98	**0.144553**	
11*	39,240,001	39,530,799	-5.93	0.106509	Gautier *et al*., 2009; Kemper *et al*., 2014¥	**19**	**42,890,001**	**43,122,753**	-6.09	**0.149988**	Chan *et al*., 2010
11*	39,550,001	39,683,044	-4.38	NA	Gautier *et al*., 2009; Kemper *et al*., 2014¥	**19**	**43,140,001**	**43,341,262**	-6.53	**0.141836**	Chan *et al*., 2010
11	75,230,001	75,441,012	-6.40	0.054879		**22***	**30,030,001**	**30,260,687**	-6.05	**0.188031**	
12	20,870,001	21,021,506	-5.64	0.079336	Gautier *et al*., 2009	**22****	**45,220,001**	**45,370,457**	-4.17	**0.149988**	Gautier *et al*., 2009; Chan *et al*., 2010; Flori *et al*., 2014
12*	21,130,001	21,320,859	-4.63	-0.05653	Gautier *et al*., 2009						

[1] ΔAZ = estimated excess/deficiency of the Asian zebu proportion.

[2] The candidate regions were cross-referenced with the ones obtained previously on tropical-adapted cattle and commercial breeds.

Bold (deviation by more than +/- 1 standard deviation from the autosomal mean ΔAZ).

¥Commercial breeds studies.

NA: No SNPs passed – $\log_{10}$ (*P*-value) = 4 of EASZ *meta-SS* analysis.

* shared between East African zebu populations (EASZ and Uganda). ** Shared between East (EASZ and Uganda) and West (Nigeria) African zebu populations.

Candidate genes and their polymorphisms

Twelve genes were selected as examples of interesting candidates within the EASZ genome-wide SNP and *Hp* overlapping candidate regions based on their biological roles (Table 3.6). These genes have functional roles directly linked to crucial traits for adaptation to the African environment, e.g., reproduction-related traits and immunological-related traits.

Table 3.6: Candidate genes considered in this chapter

Biological role	Candidate genome region	Gene ID	Gene name
Immunity	BTA 19: 40,960,001-41,450,870	*GM-CSF*	colony stimulating factor 3 (granulocyte)
	BTA 19: 40,960,001-41,450,870	*CCR7*	chemokine (C-C motif) receptor 7
Reproduction and fertility	BTA 12: 29,110,001-29,438,417	*RXFP2*	relaxin/insulin-like family peptide receptor 2
	BTA 19: 40,960,001-41,450,870	*RARA*	retinoic acid receptor, alpha
	BTA 7: 51,360,001-53,362,761	*SPATA24*	spermatogenesis associated 24
	BTA 19: 26,890,001-27,154,002	*SPAG7*	sperm associated antigen 7
Heat stress	BTA 2: 125,640,001-126,083,262	*DNAJC8*	DnaJ (Hsp40) homolog, subfamily C, member 8
	BTA 7: 51,360,001-53,362,761	*DNAJC18*	DnaJ (Hsp40) homolog, subfamily C, member 18
	BTA 7: 51,360,001-53,362,761	*HSPA9*	heat shock 70kDa protein 9
	BTA 19: 42,890,001-43,122,753	*HSPB9*	heat shock protein, alpha-crystallin-related, B9
Anatomical development	BTA 5: 48,610,001-49,021,113	*Man-1*	inner nuclear membrane protein
	BTA 7: 33,100,001-33,293,306	*LOX*	lysyl oxidase
	BTA 12: 29,110,001-29,438,417	*RXFP2*	relaxin/insulin-like family peptide receptor 2

A total of 986 SNPs and 41 indels were detected within these candidate genes (Table S3.16). After annotating these variants, four non-synonymous variants (one in *Man-1*, one in *SPAG7* and two in *RXFP2*) were identified (Table 3.7 & S3.16). The remaining variants were classified into different groups, e.g., synonymous, intron variants, 5’ UTR variants, etc. The positions of most of these variants have not been reported as polymorphic in other cattle populations, i.e., not detected previously as variants in the dbSNP database, and hence they were considered as

EASZ-specific variants. They were 830 SNPs, including the four non-synonymous variants mentioned above, and 36 indels. Based on the exome data of the 10 EASZ samples, none of the non-synonymous variants were found to be fixed for the alternative allele (Table 3.4).

Table 3.7: The reference and alternative allele frequencies for the four non-synonymous variants of the 10 EASZ exome sequences. Location (BTA: position in bp), amino acid substitution (reference amino acid, residue, alternative amino acid).

SNP location	Gene	Amino acid substitution	Reference allele frequency	Alternative allele frequency	Biological effect*
BTA 5: 48,781,846	*Man-1*	T 665 I	G = 95%	A = 5%	Probably damaging
BTA 12: 29,243,223	*RXFP2*	C 459 G	A = 60%	C = 40%	Probably damaging
BTA 12: 29,280,777	*RXFP2*	N 19 S	T = 85%	C = 15%	Benign
BTA 19: 27,072,057	*SPAG7*	R 144 Q	G = 100%	A = 0%	Benign

*Based on PolyPhen-2 online tool (Adzhubei *et al*., 2010)

Overlaps with known QTL

The *Hp* candidate sweep regions were within 1336 bovine QTL related to several biological roles, such as feed behaviour, reproduction and immunity. Examples of these QTL include the following: residual feed intake QTL, sperm motility QTL, and general disease susceptibility (Table S3.17). Six trypanotolerance QTL, defined by Hanotte *et al*. (2003), were also found (Table S3.18). Moreover, several production traits QTL intersected with these candidate regions, e.g., marbling score, milk fat percentage and milk protein yield. The EASZ genome-wide SNP and *Hp* overlapping candidate regions were within 316 QTL (Table S3.17), in which two are trypanotolerance QTL (Table S3.18). Candidate regions specific to East African zebu and African zebu cattle intersected with 138 QTL each (Table S3.17).

Identification of putative CNVs

The median depth of coverage for the regions captured by the SureSelect XT target enrichment system ranged from 36.33 reads/bp to 69.31 reads/bp (SD= 55 reads/bp to 93.99 reads/bp) for the 10 EASZ exome sequences. Their normalized values showed intervals of smaller range from 0.87 reads/bp to 0.92 reads/bp (SD= 1.2 reads/bp to 1.28 reads/bp) (Table S3.19).

A total of 17 EASZ genome-wide SNP analyses and *Hp* analysis overlapping candidate regions contained signals of CNV (multiple copies). These signals were within 31 targeted exons in 17 genes (Table S3.20). The two highest standardised values for depth of coverage were the 10th exon of *HBS1L* and the 13th exon of *Man-1*. The 9th, 14th and 15th exons of *Man-1* and the 5th exon of *RXFP2* also showed signals of CNV.

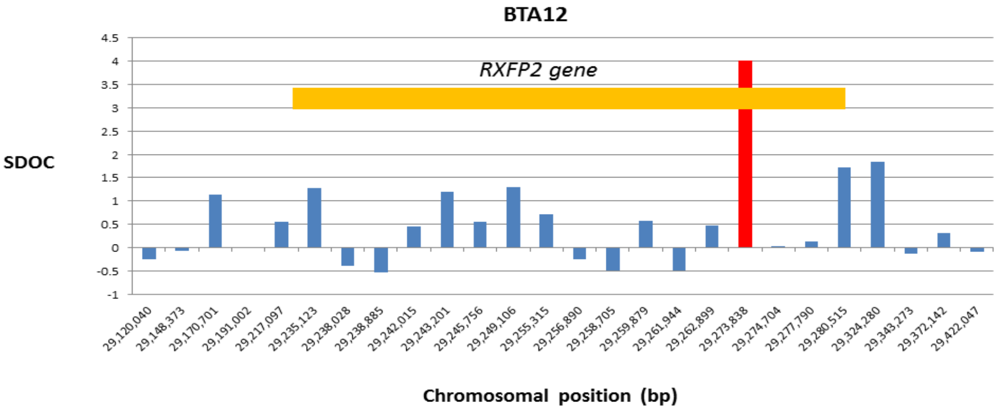

Figure 3.6: A plot showing the putative CNV within the regions captured by exome sequencing in the BTA12 candidate sweep region (29.1 – 29.4 Mb).Bar in red demonstrates a high standardized depth of coverage (SDOC) in comparison to the median depth of coverage for the whole EASZ autosomal exome.

Estimation of excesses-deficiencies in Asian zebu ancestry

The admixture analysis conducted on EASZ estimated an average genetic proportion of 0.71 ± 0.009 SD and 0.29 ± 0.009 SD of Asian zebu and African taurine ancestries. As observed in Figure 3.8, the EASZ showed low variation in the Asian zebu ancestry proportions across animal ranging from 0.69 to 0.74. This type of within-population homogeneity has been previously observed by Mbole-Kariuki *et al*. (2014) using lower density SNP array.

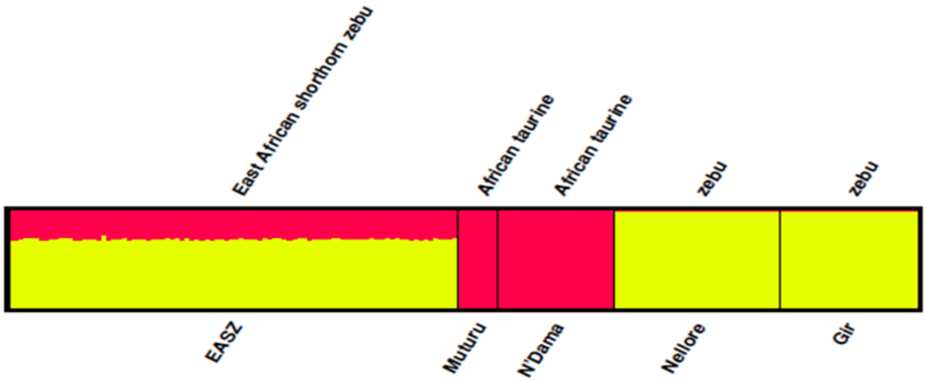

Figure 3.8: STRUCTURE bar plot of African taurine (Muturu and N'Dama) and Asian zebu (Nellore and Gir) genetic membership proportions for EASZ autosomes (K=2).

LAMP software 2.4 (Sankararaman *et al*., 2008) estimated the mean Asian zebu ancestry proportion for all SNPs in EASZ samples to be 0.76 (SD = 0.14). Based on this estimation, the mean ΔAZ in this analysis was 0 (SD = 0.14).

The majority of the candidate sweep regions showed high zebu ancestry proportion, but similar to the mean autosomal Asian zebu ancestry proportion (< +/- 1 SD) estimated by LAMP. In the 22 East and West African zebu-sharing candidate regions, six demonstrated substantial ΔAZ (≥ +/- 1 SD from the mean ΔAZ). Two of these regions showed deficiencies while four showed excesses of Asian zebu ancestry. For the 35 East African zebu-sharing candidate regions, eight revealed substantial ΔAZ, a single region showed deficiency and seven showed excesses (Table 3.1). When the genome-wide SNP and *Hp* analyses overlapping candidate regions were considered, 14 regions demonstrated substantial ΔAZ, one deficiency and 13 excesses (Table 3.5).

Discussion

Candidate genomics regions under positive selection

There are now several examples of genome-wide analyses for detection of signatures of selection in human, livestock and domestic animal populations (Sabeti *et al*., 2002, Gautier *et al*., 2009, Grossman *et al*., 2010, Rubin *et al*., 2010, Gautier and Naves, 2011, Petersen *et al*., 2013, Carneiro *et al*., 2014). These selection signals may be related to adaptation to environmental constraints (e.g., altitude and pathogens) (Flori *et al*., 2012, Ai *et al*., 2015), response to human selection pressures (Hayes *et al*., 2008, Flori *et al*., 2009, Qanbari *et al*., 2014), or even speciation and its consequence on the genome (Baker and Bradley, 2006, Nei and Nozawa, 2011). For African livestock, there have been few studies (Gautier *et al*., 2009, Flori *et al*., 2014), and until now, no East African zebu populations have been investigated.

In this chapter, we have attempted to comprehensively identify, in an ancient zebu-taurine crossbreed, candidate signatures of selection. For this purpose, we used genome-wide high-density SNP analyses, high coverage full genome sequence analysis and targeted exome sequence analysis. As mentioned in Chapter 2, the choice of EASZ population is of interest given the known history of the population's enduring natural and human selective forces in East African cattle (Rege *et al*., 2001). Also, the admixed zebu-taurine genome of this population (Mbole-Kariuki *et al*., 2014) may be another selective pressure influencing different biological pathways, e.g., immunity, fertility and development.

We have successfully identified many candidate regions of positive selection. A major issue in such studies is the presence of false positives randomly resulting from genetic drift (Akey *et al*., 2002, Qanbari and Simianer, 2014). As in Chapter 2, we attempted to minimize the effect of this issue through the comparison of our results with other previous studies in tropical-adapted admixed cattle populations, e.g., Creole cattle from Guadeloupe (Gautier and Naves, 2011), zebu-taurine admixed cattle from West Africa (Gautier *et al*., 2009, Flori *et al*., 2014, Xu *et al*., 2015), and in zebu cattle, e.g., Brahman (Ramey *et al*., 2013, Xu *et al*., 2015) and Gir (Liao *et al*., 2013, Perez O'Brien *et al*., 2014). Most importantly, analysing new zebu-taurine populations from East (Uganda) and West (Nigeria) Africa and identifying overlapping candidate regions is adding further support to our findings. This cross-validation has allowed us to narrow down the sizes of the identified candidate genome regions (Table 3.2), which may facilitate the identification of the selected causative genomic variants under selection.

Our composite statistical approach followed that of Utsunomiya *et al*. (2013b), which combines the power of the different genome-wide SNP analyses into a single test. The higher number of candidate regions defined by the EASZ *meta-SS* analysis in comparison to the individual tests supports the approach followed here (Table S3.3). This *meta-SS* analysis is different from that of the Composite of Multiple Signals (CMS) method used by Grossman *et al*. (2010), which gives a composite likelihood statistic to each variant by combining the *P*-values from five different tests. CSM was applied to simulated and empirical human genomic data from HapMap II (Grossman *et al*., 2010). Although CMS is an effective tool to localize candidate regions and to identify causal variants under selection in human genome, the method requires coalescent simulation and, hence, accurate calibrated demographic models to compute the likelihood tables (Qanbari and Simianer, 2014) that are deficient for our EASZ population.

As in Chapter 2, the reference cattle populations were pooled into a single population. Here we want to increase the reference population sample size and, hence, develop more robust EHH and allele frequency estimations as well as to breakdown population-specific LD resulting from genetic drift. A major improvement in this chapter is the analysis of the EASZ full genome sequence for signatures of selection. This provides high SNP coverage and addresses any breed ascertainment bias associated with commercially available SNP chips (Matukumalli *et al*., 2009).

Interestingly, a subset of the identified candidate regions demonstrated substantial ∆AZ (see results section). Most of these regions show excesses of Asian zebu ancestry indicating that the indicine haplotypes are more likely to be under selection in the African admixed cattle populations. This is perhaps not surprising considering the predominant zebu genomic background in the EASZ (Mbole-Kariuki *et al*., 2014; Figure 3.8), but also considering that zebu cattle were initially domesticated in the northern part of the Indian subcontinent (Chen *et al*., 2010) in an environment likely sharing more similarities with the African environment than the centre of domestication of taurine cattle in the Near East does (Loftus *et al*., 1994, Bradley *et al*., 1996). The pre-adaptation of zebu cattle may have facilitated its introgression into the local taurine populations. However, the haplotype origin assignment using the SNP array utilized in this study was only partly successful. This issue of the zebu or taurine origin of the selected SNPs or haplotypes will likely be better approached once a full indicine *de novo* reference genome is available.

Finally, it should not be forgotten that while we analysed the signatures of selection shared across three sets of populations here, many candidate signatures are population (e.g., EASZ) or geographic regions (e.g., East African)-specific. Whether or not these signals are the legacy of local selective forces remains unknown.

Potential biological functions under positive selection

The functional enrichment analyses performed using DAVID on the genes within the candidate regions emphasises the significantly enriched biological pathways under selection in EASZ. They are related to e.g., bovine adaptive and innate immunity, response to hormone stimuli, intermediate filaments and keratins, growth and development (Tables 3.3 and 3.4). Additionally, genes related to regulation of bovine immunity, fertility and reproduction, anatomical development and heat stress have also been found within these candidate regions.

Genes related to the innate and adaptive immune responses may be expected to be primary targets of selection in African tropical cattle that are exposed to a diversity of diseases and associated physiological stresses in their surrounding environment, e.g., endoparasites, haemoparasites and bacteria (de Clare Bronsvoort *et al*., 2013, Murray *et al*., 2013, Thumbi *et al*., 2014). Examples of these genes include the following: C-C chemokine receptor type 7 precursor (*CCR7*) and granulocyte macrophage-colony stimulating factor (*GM-CSF*), which are both identified within an EASZ candidate region on BTA 19 (40.96-41.45 Mb).

C-C chemokine receptor type 7 (CCR7) is sensitive to two types of chemo-attractants: CCL19 and CCL21. This receptor is involved in maturating antigen-presenting cells (dendritic cells). The mature dendritic cells will, in turn, activate T lymphocytes upon infection (Marsland *et al*., 2005, Forster *et al*., 2008). Moreover, this receptor has also demonstrated a role in regulating innate immunity by attracting macrophages to sites of infection (van Zwam *et al*., 2010). GM-CSF is a multifunctional cytokine with an important role in regulating innate immunity. This molecule acts as a positive regulator for macrophages to induce their antimicrobial and tumoricidal effects (Grabstein *et al*., 1986, Tarr, 1996).

It is interesting to point out that this functional category is being reported as significantly enriched in our three comparisons (EASZ, UGN, NGR zebu populations) as well as in other studies (Gautier *et al*., 2009, Flori *et al*., 2014), indicating that the immune response is an adaptation to the environment. Having a generic signal does not mean that the selected polymorphisms will be the same across populations, but rather and perhaps more likely that the same regions and network may be under parallel selection in different populations.

Genes related to fertility and reproduction may also be subjected to selection in African cattle populations. Examples of these candidate genes include the following: the retinoic acid receptor α (*RARA*) in an EASZ candidate region within BTA 19 (40.96-41.45 Mb). The retinoic acid receptor, which is expressed in sertoli cells in the seminiferous tubules, plays a role in transducing retinoic acid signal to maintain spermatogonia differentiation. Knocking out this receptor in experimental mice has led to defects in spermatogenesis and male sterility (Wolgemuth and Chung, 2007).

Members of the olfactory receptor gene family have been identified within *Hp* candidate regions BTA1: 42.1-42.22 Mb and BTA5: 59.5-59.6 Mb. This gene family, which is expressed in mammalian male germ cells (Vanderhaeghen *et al*., 1993, Spehr *et al*., 2003), might play a role in directing sperm to the oocyte during fertilization (Spehr *et al*., 2003, Fukuda *et al*., 2004, Guidobaldi *et al*., 2012).

More examples of genes in this category are spermatogenesis-associated 24 (*SPATA24*) within the BTA 7 (51.4 – 53.4 Mb) and sperm-associated antigen 7 (*SPAG7*). The *SPAG7* is mapped within one of the East and West African zebu-sharing candidate regions (BTA 19: 26.89-27.15 Mb). Here we are dealing with genes directly involved with male fertility. It may be argued that fertility may be linked to climatic adaptation (Hansen, 2004). Alternatively, we cannot exclude that such a signal may be the consequence of the hybridization between two cattle lineages (zebu and taurine). This requires further investigation.

Also, several examples of candidate genes identified within the candidate sweep regions are linked to biological roles associated with anatomical development. This type of gene can be considered as a target of selection to maintain optimum growth and development for an admixed cattle population. They could have been selected by human and/or the environment. *Man-1* (inner nuclear membrane protein 1) presents on an East and West African zebu-sharing candidate region (BTA 5: 48.61-49.02 Mb), which has demonstrated a critical role in heart development as discovered *via* knock-out experimental mice (Ishimura *et al*., 2008). This candidate genome region was also detected to be under selection in Brahman (Ramey *et al*., 2013) and Gir (Perez O'Brien *et al*., 2014) cattle. This region is located within a larger genomic interval associated with several traits, such as coat colour and penile sheath, in Brahman and tropical composite cattle (Porto-Neto *et al*., 2014). Moreover, a gene mapped within an EASZ candidate region within BTA 7 (33.1-33.29 Mb) *LOX* (Lysine-6-oxidase precursor) has also been linked to the development of various tissues, such as lung tissue and blood vessels (Maki *et al*., 2005).

The *RXFP2* (relaxin/insulin-like family peptide receptor 2) within an EASZ candidate region (BTA12: 29.11 – 29.44 Mb) may be classified as a candidate gene for anatomical development, fertility and reproduction categories. *RXFP2* has a role in testicular descent development, which is an adaptive physiology to maintain optimum reproduction and sperm quality when the core body temperature reaches about 34^{o}C (Gorlov *et al*., 2002, Agoulnik, 2007, Park *et al*., 2008, Feng *et al*., 2009). Interestingly, this gene has also been mapped within a candidate region in admixed Creole cattle (Gautier and Naves, 2011) and Gir zebu cattle (Liao *et al*., 2013). Although the gene has mainly been associated with the horn phenotype in sheep (Johnston *et al*., 2011, Kijas *et al*., 2012), a study by Johnston *et al*. (2013) has demonstrated an association between variants of this gene and reproductive success and survival rate in Soay sheep from St. Kilda.

Interestingly, the median ΔAZ value of this candidate region demonstrates a substantial deficiency of zebu ancestry (Table 3.5), a detailed investigation of the SNP-specific local zebu ancestry proportions within this region indicates different genomic fragments with substantial excesses or deficiencies of zebu ancestry, which is a good representation of the complexity we face when trying to define the selective haplotype origins in these candidate regions.

Genes related to the heat shock protein family have also been mapped in two EASZ candidate regions: BTA 7: 51.36-53.36 Mb and BTA 19: 42.89-43.12 Mb, e.g., *HSPA9* and *HSPB9* (heat shock proteins A9 and B9). This gene family has demonstrated a critical role in maintaining proteins folding and structure under stress (Parsell and Lindquist, 1994, Coleman *et al*., 1995). Members of the DnaJ family (DnaJ homolog subfamily C members 8 and 18), which act as cofactors for the heat shock protein 70 (Kampinga and Craig, 2010), were also found within an East African zebu-sharing candidate region (BTA 2: 125.64-126.08 Mb) and an EASZ candidate region (BTA 7: 51.36-53.36 Mb). The candidate region within BTA 7 (51.36-53.36 Mb) is the largest size sweep region showing a substantial excess of Asian zebu ancestry (Table 3.5). This region overlaps with sweep regions identified in tropically adapted Gir cattle (Liao *et al*., 2013) and West African cattle (Gautier *et al*., 2009), and showed high divergence between zebu and taurine (Porto-Neto *et al*., 2013).

Gene desert candidate regions

About 15% of the identified candidate regions were gene desert regions, i.e., no reported protein-coding or RNA genes; for example, in the East and West African zebu-sharing candidate region within BTA19 (2.5-2.7 Mb). Because of the incomplete annotation of the bovine genome, these regions may harbour transcription factor-binding sites or genes, which are not yet annotated in the UMD3.1 bovine reference genome.

Overlapping with QTL

Several QTL were located within the defined candidate sweep regions (Table S3.16). On addition to fertility and reproduction, and immune system regulation QTL, which are discussed above, interestingly we have observed some overlaps between candidate regions for positive selection in EASZ and some previously reported trypanotolerance QTL (Hanotte *et al*., 2003). The level of trypanotolerance displayed by the EASZ is unknown and undocumented. Our results might suggest some level of trypanotolerance in EASZ, as it has already been shown in other East African cattle populations (e.g., Orma Boran, Sheko and Mursi cattle) (Dolan, 1987, Mwangi *et al*., 1993, Bahbahani and Hanotte, 2015). The tolerant alleles of the BTA 7, 26 and 27 QTL are of taurine origins, whilst the parasite detection rate QTL on BTA 13 is of zebu origin (Hanotte *et al*., 2003). This may indicate the adaptive role of the EASZ admixed genome.

Also, several of the identified candidate regions overlapped with regions under positive selection in commercial dairy and beef breeds (Table 3.1, Table 3.5 and Table S3.13) (Larkin *et al*., 2012, Qanbari *et al*., 2014, Kemper *et al*., 2014). In addition, various production traits QTL, e.g., marbling score, milk fat percentage and milk fat yield, have also been found within the candidate regions. Perhaps an illustration that human selection for some production traits, e.g., milk yield, has taken place in EASZ at least in the past. It is also possible that positive selection on genes with a pleiotropic effect and/or linkage disequilibrium between loci involved in different metabolic pathways, rather than a common selection pressure, explains the overlapping candidate genome regions observed between EASZ and commercial breeds.

Putative causative variants of the candidate genes

The candidate genes selected in this chapter harbour many variants (SNPs and indels) that can be considered as causative mutations under selection.

Interestingly, 84.5% of the SNPs and 61% of the indels were specific to EASZ. Moreover, CNVs may be considered as clear targets of selection. However, the conducted depth of coverage analysis only indicates qualitative signals of CNVs that need to be further validated experimentally.

The identification of four non-synonymous putative causative variants, with two of them have probable damaging effects (Table 3.7), within three candidate genes is highly interesting. These variants, which have not been identified in other cattle breeds based on data available in the dbSNP database, could be the target of selection within the genome of EASZ. Although none of these variants exhibit fixation for the alternative alleles (Table 3.7), further experimental study of these variants is critical; for example, checking their frequencies in larger EASZ samples and in different cattle populations. Variants in the non-selected genes within the candidate regions (i.e., not candidate genes) should also be considered in future work as possible targets of selection. Other type of variants, such as variants at the 5' and 3' UTR of a gene or in non-coding regions, are also important and can play a role in regulating the expression of protein-coding genes. These variants may be argued as the probable sites under selection, as suggested by Carneiro *et al*. (2014).

Because of the different genomic affects, i.e., alteration of gene dosage and/or disrupting regulation of gene expression (Schuster-Bockler *et al*., 2010, Zhou *et al*., 2011), CNVs might be considered as targets of selection (Grossman *et al*., 2013). This type of polymorphism has been found to cover ~ 2.1% of the bovine genome (~55.6 Mb) (Bickhart *et al*., 2012), and has been linked to different traits, e.g., susceptibility to gastrointestinal nematodes in Angus cattle (Hou *et al*., 2012).

The depth of coverage analysis performed on the 10 EASZ exome sequences revealed signals of multiple copies at 17 genes within 17 candidate sweep regions overlapping between EASZ genome-wide SNP analyses and *Hp* analysis (Table S3.20). Interesting examples are *Man-1* and *RXFP2* previously linked to tropical adaptation in African cattle through their roles in anatomical development and male fertility. The putative CNVs observed on their exons, across all 10 EASZ samples, could be the result of selection besides the previously mentioned non-synonymous variants (Table 3.7). Moreover, signal of CNV was also found at *HBS1L*, a member of the GTP-binding elongation factor family, which might also be under selection in EASZ.

Genome coverage enrichment bias might result in some exome targets when the libraries were prepared for sequencing. This issue may produce false signals of CNV. Therefore, all putative CNVs will need to be confirmed by qRT-PCR analyses in different EASZ samples. Our approach in defining CNVs also did not correct for possible GC content bias in the genome that can lead to high read coverage in GC-rich regions (Dohm *et al*., 2008). However, it is unlikely that this is an issue here, as our exome data did not show any GC bias signal with a mean GC content of 48% and 10% standard deviation (Figure S3.8).

A single CNV region (multiple copies) (BTA19: 42.97 Mb–42.98 Mb) specific to the Nellore cattle genome detected by Bickhart *et al*. (2012) is overlapping with a selective sweep region identified on EASZ. This CNV is within a transcription factor (STAT5B). Our depth of coverage analysis did not reveal any signal of multiple copies on this gene indicating that any CNV, if present in the EASZ genome as well, might be in this case in the intronic region of the gene.

Origin of the selection signatures

Undoubtedly, the genome of EASZ has been shaped throughout its history by different selection pressures. The first cause of selection was stoked by the effects of domestication on the wild aurochs *Bos primigenius*, around 10,000 year BP (Loftus *et al*., 1994, Bradley *et al*., 1996) at the origins of the two ancestral genomes (taurine and zebu). This was followed by selection pressure during the formation of this admixed cattle population and adaptation to its local environment. The former may have started as early as 2000-3000 years ago.

The statistical methods applied here identify selective sweep originating from different time points (*Rsb* and *iHS* < 30,000 years, while *Hp* < 200,000 years) (Oleksyk *et al*., 2010, Utsunomiya *et al*., 2013b) (See Table 1.1 in the introduction chapter). This timeframe spans the domestication period for cattle (~ 10,000 years BP). So it possible that some of the signatures of selection will be of ancient origin, dating perhaps even from the ancestral auroch populations. Given the unavailability of the *Bos primigenius* genome, this hypothesis remains difficult to test. Other signatures of selection may be more recent. Some might date from the African taurine population present on the continent prior the arrival of zebu, others may have followed the arrival of zebu, the zebu-taurine admixture and the subsequent adaptation to the local environment.

Conclusion

The analysis of genome-wide SNP data and the full genome sequence in EASZ has revealed 101 and 165 candidate regions for positive signatures of selection. Thirty-five were detected by both approaches. The study of other African zebu x taurine admixed cattle from Uganda and Nigeria has further confirmed the candidate signatures within East African cattle and among East and West African cattle populations. The majority of these regions show similar proportions of Asian zebu ancestry, as the average estimated autosomal zebu ancestry (71%), but in a few cases ΔAZ values suggest taurine or indicine origin for the selected haplotypes. These regions harbour genes and QTL associated to different selective pressures.

Unfortunately, we were not able to identify any causative variants under selection in EASZ. Further studies are needed in this context. These will need to include sequencing and/or genotyping of more cattle populations indigenous to different environments (tropical and temperate-adapted, admixed and pure). Also, it is possible that most of the causative variants may be present in regulatory regions and/or be the results of chromosomal rearrangements. The characterization of CNV in the exome of EASZ is the first step in this direction.

References

ADZHUBEI, I. A., SCHMIDT, S., PESHKIN, L., RAMENSKY, V. E., GERASIMOVA, A., BORK, P., KONDRASHOV, A. S. & SUNYAEV, S. R. 2010. A method and server for predicting damaging missense mutations. *Nat. Methods,* 7**,** 248-9.

AGOULNIK, A. I. 2007. Relaxin and related peptides in male reproduction. *Adv. Exp. Med. Biol.,* 612**,** 49-64.

AI, H., FANG, X., YANG, B., HUANG, Z., CHEN, H., MAO, L., ZHANG, F., ZHANG, L., CUI, L., HE, W., YANG, J., YAO, X., ZHOU, L., HAN, L., LI, J., SUN, S., XIE, X., LAI, B., SU, Y., LU, Y., YANG, H., HUANG, T., DENG, W., NIELSEN, R., REN, J. & HUANG, L. 2015. Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nat. Genet.,*47**,** 217-25.

AKEY, J. M., ZHANG, G., ZHANG, K., JIN, L. & SHRIVER, M. D. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.,* 12**,** 1805-14.

AMARAL, A. J., FERRETTI, L., MEGENS, H. J., CROOIJMANS, R. P., NIE, H., RAMOS-ONSINS, S. E., PEREZ-ENCISO, M., SCHOOK, L. B. & GROENEN, M. A. 2011. Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PLoS One,* 6**,** e14782.

AULCHENKO, Y. S., RIPKE, S., ISAACS, A. & VAN DUIJN, C. M. 2007. GenABEL: an R library for genome-wide association analysis. *Bioinformatics,* 23**,** 1294-6.

BAHBAHANI, H. & HANOTTE, O. 2015. Genetic resistance – tolerance to vector-borne diseases, prospect and challenges of genomics. *OIE Scientific and Technical Review,* 34**,** 185-97.

BAKER, R. J. & BRADLEY, R. D. 2006. Speciation in Mammals and the Genetic Species Concept. *J. Mammal.,* 87**,** 643-662.

BERMAN, A. 2011. Invited review: Are adaptations present to support dairy cattle productivity in warm climates? *J. Dairy Sci.,* 94**,** 2147-2158.

BICKHART, D. M., HOU, Y., SCHROEDER, S. G., ALKAN, C., CARDONE, M. F., MATUKUMALLI, L. K., SONG, J., SCHNABEL, R. D., VENTURA, M., TAYLOR, J. F., GARCIA, J. F., VAN TASSELL, C. P., SONSTEGARD, T. S., EICHLER, E. E. & LIU, G. E. 2012. Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res.,* 22**,** 778-90.

BRADLEY, D. G., MACHUGH, D. E., CUNNINGHAM, P. & LOFTUS, R. T. 1996. Mitochondrial diversity and the origins of African and European cattle. *PNAS,* 93**,** 5131-5.

CARNEIRO, M., RUBIN, C. J., DI PALMA, F., ALBERT, F. W., ALFOLDI, J., BARRIO, A. M., PIELBERG, G., RAFATI, N., SAYYAB, S., TURNER-MAIER, J., YOUNIS, S., AFONSO, S., AKEN, B., ALVES, J. M., BARRELL, D., BOLET, G., BOUCHER, S., BURBANO, H. A., CAMPOS, R., CHANG, J. L., DURANTHON, V., FONTANESI, L., GARREAU, H., HEIMAN, D., JOHNSON, J., MAGE, R. G., PENG, Z., QUENEY, G., ROGEL-GAILLARD, C., RUFFIER, M., SEARLE, S., VILLAFUERTE, R., XIONG, A., YOUNG, S., FORSBERG-NILSSON, K., GOOD, J. M., LANDER, E. S., FERRAND, N., LINDBLAD-TOH, K. & ANDERSSON, L. 2014. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science,* 345**,** 1074-9.

CHEN, S., LIN, B. Z., BAIG, M., MITRA, B., LOPES, R. J., SANTOS, A. M., MAGEE, D. A., AZEVEDO, M., TARROSO, P., SASAZAKI, S., OSTROWSKI, S., MAHGOUB, O., CHAUDHURI, T. K., ZHANG, Y. P., COSTA, V., ROYO, L. J., GOYACHE, F., LUIKART, G., BOIVIN, N., FULLER, D. Q., MANNEN, H., BRADLEY, D. G. &

BEJA-PEREIRA, A. 2010. Zebu cattle are an exclusive legacy of the South Asia neolithic. *Mol. Biol. Evol.,* 27**,** 1-6.

COLEMAN, J. S., HECKATHORN, S. A. & HALLBERG, R. L. 1995. Heat-shock proteins and thermotolerance: linking molecular and ecological perspectives. *Trends Ecol. Evol.,* 10**,** 305-6.

DE CLARE BRONSVOORT, B. M., THUMBI, S. M., POOLE, E. J., KIARA, H., AUGUET, O. T., HANDEL, I. G., JENNINGS, A., CONRADIE, I., MBOLE-KARIUKI, M. N., TOYE, P. G., HANOTTE, O., COETZER, J. A. & WOOLHOUSE, M. E. 2013. Design and descriptive epidemiology of the Infectious Diseases of East African Livestock (IDEAL) project, a longitudinal calf cohort study in western Kenya. *BMC Vet. Res.,* 9**,** 171-192.

DOHM, J. C., LOTTAZ, C., BORODINA, T. & HIMMELBAUER, H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.,* 36**,** e105.

DOLAN, R. B. 1987. Genetics and trypanotolerance. *Parasitol. Today,* 3**,** 137-43.

ELSIK, C. G., TELLAM, R. L., WORLEY, K. C., GIBBS, R. A., MUZNY, D. M., WEINSTOCK, G. M., ADELSON, D. L., EICHLER, E. E., ELNITSKI, L., GUIGO, R., HAMERNIK, D. L., KAPPES, S. M., LEWIN, H. A., LYNN, D. J., NICHOLAS, F. W., REYMOND, A., RIJNKELS, M., SKOW, L. C., ZDOBNOV, E. M., SCHOOK, L., WOMACK, J., ALIOTO, T., ANTONARAKIS, S. E., ASTASHYN, A., CHAPPLE, C. E., CHEN, H. C., CHRAST, J., CAMARA, F., ERMOLAEVA, O., HENRICHSEN, C. N., HLAVINA, W., KAPUSTIN, Y., KIRYUTIN, B., KITTS, P., KOKOCINSKI, F., LANDRUM, M., MAGLOTT, D., PRUITT, K., SAPOJNIKOV, V., SEARLE, S. M., SOLOVYEV, V., SOUVOROV, A., UCLA, C., WYSS, C., ANZOLA, J. M., GERLACH, D., ELHAIK, E., GRAUR, D., REESE, J. T., EDGAR, R. C., MCEWAN, J. C., PAYNE, G. M., RAISON, J. M., JUNIER, T., KRIVENTSEVA, E. V., EYRAS, E., PLASS, M., DONTHU, R., LARKIN, D. M., REECY, J., YANG, M. Q., CHEN, L., CHENG, Z., CHITKO-MCKOWN, C. G., LIU, G. E., MATUKUMALLI, L. K., SONG, J., ZHU, B., BRADLEY, D. G., BRINKMAN, F. S., LAU, L. P., WHITESIDE, M. D., WALKER, A., WHEELER, T. T., CASEY, T., GERMAN, J. B., LEMAY, D. G., MAQBOOL, N. J., MOLENAAR, A. J., SEO, S., STOTHARD, P., BALDWIN, C. L., BAXTER, R., BRINKMEYER-LANGFORD, C. L., BROWN, W. C., CHILDERS, C. P., CONNELLEY, T., ELLIS, S. A., FRITZ, K., GLASS, E. J., HERZIG, C. T., IIVANAINEN, A., LAHMERS, K. K., BENNETT, A. K., DICKENS, C. M., GILBERT, J. G., HAGEN, D. E., SALIH, H., AERTS, J., CAETANO, A. R., *et al.* 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science,* 324**,** 522-8.

FENG, S., FERLIN, A., TRUONG, A., BATHGATE, R., WADE, J. D., CORBETT, S., HAN, S., TANNOUR-LOUET, M., LAMB, D. J., FORESTA, C. & AGOULNIK, A. I. 2009. INSL3/RXFP2 signaling in testicular descent. *Ann. N. Y. Acad. Sci.,* 1160**,** 197-204.

FLICEK, P., AHMED, I., AMODE, M. R., BARRELL, D., BEAL, K., BRENT, S., CARVALHO-SILVA, D., CLAPHAM, P., COATES, G., FAIRLEY, S., FITZGERALD, S., GIL, L., GARCIA-GIRON, C., GORDON, L., HOURLIER, T., HUNT, S., JUETTEMANN, T., KAHARI, A. K., KEENAN, S., KOMOROWSKA, M., KULESHA, E., LONGDEN, I., MAUREL, T., MCLAREN, W. M., MUFFATO, M., NAG, R., OVERDUIN, B., PIGNATELLI, M., PRITCHARD, B., PRITCHARD, E., RIAT, H. S., RITCHIE, G. R., RUFFIER, M., SCHUSTER, M., SHEPPARD, D., SOBRAL, D., TAYLOR, K., THORMANN, A., TREVANION, S., WHITE, S., WILDER, S. P., AKEN, B. L., BIRNEY, E., CUNNINGHAM, F., DUNHAM, I., HARROW, J., HERRERO, J., HUBBARD, T. J., JOHNSON, N., KINSELLA, R., PARKER, A., SPUDICH, G., YATES, A., ZADISSA, A. & SEARLE, S. M. 2013. Ensembl 2013. *Nucleic Acids Res.,* 41**,** D48-55.

FLORI, L., FRITZ, S., JAFFREZIC, F., BOUSSAHA, M., GUT, I., HEATH, S., FOULLEY, J. L. & GAUTIER, M. 2009. The genome response to artificial selection: a case study in dairy cattle. *PLoS One,* 4**,** e6595.

FLORI, L., GONZATTI, M. I., THEVENON, S., CHANTAL, I., PINTO, J., BERTHIER, D., ASO, P. M. & GAUTIER, M. 2012. A quasi-exclusive European ancestry in the Senepol tropical cattle breed highlights the importance of the slick locus in tropical adaptation. *PLoS One,* 7**,** e36133.

FLORI, L., THEVENON, S., DAYO, G. K., SENOU, M., SYLLA, S., BERTHIER, D., MOAZAMI-GOUDARZI, K. & GAUTIER, M. 2014. Adaptive admixture in the West African bovine hybrid zone: insight from the Borgou population. *Mol. Ecol.,* 23**,** 3241-57.

FORSTER, R., DAVALOS-MISSLITZ, A. C. & ROT, A. 2008. CCR7 and its ligands: balancing immunity and tolerance. *Nat. Rev. Immunol.,* 8**,** 362-71.

FUKUDA, N., YOMOGIDA, K., OKABE, M. & TOUHARA, K. 2004. Functional characterization of a mouse testicular olfactory receptor and its role in chemosensing and in regulation of sperm motility. *J. Cell Sci.,* 117**,** 5835-45.

GAUTIER, M., FLORI, L., RIEBLER, A., JAFFREZIC, F., LALOE, D., GUT, I., MOAZAMI-GOUDARZI, K. & FOULLEY, J. L. 2009. A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. *BMC Genomics,* 10**,** 550.

GAUTIER, M. & NAVES, M. 2011. Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Mol. Ecol.,* 20**,** 3128-43.

GAUTIER, M. & VITALIS, R. 2012. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics,* 28**,** 1176-7.

GODDARD, M. E. & HAYES, B. J. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.,* 10**,** 381-91.

GORLOV, I. P., KAMAT, A., BOGATCHEVA, N. V., JONES, E., LAMB, D. J., TRUONG, A., BISHOP, C. E., MCELREAVEY, K. & AGOULNIK, A. I. 2002. Mutations of the GREAT gene cause cryptorchidism. *Hum. Mol. Genet.,* 11**,** 2309-18.

GRABSTEIN, K. H., URDAL, D. L., TUSHINSKI, R. J., MOCHIZUKI, D. Y., PRICE, V. L., CANTRELL, M. A., GILLIS, S. & CONLON, P. J. 1986. Induction of macrophage tumoricidal activity by granulocyte-macrophage colony-stimulating factor. *Science,* 232**,** 506-8.

GROSSMAN, S. R., ANDERSEN, K. G., SHLYAKHTER, I., TABRIZI, S., WINNICKI, S., YEN, A., PARK, D. J., GRIESEMER, D., KARLSSON, E. K., WONG, S. H., CABILI, M., ADEGBOLA, R. A., BAMEZAI, R. N., HILL, A. V., VANNBERG, F. O., RINN, J. L., LANDER, E. S., SCHAFFNER, S. F. & SABETI, P. C. 2013. Identifying recent adaptations in large-scale genomic data. *Cell,* 152**,** 703-13.

GROSSMAN, S. R., SHLYAKHTER, I., KARLSSON, E. K., BYRNE, E. H., MORALES, S., FRIEDEN, G., HOSTETTER, E., ANGELINO, E., GARBER, M., ZUK, O., LANDER, E. S., SCHAFFNER, S. F. & SABETI, P. C. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science,* 327**,** 883-6.

GUIDOBALDI, H. A., TEVES, M. E., UNATES, D. R. & GIOJALAS, L. C. 2012. Sperm transport and retention at the fertilization site is orchestrated by a chemical guidance and oviduct movement. *Reproduction,* 143**,** 587-96.

HANOTTE, O., RONIN, Y., AGABA, M., NILSSON, P., GELHAUS, A., HORSTMANN, R., SUGIMOTO, Y., KEMP, S., GIBSON, J., KOROL, A., SOLLER, M. & TEALE, A. 2003. Mapping of quantitative trait loci controlling trypanotolerance in a cross of tolerant West African N'Dama and susceptible East African Boran cattle. *PNAS,* 100**,** 7443-8.

HANSEN, P. J. 2004. Physiological and cellular adaptations of zebu cattle to thermal stress. *Anim. Reprod. Sci.,* 82-83**,** 349-60.

HAYES, B., CHAMBERLAIN, A., MACEACHERN, S., SAVIN, K., MCPARTLAN, H., MACLEOD, I., SETHURAMAN, L. & GODDARD, M. 2008. A genome map of divergent artificial selection between *Bos taurus* dairy cattle and *Bos taurus* beef cattle. *Anim. Genet.,* 40**,** 176 - 184.

HOU, Y., LIU, G. E., BICKHART, D. M., MATUKUMALLI, L. K., LI, C., SONG, J., GASBARRE, L. C., VAN TASSELL, C. P. & SONSTEGARD, T. S. 2012. Genome regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. *Functional & Integrative Genomics,* 12**,** 81-92.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.,* 37**,** 1-13.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.,* 4**,** 44-57.

ISHIMURA, A., CHIDA, S. & OSADA, S. 2008. Man1, an inner nuclear membrane protein, regulates left-right axis formation by controlling nodal signaling in a node-independent manner. *Dev. Dyn.,* 237**,** 3565-76.

JAKOBSSON, M. & ROSENBERG, N. A. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics,* 23**,** 1801-6.

JOHNSTON, S., MCEWAN, J., PICKERING, N., KIJAS, J., BERALDI, D., PILKINGTON, J., PEMBERTON, J. & SLATE, J. 2011. Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. *Mol. Ecol.,* 20**,** 2555 - 2566.

JOHNSTON, S. E., GRATTEN, J., BERENOS, C., PILKINGTON, J. G., CLUTTON-BROCK, T. H., PEMBERTON, J. M. & SLATE, J. 2013. Life history trade-offs at a single locus maintain sexually selected genetic variation. *Nature,* 502**,** 93-5.

KAMPINGA, H. H. & CRAIG, E. A. 2010. The HSP70 chaperone machinery: J proteins as drivers of functional specificity. *Nat. Rev. Mol. Cell Biol.,* 11**,** 579-92.

KEIGHTLEY, P. D. & EYRE-WALKER, A. 2000. Deleterious mutations and the evolution of sex. *Science,* 290**,** 331-3.

KEMPER, K. E., SAXTON, S. J., BOLORMAA, S., HAYES, B. J. & GODDARD, M. E. 2014. Selection for complex traits leaves little or no classic signatures of selection. *BMC Genomics,* 15**,** 246.

KIJAS, J. W., LENSTRA, J. A., HAYES, B., BOITARD, S., PORTO NETO, L. R., SAN CRISTOBAL, M., SERVIN, B., MCCULLOCH, R., WHAN, V., GIETZEN, K., PAIVA, S., BARENDSE, W., CIANI, E., RAADSMA, H., MCEWAN, J. & DALRYMPLE, B. 2012. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol,* 10**,** e1001258.

KINSELLA, R. J., KAHARI, A., HAIDER, S., ZAMORA, J., PROCTOR, G., SPUDICH, G., ALMEIDA-KING, J., STAINES, D., DERWENT, P., KERHORNOU, A., KERSEY, P. & FLICEK, P. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford),* 2011**,** bar030.

LARKIN, D. M., DAETWYLER, H. D., HERNANDEZ, A. G., WRIGHT, C. L., HETRICK, L. A., BOUCEK, L., BACHMAN, S. L., BAND, M. R., AKRAIKO, T. V., COHEN-ZINDER, M., THIMMAPURAM, J., MACLEOD, I. M., HARKINS, T. T., MCCAGUE, J. E., GODDARD, M. E., HAYES, B. J. & LEWIN, H. A. 2012. Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. *PNAS,* 109**,** 7693-8.

LIAO, X., PENG, F., FORNI, S., MCLAREN, D., PLASTOW, G. & STOTHARD, P. 2013. Whole genome sequencing of Gir cattle for identifying polymorphisms and loci under selection. *Genome,* 56**,** 592-8.

LOFTUS, R. T., MACHUGH, D. E., BRADLEY, D. G., SHARP, P. M. & cunningham, p. 1994. Evidence for 2 independent domestications of cattle. *PNAS,* 91**,** 2757-2761.

MAKI, J. M., SORMUNEN, R., LIPPO, S., KAARTEENAHO-WIIK, R., SOININEN, R. & MYLLYHARJU, J. 2005. Lysyl oxidase is essential for normal development and function of the respiratory system and for the integrity of elastic and collagen fibers in various tissues. *Am. J. Pathol.,* 167**,** 927-36.

MARSLAND, B. J., BATTIG, P., BAUER, M., RUEDL, C., LASSING, U., BEERLI, R. R., DIETMEIER, K., IVANOVA, L., PFISTER, T., VOGT, L., NAKANO, H., NEMBRINI, C., SAUDAN, P., KOPF, M. & BACHMANN, M. F. 2005. CCL19 and CCL21 induce a potent proinflammatory differentiation program in licensed dendritic cells. *Immunity,* 22**,** 493-505.

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., SONSTEGARD, T. S. & VAN TASSELL, C. P. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One,* 4**,** e5350.

MBOLE-KARIUKI, M. N., SONSTEGARD, T., ORTH, A., THUMBI, S. M., BRONSVOORT, B. M., KIARA, H., TOYE, P., CONRADIE, I., JENNINGS, A., COETZER, K., WOOLHOUSE, M. E., HANOTTE, O. & TAPIO, M. 2014. Genome-wide analysis reveals the ancient and recent admixture history of East African Shorthorn Zebu from Western Kenya. *Heredity (Edinb),* 113**,** 297-305.

MCKAY, S. D., SCHNABEL, R. D., MURDOCH, B. M., MATUKUMALLI, L. K., AERTS, J., COPPIETERS, W., CREWS, D., DIAS NETO, E., GILL, C. A., GAO, C., MANNEN, H., STOTHARD, P., WANG, Z., VAN TASSELL, C. P., WILLIAMS, J. L., TAYLOR, J. F. & MOORE, S. S. 2007. Whole genome linkage disequilibrium maps in cattle. *BMC Genet.,* 8**,** 74.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.,* 20**,** 1297-303.

METZKER, M. L. 2010. Sequencing technologies - the next generation. *Nat. Rev. Genet.,* 11**,** 31-46.

MIKAWA, S., MOROZUMI, T., SHIMANUKI, S., HAYASHI, T., UENISHI, H., DOMUKAI, M., OKUMURA, N. & AWATA, T. 2007. Fine mapping of a swine quantitative trait locus for number of vertebrae and analysis of an orphan nuclear receptor, germ cell nuclear factor (NR6A1). *Genome Res.,* 17**,** 586-93.

MURRAY, G. G., WOOLHOUSE, M. E., TAPIO, M., MBOLE-KARIUKI, M. N., SONSTEGARD, T. S., THUMBI, S. M., JENNINGS, A. E., VAN WYK, I. C., CHASE-TOPPING, M., KIARA, H., TOYE, P., COETZER, K., DEC BRONSVOORT, B. M. & HANOTTE, O. 2013. Genetic susceptibility to infectious disease in East African Shorthorn Zebu: a genome-wide analysis of the effect of heterozygosity and exotic introgression. *BMC Evol. Biol.,* 13**,** 246-253.

MWANGI, E. K., STEVENSON, P., GETTINBY, G. & MURRAY, M. Variation in susceptibility to tsetse-borne trypanosomiasis among *Bos indicus* cattle breeds in East Africa. *In:* ROWLANDS, G. J. & TEALE, A. J., eds. Towards increased use of trypanotolerance: current research and future directions, 1993 Nairobi, Kenya. 81-86.

NEI, M. & NOZAWA, M. 2011. Roles of mutation and selection in speciation: from Hugo de Vries to the modern genomic era. *Genome Biol. Evol.,* 3**,** 812-29.

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 Res.,* 18**,** 1282-93.

OLEKSYK, T. K., SMITH, M. W. & O'BRIEN, S. J. 2010. Genome-wide scans for footprints of natural selection. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.,* 365**,** 185-205.

PARK, J. I., SEMYONOV, J., CHANG, C. L., YI, W., WARREN, W. & HSU, S. Y. 2008. Origin of INSL3-mediated testicular descent in therian mammals. *Genome Res.,* 18**,** 974-85.

PARSELL, D. A. & LINDQUIST, S. 1994. *In The Biology of Heat Schock Proteins and Molecular Chaperones,* Cold Spring Harbor Laboratory Press.

PEREZ O'BRIEN, A. M., UTSUNOMIYA, Y. T., MESZAROS, G., BICKHART, D. M., LIU, G. E., VAN TASSELL, C. P., SONSTEGARD, T. S., DA SILVA, M. V., GARCIA, J. F. & SOLKNER, J. 2014. Assessing signatures of selection through variation in linkage disequilibrium between taurine and indicine cattle. *Genet. Sel. Evol.,* 46**,** 19.

PETERSEN, J. L., MICKELSON, J. R., RENDAHL, A. K., VALBERG, S. J., ANDERSSON, L. S., AXELSSON, J., BAILEY, E., BANNASCH, D., BINNS, M. M., BORGES, A. S., BRAMA, P., DA CAMARA MACHADO, A., CAPOMACCIO, S., CAPPELLI, K., COTHRAN, E. G., DISTL, O., FOX-CLIPSHAM, L., GRAVES, K. T., GUERIN, G., HAASE, B., HASEGAWA, T., HEMMANN, K., HILL, E. W., LEEB, T., LINDGREN, G., LOHI, H., LOPES, M. S., MCGIVNEY, B. A., MIKKO, S., ORR, N., PENEDO, M. C., PIERCY, R. J., RAEKALLIO, M., RIEDER, S., ROED, K. H., SWINBURNE, J., TOZAKI, T., VAUDIN, M., WADE, C. M. & MCCUE, M. E. 2013. Genome-wide analysis reveals selection for important traits in domestic horse breeds. *PLoS Genet.,* 9**,** e1003211.

PORTO-NETO, L. R., REVERTER, A., PRAYAGA, K. C., CHAN, E. K., JOHNSTON, D. J., HAWKEN, R. J., FORDYCE, G., GARCIA, J. F., SONSTEGARD, T. S., BOLORMAA, S., GODDARD, M. E., BURROW, H. M., HENSHALL, J. M., LEHNERT, S. A. & BARENDSE, W. 2014. The genetic architecture of climatic adaptation of tropical cattle. *PLoS One,* 9**,** e113284.

PORTO-NETO, L. R., SONSTEGARD, T. S., LIU, G. E., BICKHART, D. M., DA SILVA, M. V., MACHADO, M. A., UTSUNOMIYA, Y. T., GARCIA, J. F., GONDRO, C. & VAN TASSELL, C. P. 2013. Genomic divergence of zebu and taurine cattle identified through high-density SNP genotyping. *BMC Genomics,* 14**,** 876.

PRITCHARD, J. K., STEPHENS, M. & DONNELLY, P. 2000. Inference of population structure using multilocus genotype data. *Genetics,* 155**,** 945-59.

QANBARI, S., PAUSCH, H., JANSEN, S., SOMEL, M., STROM, T. M., FRIES, R., NIELSEN, R. & SIMIANER, H. 2014. Classic selective sweeps revealed by massive sequencing in cattle. *PLoS Genet.,* 10**,** e1004148.

QANBARI, S. & SIMIANER, H. 2014. Mapping signatures of positive selection in the genome of livestock. *Livestock Science,* 166**,** 133-143.

QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-2.

RAMEY, H. R., DECKER, J. E., MCKAY, S. D., ROLF, M. M., SCHNABEL, R. D. & TAYLOR, J. F. 2013. Detection of selective sweeps in cattle using genome-wide SNP data. *BMC Genomics,* 14**,** 382.

R Development Core Team 2012. R: A language and environment for statistical computing. Vienna, Austria.

REGE, J. E. O., KAHI, A., M., O.-A., MWACHARO, J. & HANOTTE, O. 2001. *Zebu cattle of Kenya: Uses, performance, farmer preferences and measures of genetic diversity*. Nairobi, Kenya, International Livestock Reaserch Institute.

RINCON, G., WEBER, K. L., EENENNAAM, A. L., GOLDEN, B. L. & MEDRANO, J. F. 2011. Hot topic: performance of bovine high-density genotyping platforms in Holsteins and Jerseys. *J Dairy Sci,* 94**,** 6116-21.

ROSENBERG, N. A. 2004. DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes,* 4**,** 137-138.

RUBIN, C. J., MEGENS, H. J., MARTINEZ BARRIO, A., MAQBOOL, K., SAYYAB, S., SCHWOCHOW, D., WANG, C., CARLBORG, O., JERN, P., JORGENSEN, C. B., ARCHIBALD, A. L., FREDHOLM, M., GROENEN, M. A. & ANDERSSON, L. 2012. Strong signatures of selection in the domestic pig genome. *PNAS,* 109**,** 19529-36.

RUBIN, C. J., ZODY, M. C., ERIKSSON, J., MEADOWS, J. R., SHERWOOD, E., WEBSTER, M. T., JIANG, L., INGMAN, M., SHARPE, T., KA, S., HALLBOOK, F., BESNIER, F., CARLBORG, O., BED'HOM, B., TIXIER-BOICHARD, M., JENSEN, P., SIEGEL, P., LINDBLAD-TOH, K. & ANDERSSON, L. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature,* 464**,** 587-91.

SABETI, P. C., REICH, D. E., HIGGINS, J. M., LEVINE, H. Z., RICHTER, D. J., SCHAFFNER, S. F., GABRIEL, S. B., PLATKO, J. V., PATTERSON, N. J., MCDONALD, G. J., ACKERMAN, H. C., CAMPBELL, S. J., ALTSHULER, D., COOPER, R., KWIATKOWSKI, D., WARD, R. & LANDER, E. S. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature,* 419**,** 832-7.

SANKARARAMAN, S., SRIDHAR, S., KIMMEL, G. & HALPERIN, E. 2008. Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.,* 82**,** 290-303.

SANTANA, M. H., UTSUNOMIYA, Y. T., NEVES, H. H., GOMES, R. C., GARCIA, J. F., FUKUMASU, H., SILVA, S. L., OLIVEIRA JUNIOR, G. A., ALEXANDRE, P. A., LEME, P. R., BRASSALOTI, R. A., COUTINHO, L. L., LOPES, T. G., MEIRELLES, F. V., ELER, J. P. & FERRAZ, J. B. 2014. Genome-wide association analysis of feed intake and residual feed intake in Nellore cattle. *BMC Genet.,* 15**,** 21.

SCHEET, P. & STEPHENS, M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.,* 78**,** 629-44.

SCHUSTER-BOCKLER, B., CONRAD, D. & BATEMAN, A. 2010. Dosage sensitivity shapes the evolution of copy-number varied regions. *PLoS One,* 5**,** e9474.

SHERRY, S. T., WARD, M. H., KHOLODOV, M., BAKER, J., PHAN, L., SMIGIELSKI, E. M. & SIROTKIN, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.,* 29**,** 308-11.

SPEHR, M., GISSELMANN, G., POPLAWSKI, A., RIFFELL, J. A., WETZEL, C. H., ZIMMER, R. K. & HATT, H. 2003. Identification of a testicular odorant receptor mediating human sperm chemotaxis. *Science,* 299**,** 2054-8.

STOUFFER, S. A., SUCHMAN, E. A., DEVINNEY, L. C., STAR, S. A. & WILLIAMS, R. M. 1949. *The American soldier, Vol. 1: Adjustment during Army Life* Princeton, Princeton University Press.

SU, G., BRØNDUM, R. F., MA, P., GULDBRANDTSEN, B., AAMAND, G. P. & LUND, M. S. 2012. Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. *J. Dairy Sci.,* 95**,** 4657-4665.

TANG, K., THORNTON, K. R. & STONEKING, M. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.,* 5**,** e171.

TARR, P. E. 1996. Granulocyte-macrophage colony-stimulating factor and the immune system. *Med. Oncol.,* 13**,** 133-40.

THUMBI, S. M., BRONSVOORT, B. M., POOLE, E. J., KIARA, H., TOYE, P. G., MBOLE-KARIUKI, M. N., CONRADIE, I., JENNINGS, A., HANDEL, I. G., COETZER, J. A., STEYL, J. C., HANOTTE, O. & WOOLHOUSE, M. E. 2014. Parasite co-infections and their impact on survival of indigenous cattle. *PLoS One,* 9**,** e76324.

TIJJANI, A. 2013. *Genome-wide characterization of diversity and admixture in African cattle breeds using high density SNP markers.* MSc, University of Nottingham.

UTSUNOMIYA, Y. T., DO CARMO, A. S., CARVALHEIRO, R., NEVES, H. H., MATOS, M. C., ZAVAREZ, L. B., PEREZ O'BRIEN, A. M., SOLKNER, J., MCEWAN, J. C., COLE, J. B., VAN TASSELL, C. P., SCHENKEL, F. S., DA SILVA, M. V., PORTO NETO, L. R., SONSTEGARD, T. S. & GARCIA, J. F. 2013a. Genome-wide association study for birth weight in Nellore cattle points to previously described orthologous genes affecting human and bovine height. *BMC Genet.,* 14**,** 52.

UTSUNOMIYA, Y. T., PEREZ O'BRIEN, A. M., SONSTEGARD, T. S., VAN TASSELL, C. P., DO CARMO, A. S., MESZAROS, G., SOLKNER, J. & GARCIA, J. F. 2013b. Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. *PLoS One,* 8**,** e64280.

VAN ZWAM, M., WIERENGA-WOLF, A. F., MELIEF, M. J., SCHRIJVER, B., LAMAN, J. D. & BOVEN, L. A. 2010. Myelin ingestion by macrophages promotes their motility and capacity to recruit myeloid cells. *J Neuroimmunol.,* 225**,** 112-7.

VANDERHAEGHEN, P., SCHURMANS, S., VASSART, G. & PARMENTIER, M. 1993. Olfactory receptors are displayed on dog mature sperm cells. *J. Cell Biol.,* 123**,** 1441-52.

VOIGHT, B. F., KUDARAVALLI, S., WEN, X. & PRITCHARD, J. K. 2006. A map of recent positive selection in the human genome. *PLoS Biol.,* 4**,** e72.

WHITLOCK, M. C. 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.,* 18**,** 1368-73.

WOLGEMUTH, D. J. & CHUNG, S. S. 2007. Retinoid signaling during spermatogenesis as revealed by genetic and metabolic manipulations of retinoic acid receptor alpha. *Soc. Reprod. Fertil. Suppl.,* 63**,** 11-23.

XU, L., BICKHART, D. M., COLE, J. B., SCHROEDER, S. G., SONG, J., VAN TASSELL, C. P., SONSTEGARD, T. S. & LIU, G. E. 2015. Genomic signatures reveal new evidences for selection of important traits in domestic cattle. *Mol. Biol. Evol.,* 32**,** 711-25.

ZHOU, J., LEMOS, B., DOPMAN, E. B. & HARTL, D. L. 2011. Copy-number variation: the balance between gene dosage and expression in Drosophila melanogaster. *Genome Biol. Evol.,* 3**,** 1014-24.

Chapter four

Signatures of positive selection in the East African shorthorn zebu sex chromosome (BTA X)

Abstract

The lower effective population size, recombination and mutation rates of the sex chromosome X in comparison to the autosomes are the main reasons to analyse this chromosome separately from the autosomes. In this chapter, 22,778 single nucleotide polymorphisms on the X chromosome and the full chromosome sequence of 10 pooled EASZ were analysed with two Extended Haplotype Homozygosity (EHH)-based (*Rsb* and *iHS*) indices and a pooled heterozygosity (*Hp*) index, respectively, to identify candidate signatures of positive selection. Admixture analysis indicated a similar zebu–taurine ancestry compared to autosomes, though within-population homogeneity is significantly lower (Mann-Whitney U test; *P*-value < 2.2×10^{-16}). A total of 20 candidate regions (six by *Rsb*, two by *iHS*, and twelve by *Hp* analysis) were identified. Two were also present in zebu cattle populations from Uganda, while two additional were found in Nigerian zebu cattle. None are shared across all populations. Seven regions showed substantial deviation from the mean chromosome X zebu ancestry (≥ 1 standard deviation from the mean), indicative of an indicine (n = 3) or a taurine (n = 4) origin. A total of 49 genes were identified within the 20 candidate regions and five considered as candidate genes based on their biological roles (e.g. immune response and fertility). Five non-synonymous variants and two putative copy number variants represent possible causative polymorphisms. These potential targets of positive selection might be related to the adaptation of EASZ to their surrounding environment and/or the admixed genomic structure of EASZ.

Introduction

In Chapter 3, the bovine sex chromosome (BTA X) was not analysed with the autosomes, given its genetic characteristics, particularly its lower effective population size (Ne), three quarters of bovine autosomal Ne (Schaffner, 2004), and its lower recombination rate compared to autosomes (e.g., two thirds of the genome average in humans (Kong *et al*., 2002)). Additionally, since BTA X spends two thirds of its history in females, which show a lower mutation rate than males (Bohossian *et al*., 2000), the genetic diversity in this chromosome is expected to be lower than in autosomes (Schaffner, 2004). Apart from the pseudoautosomal region (PAR) shared between the X and Y chromosomes, no recombination occurs in this chromosome in males (Van Laere *et al*., 2008, Graves, 2010). All of these factors mean that population genetic forces (selection, genetic drift and population structure) acting on the BTA X will have different dynamics compared to autosomes (Harris and Hey, 1999, Ebersberger *et al*., 2002, Schaffner, 2004, Amato *et al*., 2009, Yang *et al*., 2014).

BTA X contains genes with different biological roles. Some of these genes are associated with bovine immunity (e.g., interleukin 13 receptor alpha 2 (*IL13RA2*) and interleukin 2 receptor gamma (*IL2RG*)), cattle reproduction and fertility (e.g., foetal and adult testis expressed 1 (*FATE1*)) and brain development (e.g., brain expressed X-linked 2 (*BEX2*)). Due to the difference in number of X chromosome copies between sexes, genes demonstrate dosage compensation, i.e., doubling expression in somatic tissues with random inactivation of a single copy in females, to balance the expression of autosomal genes (Nguyen and Disteche, 2006, Graves, 2006, Graves, 2010).

Several studies on different livestock species have defined candidate regions on their genomes subjected to positive selection (natural and artificial) (Gautier *et al*., 2009, Chan *et al*., 2010, Gautier and Naves, 2011, Flori *et al*., 2012, Rubin *et al*., 2012, Ai *et al*., 2013, Yang *et al*., 2014). However, the sex chromosome has rarely been investigated by these studies. Recently, in pigs, several genome regions on chromosome X were identified to be under positive selection (Yang *et al*., 2014, Ma *et al*., 2014), with candidate genes associated with haematological traits, reproduction, immunity and meat quality reported in a study by Ma *et al*. (2014). In cattle, Chan *et al*. (2010) have identified three candidate genome regions on BTA X with signatures of positive selection in zebu and taurine cattle. Several genes related to immunity, e.g., *IL2R2* and *IL2RG*, solute carriers, e.g., *SLC35A2* and *SLC7A3*, were considered candidates. *IL2RG* has also been reported to be under diversifying selection when zebu cattle were compared to taurine breeds (Porto-Neto *et al*., 2013).

In this chapter, we aim to understand the genetic structure of the EASZ BTA X. We also try to detect candidate genome regions with signatures of positive selection on this chromosome and to identify putative causative mutations (SNPs and CNVs) for the detected signatures.

Materials and Methods

Genotyped SNPs and cattle samples quality controls

SNPs mapped on BTA X (39,367 SNPs) in the Illumina BovineHD Genotyping BeadChip (Rincon *et al*., 2011) were filtered out *via* the *check.marker* function of the GenABEL package (Aulchenko *et al*., 2007) in R software version 2.15.1 (R Development Core Team, 2012). Prior to quality control (QC), possibly duplicate samples based on the autosomal identity-by-state (IBS) QC step (see Materials and Methods section in Chapter 3) were eliminated. SNPs with minor allele frequency (MAF) less than 5% (2,043 SNPs) and genotyping call rate less than 95% (1,194 SNPs) were excluded. Some of the pruned out SNPs (104 SNPs) did not pass both of the two criteria, so we ended up with 36,234 SNPs. From this final SNP list, we included only markers with their ancestral allelic status determined previously by Utsunomiya *et al*. (2013), leaving a total of 22,778 SNPs (mean gap size = 6.5 kb, median gap size = 3.6 kb, SD = 19.3 kb). Based on the genotyping data of reference cattle populations, about 1,000 SNPs might be in the PAR.

Cattle samples genotyped with less than 95% of the SNPs (two Red bororo samples) were also excluded from the analyses. Thus, the total number of samples analysed were EASZ (n = 92), Ugandan (UGN) zebu cattle (n = 77) (25 Ankole (AO), 16 Karamojong zebu (KR), 23 Nanda (NG), 13 Serere zebu (ZS)), Nigerian

(NGR) zebu cattle (n = 99) (23 Adamawa Gudali (AG), 1 Azawak (AZ), 22 Bunaji (BJ), 20 Red bororo (OR), 19 Sokoto Gudali (SO), 2 Wadara (WD) and 12 Yakanaji (YK)), 8 Muturu (MT), 24 N'Dama (NDM), 59 Holstein-Friesian (HOL), 32 Jersey (JER), 34 Nellore (NEL) and 28 Gir (GIR).

EASZ BTA X admixture analysis

Admixture analysis *via* a Bayesian clustering method implemented in STRUCTURE software version 2.3 (Pritchard *et al*., 2000) was conducted on BTA X for the EASZ, NDM, MT, NEL and GIR. Three independent replicates of an admixed model with independent allele frequencies were run for a burn-in period of 25,000 iterations and 50,000 Markov Chain Monte Carlo steps for K = 2. The mean output file was generated using *CLUMPP* software version 1.1.2 (Jakobsson and Rosenberg, 2007) and graphically displayed by *Distruct* software version 1.1 (Rosenberg, 2004).

Extended Haplotype Homozygosity (EHH)-based methods (*Rsb* and *iHS*)

Rsb analyses (Tang *et al*., 2007) were performed on each of the African zebu-taurine admixed cattle populations (Tijjani, 2013, Mbole-Kariuki *et al*., 2014) (EASZ, combined UGN zebu cattle populations and combined NGR zebu cattle populations) and on the combined reference cattle populations using the *rehh* package (Gautier and Vitalis, 2012) of the R software (R Development Core Team, 2012). *iHS* (Voight *et al*., 2006) analyses were carried on EASZ, combined UGN zebu cattle and combined NGR zebu cattle populations using the *rehh* package of the R software.

Given the normal distribution of the standardised values of the tests (Figure S4.1), one- and two-tailed Z tests were applied as described in Chapter 3 (Materials and Methods section) considering - $\log_{10}$ (*P*-value) = 3, which corresponds to a *P*-value equal to 0.001, as a significant threshold. At least three consecutive significant SNPs (not separated by more than 500 kb) were required to specify candidate region intervals.

Selective sweep analysis (pooled heterozygosity *Hp*)

The pooled heterozygosity (*Hp*) of the SNPs identified in the EASZ pooled BTA X sequence (see Material and Methods section in Chapter 3) were calculated on 100 kb sliding windows with 10 kb incremental steps. A ZHp ≤ - 4 was applied as a threshold to specify windows carrying a selective sweep as in the study by Liao *et al*. (2013). Overlapping candidate windows were merged into a single region as described (Rubin *et al*., 2010, Liao *et al*., 2013).

Functional characterization of the candidate regions

Genes within the identified candidate regions (BTA X only and autosomal with BTA X combined), obtained from the *Ensembl Genes 73* database (Flicek *et al*., 2013), were processed using the functional annotation tool implemented in *DAVID* Bioinformatics resources 6.7 to determine the over-represented (enriched) functional terms (Huang da *et al*., 2009a, Huang da *et al*., 2009b). An enrichment score of 1.3, which is equivalent to the Fisher exact test *P*-value = 0.05, was used as a threshold to define the significantly enriched functional terms in comparison to the whole bovine reference genome background.

Variants (SNPs and indels) in the genes were annotated using the variant effect predictor tool on the Ensembl website (Flicek *et al.*, 2013). Comparisons with the previously discovered bovine variants listed in the dbSNP database (Sherry *et al*., 2001) (http://www.ncbi.nlm.nih.gov/SNP/) classified these variants into EASZ-specific and general bovine variants. The biological effects of the candidate non-synonymous variants were predicted by the PolyPhen-2 online tool (Adzhubei *et al*., 2010).

The bovine Quantitative Trait Loci (QTL) (http://www.animalgenome.org/cgi-bin/QTLdb/BT/index) spanning the candidate regions were also identified using the *intersectBed* function from the *BedTools* software (Quinlan and Hall, 2010).

Copy Number Variation (CNV) analysis

Putative CNVs (multiple copies) were identified in the 10 sequenced EASZ BTA X exome using the same approach described previously (Chapter 3). Figure S4.2 demonstrates the positive skew of the normalized depth of coverage values. An SDOC value equal to three was arbitrarily chosen to define targeted exons with multiple copies. The GC content of the ten EASZ BTA X exome sequences was calculated using GATK (McKenna *et al*., 2010).

Estimation of excess or deficiency in Asian zebu ancestry at candidate regions

LAMP software version 2.4 (Sankararaman *et al*., 2008) was used to estimate the Asian zebu and African taurine ancestry proportions of the genotyped SNPs on EASZ samples as in Chapter 3. The BTA X zebu ancestry proportion (68%) and African taurine ancestry proportion (32%) of EASZ were obtained from the admixture proportions α of the STRUCTURE analysis. The median ΔAZ for the significant SNPs of EASZ *iHS* and *Rsb* analyses within candidate regions were considered. For the *Hp* sequence candidate regions, the median ΔAZ of all the SNPs within these regions were considered.

Results

EASZ BTA X admixture analysis

The admixture analysis conducted on EASZ BTA X estimated an average genetic proportion of 0.68 ± 0.085 SD and 0.32 ± 0.085 SD for Asian zebu and African taurine ancestries, respectively. As observed in Figure 4.1, the EASZ showed high variation in the Asian zebu ancestry proportions across animals, ranging from 0.43 to 0.84 in males and 0.49 to 0.79 in females. This variation is significantly higher than the within-population variation in zebu ancestry proportions observed in EASZ autosomes (see Results section in Chapter 3) (Mann-Whitney U test; *P*-value < 2.2×10^{-16}).

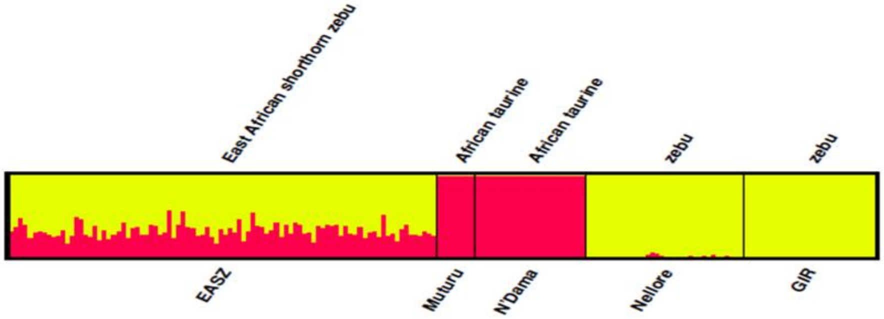

Figure 4.1: STRUCTURE bar plot of African taurine (Muturu and N’Dama) and Asian zebu (Nellore and Gir) genetic membership proportions for EASZ BTA X (K = 2).

Rsb and *iHS* analyses

The *Rsb* analyses resulted in six, two and five candidate regions on EASZ, UGN and NGR zebu cattle, respectively (Figure 4.2, Table 4.1 and S4.1). Whilst the *iHS* analyses indicated two, three and no candidate regions in EASZ, UGN zebu cattle and NGR zebu cattle (Figure 4.3, Table 4.1 and Table S4.1). Out of the total eight EASZ candidates, two were characterised as East African zebu-sharing (86.75 – 87.1 Mb and 64.64 – 64.8 Mb) (Table S4.1). A single region (68.41 – 68.49 Mb) was shared between EASZ and NGR zebu cattle.

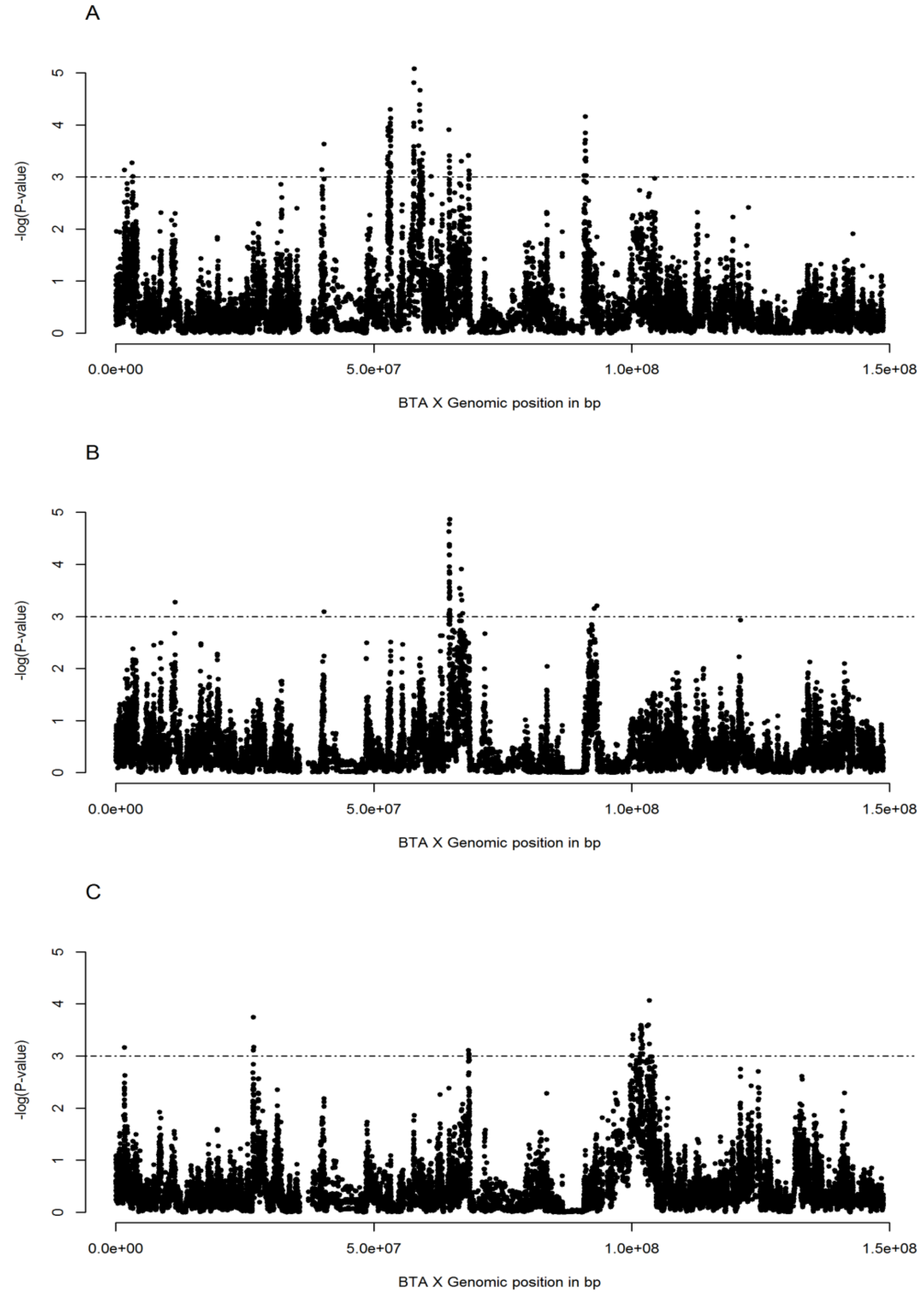

Figure 4.2: Manhattan plots of BTA X *Rsb* analyses between (A) EASZ, (B) zebu cattle from Uganda, (C) zebu cattle from Nigeria and all combined reference cattle populations combined (HOL, JER, NDM, MT, NEL and GIR) The significant threshold is at -$\log_{10}$ *P*-value = 3 (one-tailed Z-test).

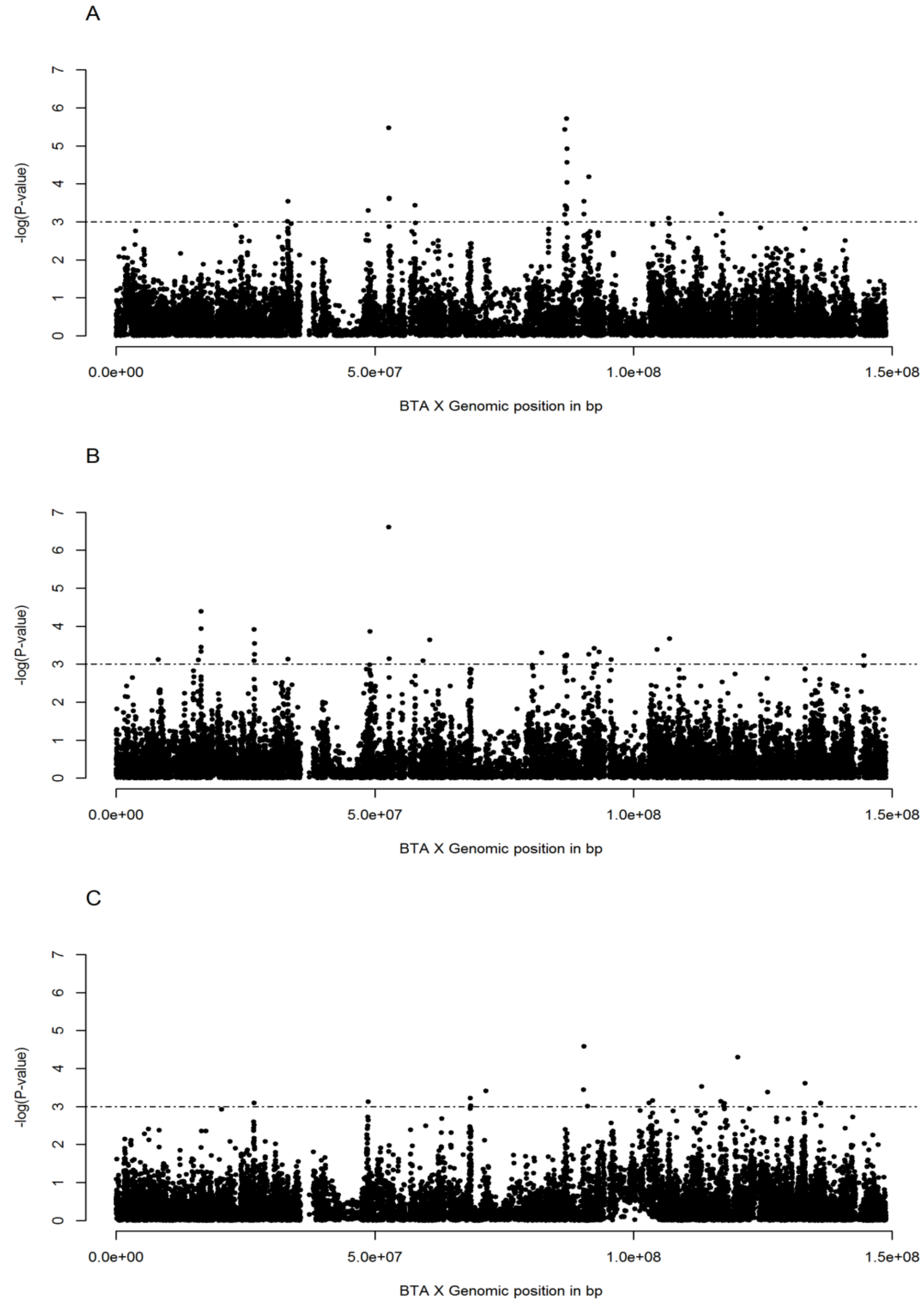

Figure 4.3: Manhattan plots of BTA X *iHS* analyses on (A) EASZ, (B) zebu cattle from Uganda, (C) zebu cattle from Nigeria. The significant threshold at -$\log_{10}$ *P*-value = 3 (two-tailed Z-tests).

Hp sweep analysis

Sequencing the full EASZ BTA X generated 27,925,062 reads with MAPQ ≥ 20. These reads covered ~ 95% of the reference UMD 3.1 BTA X with an average depth of coverage ~ 9 folds. A total of 309,050 SNPs (205,530 heterozygotes and 103,520 homozygotes) were identified. The distribution of the number of SNPs in the 100 kb windows (Figure S4.3a) was similar to the one obtained for the autosomal SNPs (Figure S3.7), with a mean of 183 SNPs per window. The resulting mean *Hp* value was 0.386 (SD = 0.073). A total of 205 windows out of 14,870 windows passed the ZHp threshold of -4 (~ 1% tail of the ZHp distribution) (Figure 4.4, S4.3b, and Table S4.2). These windows were merged into 12 candidate sweep regions. None of these regions overlapped with the EASZ candidate regions obtained by the *Rsb* and *iHS* analyses (Table 4.1). However, a single sweep region (100.11 – 100.37 Mb) overlapped with an *Rsb* candidate region identified in zebu cattle from Nigeria (Table S4.1).

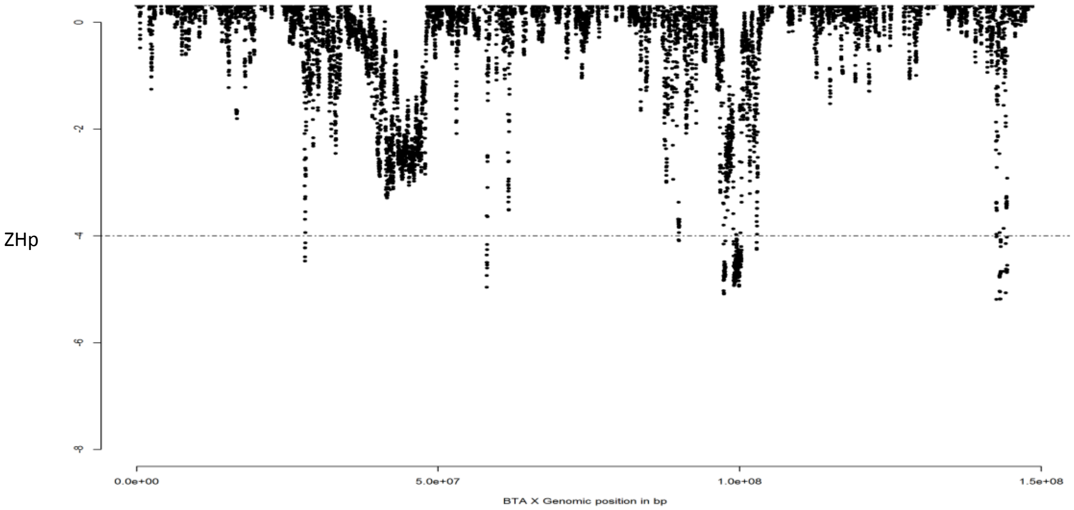

Figure 4.4: Manhattan plot of EASZ BTA X *Hp* analysis. Each point represents a 100 kb window. Threshold ZHp = -4.

Table 4.1: Candidate regions on BTA X of EASZ defined by *Rsb*, *iHS* and *Hp* analysis. ΔAZ: excess/deficiency of Asian zebu ancestry.

start	end	Value[1]	Median ΔAZ	start	end	Value[1]	Median ΔAZ	start	end	Value[1]	Median ΔAZ	Ref[2]
iHS				***Rsb***				***Hp***				
52,724,365	52,831,886	5.47	-0.016	**52,724,365 ¥**	**53,286,404**	**4.3**	**-0.152**	**27,900,001**	**28,042,935**	**-4.29**	**-0.141**	
86,750,029 *	87,174,478	5.71	0.114	**57,753,489**	**57,862,192**	**5.07**	**-0.158**	58,030,001	58,239,841	-4.52	NA	
				58,699,877 ¥	**59,561,015**	**4.67**	**-0.147**	89,870,001 ¥	89,983,990	-4.08	0.087	
				64,636,360 *	64,803,846	3.4	-0.049	**97,300,001 ¥**	**97,720,659**	**-4.67**	**0.12**	Porto-Neto *et al.* (2013)
				68,399,979¤	**68,541,416**	**3.42**	**0.147**	98,950,001	100,085,162	-4.55	0.114	Porto-Neto *et al.* (2013)
				90,841,068	91,235,638	4.1	-0.049	100,110,001¤	100,366,965	-4.42	0.114	Porto-Neto *et al.* (2013)
								102,800,001	**102,926,979**	**-4.21**	**0.174**	
								142,500,001	142,633,019	-4.61	NA	
								143,050,001	143,390,762	-4.78	NA	
								143,730,001	143,845,721	-4.67	0.054	
								144,080,001	144,257,602	-4.61	-0.0162	
								144,260,001	144,424,871	-4.54	NA	

[1] – $\log_{10}$ (*P*-value) of the most significant SNP in *iHS* and *Rsb* candidate regions. Mean ZHp value for the *Hp* candidate windows.

[2] previous studies identifying signatures of selection in bovine BTA X.

***** East African zebu-sharing regions

¤ Overlap with candidate regions in zebu cattle populations from Nigeria

¥ Overlap with CNVs identified on Nellore by Bickhart *et al.*, 2012.

Bold (deviation by more than +/- 1 SD from the mean ΔAZ)

NA: No genotyped SNPs on Illumina BovineHD Bead chip (after QC).

Functional characterization of the candidate regions

All the identified EASZ candidate regions, except four identified by *Hp* analysis (Table S4.3), contain genes based on the UMD3.1 reference genome annotation. A total of 49 genes were identified; 22 genes in *iHS* and *Rsb* candidate regions and 27 genes in *Hp* candidate regions (Tables 4.2 and S4.4). DAVID functional cluster analysis revealed three non-significantly enriched functional clusters: nucleotide binding (enrichment score = 0.99), non-membrane-bounded organelle (enrichment score = 0.46) and ion binding (enrichment score = 0.29). No enriched functional clusters were identified within the genes in the East African zebu-sharing regions and EASZ-Nigerian zebu cattle overlapping regions. No bovine QTL overlapped with any of the identified candidate regions. DAVID analyses were also conducted on genes i) within EASZ autosomal and BTA X candidate regions combined and ii) within EASZ BTA X and autosomal SNPs and *Hp* analyses overlapping candidate regions. A total of 149 functional term clusters were identified in the first analysis (Table S4.5), with seven significantly enriched (Table 4.3). The second analysis identified 25 functional term clusters (Table S4.5) with three significantly enriched (Table 4.3).

Table 4.2: Genes mapped within the identified EASZ *iHS*, *Rsb* and *Hp* BTA X candidate regions.

Gene Start (bp)	**Gene End (bp)**	**Gene Name**
iHS* and *Rsb		
52,793,107	52,794,418	ENSBTAG00000009457
53,200,774	53,201,043	ENSBTAG00000045756
57,744,030	57,756,166	MORF4L2
57,773,646	57,798,407	GLRA4
58,699,812	58,706,139	ESX1
58,997,460	58,998,076	ENSBTAG00000047884
59,131,168	59,134,535	ENSBTAG00000034644
64,404,549	64,659,083	PAK3*
64,684,214	64,708,869	CAPN6*
64,755,904	64,861,282	DCX*
68,428,002	68,491,176	LHFPL1¤
86,944,028	86,956,186	EFNB1*

87,050,780	87,083,091	STARD8*
90,831,727	90,850,346	UBA1
90,853,944	90,865,344	CDK16
90,868,233	90,882,828	USP11
90,997,109	91,000,794	ZNF157
91,031,761	91,044,478	ZNF41
91,141,062	91,141,158	5S_rRNA
91,180,863	91,191,308	ARAF
91,223,900	91,228,056	SYN1
91,232,235	91,236,073	TIMP1
Hp		
89,907,348	89,908,678	ENSBTAG00000047411
97,359,957	97,364,910	TSR2
97,364,883	97,402,545	FGD1
97,425,377	97,426,021	RAB21
97,432,890	97,461,194	GNL3L
97,604,641	97,605,063	RPL39
98,958,218	98,960,092	UBQLN2
99,440,294	99,442,061	SPIN2
99,507,839	99,509,617	SPIN2B
99,522,364	99,529,359	ENSBTAG00000038387
99,529,831	99,530,700	ENSBTAG00000045616
99,588,579	99,589,450	ENSBTAG00000047279
99,644,622	99,768,588	HEPH
99,894,566	99,933,646	VSIG4
99,936,305	99,936,412	bta-mir-223
100,070,406	100,162,454	MSN¤
100,346,827	100,366,716	LAS1L¤
102,851,711	102,885,917	ZNF674
143,103,987	143,315,918	ENSBTAG00000046123
143,125,004	143,125,384	ENSBTAG00000046518
143,364,826	143,375,658	DDX3Y
143,378,430	143,385,607	ENSBTAG00000047068
143,736,738	143,800,954	ENSBTAG00000048102
143,750,456	143,750,560	U6
143,779,555	143,779,660	U6
143,827,184	143,930,209	ENSBTAG00000000211
144,159,056	144,234,222	ENSBTAG00000045544

*****in East African zebu-sharing candidate regions

¤ in EASZ-Nigerian zebu cattle overlapping candidate regions

Table 4.3: Significantly enriched functional term clusters of genes mapped within (A) EASZ autosomal and BTA X candidate regions combined (B) EASZ BTA X and autosomal SNPs and *Hp* analyses of overlapping candidate regions.

A		B	
Functional term cluster	**Score***	**Functional term cluster**	**Score***
Intermediate protein filaments and keratin	3.13	Response to hormones stimuli (e.g., growth hormones)	1.87
Enzyme inhibitor activity	2.38	Regulation of lymphocytes activation and proliferation	1.39
Cell-cell adhesion	1.94	Cell-cell junction	1.35
Protein transport and localization	1.88		
Nuclear lumen and nucleoplasm	1.63		
Cytoskeleton	1.53		
Cell-cell junction	1.49		

*Enrichment score following DAVID analysis (a score equals to 1.3, equivalent to Fisher exact test *P*-value = 0.05, was used as a significant threshold).

Potential causative variants under selection

A total of 3,273 SNPs (1,971 EASZ-specific) and 153 indels (124 EASZ-specific) were detected within the genes in the candidate regions. After annotation, 31 non-synonymous mutations were identified in 15 genes (17 EASZ-specific mutations in 11 genes) (Table S4.6). Five genes (*MSN*, *ESX1*, *VSIG4*, *SPIN2* and *SPIN2B*) were considered as candidates given their functions putatively linked to adaptation to the African environment (see Discussion). Five EASZ-specific non-synonymous variants were identified within two of these candidate genes (Table 4.4). Based on the exome data of the 10 EASZ samples, none of the non-synonymous variants were completely, or even nearly, fixed for the alternative allele (Table 4.4)

Table 4.4: The reference and alternative allele frequencies for the non-synonymous variants at candidate genes. Exome data from 10 unrelated EASZ. Location (variant position in bp on BTA X), amino acid substitution (reference amino acid, residue, alternative amino acid).

Location (UMD 3.1)	Gene	Amino acid substitution	Reference allele frequency (%)	Alternative allele frequency (%)	Biological effect*
BTA X: 99522547	*SPIN2B*	V 255 M	T = 100%	C = 0%	Possibly damaging
BTA X: 99522715	*SPIN2B*	E 199 K	C = 80%	T = 20%	Benign
BTA X: 99523015	*SPIN2B*	H 99 N	T = 100%	G = 0%	Benign
BTA X: 99523039	*SPIN2B*	D 91 H	C = 100%	G = 0%	Probably damaging
BTA X: 99927182	*VSIG4*	T 160 A	A = 67%	G = 33%	Possibly damaging

*Based on PolyPhen-2 online tool (Adzhubei *et al*., 2010)

Putative CNVs in candidate regions

The median depth of coverage for the BTA X exome regions captured by the SureSelect XT target enrichment system ranged from 25.9 reads/bp to 43.63 reads/bp (SD = 38.3 reads/bp to 68.17 reads/bp) for males and 44.45 reads/bp to 59.8 reads/bp (SD= 69.29 reads/bp to 97 reads/bp) for females. The normalised depth of coverage showed similar values in each sex: 1.1 reads/bp (SD= 1.7 reads/bp) for females and 0.58 reads/bp (SD = 0.87 reads/bp to 0.95 reads/bp) for males (Table S4.7).

Two candidate *Hp* sweep regions (89.87 – 89.98 Mb and 97.3 – 97.72 Mb) showed signals of CNV. These signals were within two targeted regions in two genes: the first exon of an uncharacterised gene (ENSBTAG00000047411) and the 3’ end of *FGD1* (Table S4.8). Moreover, four of the EASZ candidate regions overlapped with CNV regions (multiple copies) on Nellore BTA X (Bickhart *et al*., 2012) (Table 4.1).

Excess/deficiency of Asian zebu ancestry

LAMP software 2.4 (Sankararaman *et al*., 2008) estimated the mean Asian zebu ancestry proportion for all the QC-filtered BTA X SNPs on EASZ samples to be 0.76 (SD = 0.12). Based on this estimation, the mean ΔAZ for all of these SNPs was 0 (SD = 0.12). A total of seven regions (four showed deficiencies and three showed excesses) revealed substantial ΔAZ (define as more than +/- 1 SD from the mean ΔAZ) (Table 4.1).

Discussion

This chapter focuses on assessing the genetic structure of EASZ BTA X. Following the work conducted in the previous two chapters, we aimed to identify signatures of positive selection and candidate putative causative mutations for these signals.

EASZ BTA X genetic admixture

EASZ is likely the outcome of male-mediated Asian zebu introgression into Africa and crossbreeding with the native African taurine (Hanotte *et al.*, 2002) (see also Chapter 5). Sex ratio in adult domestic cattle is highly skewed with more cows than bulls. Consequently, it may be expected that the process of zebu introgression on the X chromosome was slower compared to that of autosomes. Our results indicate that EASZ BTA X shows Asian zebu and African taurine genetic admixture at comparable proportions to what has been estimated in autosomes (Figure 4.1 and Figure 3.8), supporting an ancient admixture (Decker *et al.*, 2014). However, in contrast to autosomes, EASZ BTA X indicates a lower level of admixture homogeneity across animals (Mann-Whitney U test; *P*-value < $2.2x10^{-16}$), perhaps a consequence of a lower recombination rate in sex chromosome in comparison to autosomes (Schaffner, 2004). It means that more time might be needed for fixation of the beneficial introgressed genomic blocks.

Signatures of selection on EASZ BTA X

As for the EASZ autosomes, the sex chromosome also shows signatures of positive selection with a total of 20 candidate regions identified. Three out of these regions overlap with genetically differentiated regions between taurine and zebu cattle on BTA X (Porto-Neto *et al*., 2013), while four regions were also identified in zebu cattle from Uganda and Nigeria (Table 4.1). These observations support the role of positive selection, rather than genetic drift, in shaping the diversity of these candidate regions.

Only a single candidate region overlaps between *iHS* and *Rsb* results. The lack of overlap between the three conducted analyses (*iHS, Rsb* and *Hp*) might be explained as further detailed in the discussion sections of Chapters 2 and 3, based on the different algorithms used for each test, as well as the different selection time-scales targeted by these approaches.

A critical issue that needs to be considered in this type of analysis is the inaccurate assembly of the bovine BTA X. Previous observations have demonstrated male heterozygosity in multiple locations on BTA X, suggesting more than one pseudoautosomal region (PAR) (personal communication with Dr. Tad Sonstegard, USDA-ARS Maryland). Given that only a single PAR is confirmed to be present on BTA X (Van Laere *et al*., 2008, Das *et al*., 2009), these observations may indicate that the BTA X assembly has portions of its genome region misplaced; in other words, the BTA assembly is not fully correct.

Interestingly, we have noted a close balance between the numbers of candidate regions with substantial excesses (three regions) and deficiencies (four regions) in Asian zebu ancestry. Therefore, both indicine and taurine haplotypes might be under selection here, supporting the hypothesis of adaptive admixture in EASZ.

The potential biological pathways under selection

The functional annotation clustering analyses conducted on the genes within the BTA X candidate regions alone did not indicate significantly enriched functional clusters. However, five candidate genes related to the regulation of the immune system (e.g., *VSIG4* and *MSN*) and cattle reproduction and fertility (e.g., *ESX1*, *SPIN2* and *SPIN2B*) were found.

VSIG4 (V-set and Ig domain-containing 4) is a macrophage complement receptor, which has a role in regulating innate immunity *via* the suppression of inflammatory responses (Helmy *et al*., 2006), and adaptive immunity by reducing T-cell response induction (Vogt *et al*., 2006). This gene has been previously considered as a candidate gene under diversifying selection between taurine and zebu cattle (Porto-Neto *et al*., 2013). *MSN* (moesin) is a member of the ERM domain in the protein tyrosine phosphatase (PTP). This protein plays a role in regulating lymphocyte activity (Mustelin *et al*., 2005).

Positive selection on favourable advantageous variants within immunity-related genes (innate and adaptive) is very important for EASZ cattle to face the pathological challenges in their surrounding environment (Di Giulio *et al*., 2009, Thumbi *et al*., 2014, de Clare Bronsvoort *et al*., 2013, Bahbahani and Hanotte, 2015). We have identified a single EASZ-specific non-synonymous variant with a possible damaging effect on *VSIG4*, based on PolyPhen-2 prediction. Although the alternative allele frequency in the 10 EASZ exome sequences is not high (0.33), further genotyping of this variant in EASZ and other African cattle populations is required before drawing any conclusion regarding the putative adaptive role of this variant.

The two spindlin family member genes (*SPIN2* and *SPIN2B*) are associated with gamete generation and sexual reproduction following the gene ontology specified in *DAVID* Bioinformatics resources 6.7. More specifically, these encoded proteins regulate the progression of the oocyte cycle after fertilization (Oh *et al*., 1997). These genes were also previously considered as candidates under selection by Porto-Neto *et al*. (2013). Another gene that may be classified into the cattle fertility and sexual reproduction category is the ESX homeobox 1 (*ESX1*) gene, initially called *Spx1*. This gene is expressed in adult mice and human testes (Branford *et al*., 1997, Li *et al*., 1997, Fohn and Behringer, 2001). Although *ESX1* is not essential for spermatogenesis in mice (Li and Behringer, 1998), it is involved in controlling spermatogenesis in primates by regulating male germ cell division (Ozawa *et al*., 2004). This explains the signature of positive selection identified on this gene during primate evolution (Wang and Zhang, 2007). Four EASZ-specific non-synonymous variants were identified in *SPIN2B*, with two predicted to have damaging effects on gene function (Table 4.4). As for the *VSIG4* non-synonymous variant, they need to be validated in a large number of samples and populations prior to consideration as targets of positive selection.

Several housekeeping genes associated with cell cycle progression, e.g. *CDK16* (cyclin-dependent kinase 16) (Grana and Reddy, 1995), and mRNA splicing processing, e.g. *U6* (Lamond, 1991), have also been found within the candidate regions. These genes may have been targeted by selection to optimize the cellular machinery in EASZ.

Last but not least, several uncharacterised genes were found within the candidate regions, e.g. 143.05 – 144.42 Mb, which covers the bovine PAR (Van Laere *et al*., 2008, Das *et al*., 2009). These genes need to be further characterised to understand their biological roles. Moreover, due to the limited annotation of the bovine reference genome, gene desert regions (Table S4.3) might also reveal valuable unknown features under selection, e.g. unannotated genes and transcription factor binding sites.

Putative CNVs in the candidate regions

The depth of coverage analysis showed qualitative signals of CNV regions overlapping with the identified candidate selection regions. This is clearly noticed on the *Hp* candidate region 89.87 – 89.98 Mb (Table S4.8). This signal spans a very large exon (856 bp) in an uncharacterised gene. The size of this targeted region is larger than the mean size of the regions targeted by the Agilent SureSelect^{XT} target enrichment kit (~ 250 bp). Because of the possible enrichment bias associated with the sequencing of this exon, this signal needs to be further validated experimentally using qRT-PCR. Although the BTA X exome sequence does not show a signal of GC content bias with a mean GC content of 46% and 9% standard deviation (Figure S4.4), the above-mentioned region demonstrates higher GC content (59%) than the mean, questioning the validity of the CNV (Dohm *et al*., 2008). For unknown reasons, both signals identified show higher standardised depth of coverage in females than in males (Table S4.8).

Although four of the identified candidate regions overlap with CNV regions identified by Bickhart *et al*. (2012) on the BTA X of Nellore, these regions did not show clear CNV signals here. Thus, CNV may not be present in EASZ or was not present in the coding regions analysed here.

Conclusion

To our knowledge, this is the first time the BTA X of an indigenous African cattle population has been analysed for signatures of selection leading to the identification of 20 candidate regions, in which two are East African zebu-sharing. The logical follow-up of this chapter is to define the causative mutations under selection in BTA X. This will need, as a first step, genotyping of more cattle populations for the identified variants. Moreover, the availability of the zebu reference genome will be a great opportunity to accurately define the ancestral origin of the candidate haplotypes as well as to analyse sequence contigs not mapped on the reference taurine genome. Analysing the depth of coverage in separate EASZ full genome sequences, in parallel with qRT-PCR, will fully characterise the genome of EASZ for CNVs and validate the putative ones defined here.

References

ADZHUBEI, I. A., SCHMIDT, S., PESHKIN, L., RAMENSKY, V. E., GERASIMOVA, A., BORK, P., KONDRASHOV, A. S. & SUNYAEV, S. R. 2010. A method and server for predicting damaging missense mutations. *Nat Methods,* 7**,** 248-9.

AI, H., HUANG, L. & REN, J. 2013. Genetic diversity, linkage disequilibrium and selection signatures in chinese and Western pigs revealed by genome-wide SNP markers. *PLoS One,* 8**,** e56001.

AMATO, R., PINELLI, M., MONTICELLI, A., MARINO, D., MIELE, G. & COCOZZA, S. 2009. Genome-wide scan for signatures of human population differentiation and their relationship with natural selection, functional pathways and diseases. *PLoS One,* 4**,** e7927.

AULCHENKO, Y. S., RIPKE, S., ISAACS, A. & VAN DUIJN, C. M. 2007. GenABEL: an R library for genome-wide association analysis. *Bioinformatics,* 23**,** 1294-6.

BAHBAHANI, H. & HANOTTE, O. 2015. Genetic resistance – tolerance to vector-borne diseases, prospect and challenges of genomics. *OIE Scientific and Technical Review,* 34**,** 185-97.

BICKHART, D. M., HOU, Y., SCHROEDER, S. G., ALKAN, C., CARDONE, M. F., MATUKUMALLI, L. K., SONG, J., SCHNABEL, R. D., VENTURA, M., TAYLOR, J. F., GARCIA, J. F., VAN TASSELL, C. P., SONSTEGARD, T. S., EICHLER, E. E. & LIU, G. E. 2012. Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res.*, 22**,** 778-90.

BOHOSSIAN, H. B., SKALETSKY, H. & PAGE, D. C. 2000. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. *Nature,* 406**,** 622-5.

BRANFORD, W. W., ZHAO, G.-Q., VALERIUS, M. T., WEINSTEIN, M., BIRKENMEIER, E. H., ROWE, L. B. & POTTER, S. S. 1997. *Spx1*, a novel X-linked homeobox gene expressed during spermatogenesis. *Mechanisms of Development,* 65**,** 87-98.

CHAN, E. K., NAGARAJ, S. H. & REVERTER, A. 2010. The evolution of tropical adaptation: comparing taurine and zebu cattle. *Anim. Genet.,* 41**,** 467-77.

DAS, P. J., CHOWDHARY, B. P. & RAUDSEPP, T. 2009. Characterization of the bovine pseudoautosomal region and comparison with sheep, goat, and other mammalian pseudoautosomal regions. *Cytogenet. Genome Res.,* 126**,** 139-47.

DE CLARE BRONSVOORT, B. M., THUMBI, S. M., POOLE, E. J., KIARA, H., AUGUET, O. T., HANDEL, I. G., JENNINGS, A., CONRADIE, I., MBOLE-KARIUKI, M. N., TOYE, P. G., HANOTTE, O., COETZER, J. A. & WOOLHOUSE, M. E. 2013. Design and descriptive epidemiology of the Infectious Diseases of East African Livestock (IDEAL) project, a longitudinal calf cohort study in western Kenya. *BMC Vet. Res.,* 9**,** 171-192.

DECKER, J. E., MCKAY, S. D., ROLF, M. M., KIM, J., MOLINA ALCALA, A., SONSTEGARD, T. S., HANOTTE, O., GOTHERSTROM, A., SEABURY, C. M., PRAHARANI, L., BABAR, M. E., CORREIA DE ALMEIDA REGITANO, L., YILDIZ, M. A., HEATON, M. P., LIU, W. S., LEI, C. Z., REECY, J. M., SAIF-UR-REHMAN, M., SCHNABEL, R. D. & TAYLOR, J. F. 2014. Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. *PLoS Genet.,* 10**,** e1004254.

DI GIULIO, G., LYNEN, G., MORZARIA, S., OURA, C. & BISHOP, R. 2009. Live immunization against East Coast fever--current status. *Trends Parasitol.,* 25**,** 85-92.

DOHM, J. C., LOTTAZ, C., BORODINA, T. & HIMMELBAUER, H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res,* 36**,** e105.

EBERSBERGER, I., METZLER, D., SCHWARZ, C. & PAABO, S. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.,* 70**,** 1490-7.

FLICEK, P., AHMED, I., AMODE, M. R., BARRELL, D., BEAL, K., BRENT, S., CARVALHO-SILVA, D., CLAPHAM, P., COATES, G., FAIRLEY, S., FITZGERALD, S., GIL, L., GARCIA-GIRON, C., GORDON, L., HOURLIER, T., HUNT, S., JUETTEMANN, T., KAHARI, A. K., KEENAN, S., KOMOROWSKA, M., KULESHA, E., LONGDEN, I., MAUREL, T., MCLAREN, W. M., MUFFATO, M., NAG, R., OVERDUIN, B., PIGNATELLI, M., PRITCHARD, B., PRITCHARD, E., RIAT, H. S., RITCHIE, G. R., RUFFIER, M., SCHUSTER, M., SHEPPARD, D., SOBRAL, D., TAYLOR, K., THORMANN, A., TREVANION, S., WHITE, S., WILDER, S. P., AKEN, B. L., BIRNEY, E., CUNNINGHAM, F., DUNHAM, I., HARROW, J., HERRERO, J., HUBBARD, T. J., JOHNSON, N., KINSELLA, R., PARKER, A., SPUDICH, G., YATES, A., ZADISSA, A. & SEARLE, S. M. 2013. Ensembl 2013. *Nucleic Acids Res.,* 41**,** D48-55.

FLORI, L., GONZATTI, M. I., THEVENON, S., CHANTAL, I., PINTO, J., BERTHIER, D., ASO, P. M. & GAUTIER, M. 2012. A quasi-exclusive European ancestry in the Senepol tropical cattle breed highlights the importance of the slick locus in tropical adaptation. *PLoS One,* 7**,** e36133.

FOHN, L. E. & BEHRINGER, R. R. 2001. ESX1L, a novel X chromosome-linked human homeobox gene expressed in the placenta and testis. *Genomics,* 74**,** 105-8.

GAUTIER, M., FLORI, L., RIEBLER, A., JAFFREZIC, F., LALOE, D., GUT, I., MOAZAMI-GOUDARZI, K. & FOULLEY, J. L. 2009. A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. *BMC Genomics,* 10**,** 550.

GAUTIER, M. & NAVES, M. 2011. Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Mol. Ecol.,* 20**,** 3128-43.

GAUTIER, M. & VITALIS, R. 2012. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics,* 28**,** 1176-7.

GRANA, X. & REDDY, E. P. 1995. Cell cycle control in mammalian cells: role of cyclins, cyclin dependent kinases (CDKs), growth suppressor genes and cyclin-dependent kinase inhibitors (CKIs). *Oncogene,* 11**,** 211-9.

GRAVES, J. A. 2006. Sex chromosome specialization and degeneration in mammals. *Cell,* 124**,** 901-14.

GRAVES, J. A. 2010. Review: Sex chromosome evolution and the expression of sex-specific genes in the placenta. *Placenta,* 31 Suppl**,** S27-32.

HANOTTE, O., BRADLEY, D. G., OCHIENG, J. W., VERJEE, Y., HILL, E. W. & REGE, J. E. 2002. African pastoralism: genetic imprints of origins and migrations. *Science,* 296**,** 336-9.

HARRIS, E. E. & HEY, J. 1999. X chromosome evidence for ancient human histories. *PNAS,* 96**,** 3320-4.

HELMY, K. Y., KATSCHKE, K. J., JR., GORGANI, N. N., KLJAVIN, N. M., ELLIOTT, J. M., DIEHL, L., SCALES, S. J., GHILARDI, N. & VAN LOOKEREN CAMPAGNE, M. 2006. CRIg: a macrophage complement receptor required for phagocytosis of circulating pathogens. *Cell,* 124**,** 915-27.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.,* 37**,** 1-13.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.,* 4**,** 44-57.

JAKOBSSON, M. & ROSENBERG, N. A. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics,* 23**,** 1801-6.

KONG, A., GUDBJARTSSON, D. F., SAINZ, J., JONSDOTTIR, G. M., GUDJONSSON, S. A., RICHARDSSON, B., SIGURDARDOTTIR, S., BARNARD, J., HALLBECK, B., MASSON, G., SHLIEN, A., PALSSON, S. T., FRIGGE, M. L., THORGEIRSSON, T. E., GULCHER, J. R. & STEFANSSON, K. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.,* 31**,** 241-7.

LAMOND, A. I. 1991. Nuclear RNA processing. *Curr. Opin. Cell Biol.,* 3**,** 493-501.

LI, Y. & BEHRINGER, R. R. 1998. Esx1 is an X-chromosome-imprinted regulator of placental development and fetal growth. *Nat. Genet.,* 20**,** 309-11.

LI, Y., LEMAIRE, P. & BEHRINGER, R. R. 1997. Esx1,a Novel X Chromosome-Linked Homeobox Gene Expressed in Mouse Extraembryonic Tissues and Male Germ Cells. *Dev. Biol.,* 188**,** 85-95.

LIAO, X., PENG, F., FORNI, S., MCLAREN, D., PLASTOW, G. & STOTHARD, P. 2013. Whole genome sequencing of Gir cattle for identifying polymorphisms and loci under selection. *Genome,* 56**,** 592-8.

MA, Y., ZHANG, H., ZHANG, Q. & DING, X. 2014. Identification of selection footprints on the X chromosome in pig. *PLoS One,* 9**,** e94911.

MBOLE-KARIUKI, M. N., SONSTEGARD, T., ORTH, A., THUMBI, S. M., BRONSVOORT, B. M., KIARA, H., TOYE, P., CONRADIE, I., JENNINGS, A., COETZER, K., WOOLHOUSE, M. E., HANOTTE, O. & TAPIO, M. 2014. Genome-wide analysis reveals the ancient and recent admixture history of East African Shorthorn Zebu from Western Kenya. *Heredity (Edinb),* 113**,** 297-305.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.,* 20**,** 1297-303.

MUSTELIN, T., VANG, T. & BOTTINI, N. 2005. Protein tyrosine phosphatases and the immune response. *Nat. Rev. Immunol.,* 5**,** 43-57.

NGUYEN, D. K. & DISTECHE, C. M. 2006. Dosage compensation of the active X chromosome in mammals. *Nature Genetics,* 38**,** 47-53.

OH, B., HWANG, S. Y., SOLTER, D. & KNOWLES, B. B. 1997. Spindlin, a major maternal transcript expressed in the mouse during the transition from oocyte to embryo. *Development,* 124**,** 493-503.

OZAWA, H., ASHIZAWA, S., NAITO, M., YANAGIHARA, M., OHNISHI, N., MAEDA, T., MATSUDA, Y., JO, Y., HIGASHI, H., KAKITA, A. & HATAKEYAMA, M. 2004. Paired-like homeodomain protein ESXR1 possesses a cleavable C-terminal region that inhibits cyclin degradation. *Oncogene,* 23**,** 6590-602.

PORTO-NETO, L. R., SONSTEGARD, T. S., LIU, G. E., BICKHART, D. M., DA SILVA, M. V., MACHADO, M. A., UTSUNOMIYA, Y. T., GARCIA, J. F., GONDRO, C. & VAN TASSELL, C. P. 2013. Genomic divergence of zebu and taurine cattle identified through high-density SNP genotyping. *BMC Genomics,* 14**,** 876.

PRITCHARD, J. K., STEPHENS, M. & DONNELLY, P. 2000. Inference of population structure using multilocus genotype data. *Genetics,* 155**,** 945-59.

QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-2.

R Development Core Team 2012. R: A language and environment for statistical computing. Vienna, Austria.

RINCON, G., WEBER, K. L., EENENNAAM, A. L., GOLDEN, B. L. & MEDRANO, J. F. 2011. Hot topic: performance of bovine high-density genotyping platforms in Holsteins and Jerseys. *J. Dairy Sci.,* 94**,** 6116-21.

ROSENBERG, N. A. 2004. Distruct: a program for the graphical display of population structure. *Molecular Ecology Notes,* 4**,** 137-138.

RUBIN, C. J., MEGENS, H. J., MARTINEZ BARRIO, A., MAQBOOL, K., SAYYAB, S., SCHWOCHOW, D., WANG, C., CARLBORG, O., JERN, P., JORGENSEN, C. B., ARCHIBALD, A. L., FREDHOLM, M., GROENEN, M. A. & ANDERSSON, L. 2012. Strong signatures of selection in the domestic pig genome. *PNAS,* 109**,** 19529-36.

RUBIN, C. J., ZODY, M. C., ERIKSSON, J., MEADOWS, J. R., SHERWOOD, E., WEBSTER, M. T., JIANG, L., INGMAN, M., SHARPE, T., KA, S., HALLBOOK, F., BESNIER, F., CARLBORG, O., BED'HOM, B., TIXIER-BOICHARD, M., JENSEN, P., SIEGEL, P., LINDBLAD-TOH, K. & ANDERSSON, L. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature,* 464**,** 587-91.

SANKARARAMAN, S., SRIDHAR, S., KIMMEL, G. & HALPERIN, E. 2008. Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.,* 82**,** 290-303.

SCHAFFNER, S. F. 2004. The X chromosome in population genetics. *Nature Reviews Genetics,* 5**,** 43-51.

TANG, K., THORNTON, K. R. & STONEKING, M. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.,* 5**,** e171.

THUMBI, S. M., BRONSVOORT, B. M., POOLE, E. J., KIARA, H., TOYE, P. G., MBOLE-KARIUKI, M. N., CONRADIE, I., JENNINGS, A., HANDEL, I. G., COETZER, J. A., STEYL, J. C., HANOTTE, O. & WOOLHOUSE, M. E. 2014. Parasite co-infections and their impact on survival of indigenous cattle. *PLoS One,* 9**,** e76324.

TIJJANI, A. 2013. *Genome-wide characterization of diversity and admixture in African cattle breeds using high density SNP markers.* MSc, University of Nottingham.

UTSUNOMIYA, Y. T., PEREZ O'BRIEN, A. M., SONSTEGARD, T. S., VAN TASSELL, C. P., DO CARMO, A. S., MESZAROS, G., SOLKNER, J. & GARCIA, J. F. 2013. Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. *PLoS One,* 8**,** e64280.

VAN LAERE, A. S., COPPIETERS, W. & GEORGES, M. 2008. Characterization of the bovine pseudoautosomal boundary: Documenting the evolutionary history of mammalian sex chromosomes. *Genome Res.,* 18**,** 1884-95.

VOGT, L., SCHMITZ, N., KURRER, M. O., BAUER, M., HINTON, H. I., BEHNKE, S., GATTO, D., SEBBEL, P., BEERLI, R. R., SONDEREGGER, I., KOPF, M., SAUDAN, P. & BACHMANN, M. F. 2006. *VSIG4,* a B7 family-related protein, is a negative regulator of T cell activation. *J. Clin. Invest.,* 116**,** 2817-26.

VOIGHT, B. F., KUDARAVALLI, S., WEN, X. & PRITCHARD, J. K. 2006. A map of recent positive selection in the human genome. *PLoS Biol.,* 4**,** e72.

WANG, X. & ZHANG, J. 2007. Rapid evolution of primate *ESX1*, an X-linked placenta- and testis-expressed homeobox gene. *Hum. Mol. Genet.,* 16**,** 2053-60.

YANG, S., LI, X., LI, K., FAN, B. & TANG, Z. 2014. A genome-wide scan for signatures of selection in Chinese indigenous and commercial pig breeds. *BMC Genet.,* 15**,** 7.

Chapter five

Mitochondrial DNA (mtDNA) of East African shorthorn zebu: Haplotype diversity and signatures of selection[1]

[1] The EASZ mitochondrial DNA sequences were generated by Dr. Joram Mwacharo, University of Nottingham, and have been included to main text.

Abstract

Mitochondrial DNA (mtDNA) is a marker commonly used in population diversity studies. The complete sequences of domestic cattle mtDNA divide them into two major haplogroups: the T (taurine) and I (indicine). Given the importance of mitochondria in energy production, mtDNA protein-coding genes may be subjected to purifying and positive natural selection. Here, we report the comparison of the full mtDNA sequences of 13 indigenous EASZ cattle with 63 worldwide domestic cattle (Europe, n = 28; Africa, n = 19; and Asia, n = 16), two yaks and a single European auroch. This work addresses three issues: i) the extent of EASZ mtDNA genetic diversity; ii) the presence of signatures of selection in taurine mtDNA compared to zebu mtDNA; and iii) within African cattle. Neighbour-joining and median-joining network analyses indicate that EASZ mtDNA belongs to the taurine T1a, T1b and T1b1 sub-haplogroups. Nineteen taurine-zebu non-synonymous variants were detected, but none seem to be associated with a selective advantage for taurine mtDNA. Averaged and site-specific ω ratio analyses indicate that purifying selection is the main selection pressure on taurine mtDNA, with less selective constraint on the *ATP6* and *ATP8* genes. Interestingly, within African cattle, we identified a signal of positive selection in the *Cox-2* gene in the T1b/T1b1 sub-haplogroups, together the most common sub-haplogroups in the continent. We conclude that male-mediated introgression of zebu cattle into African taurine cattle remains the most likely explanation for the absence of zebu mtDNA on the continent.

Introduction

This chapter completes the picture regarding signatures of selection in the EASZ genome investigated in previous chapters. Mammalian mitochondrial DNA (mtDNA), which is an ~ 16 kb, maternally transmitted, small circular molecule, contains 37 genes (RNA and protein-coding), in addition to the hypervariable control region (D-loop). The encoded proteins are part of the electron transport chain (Complex I – Complex V) embedded in the inner-membrane of each mitochondrion, which conducts several oxidation-reduction (redox) reactions in an oxidative phosphorylation pathway to meet two primary physiological functions: i) producing ATP molecules as an energy source; and ii) generating heat to maintain body temperature (Figure 5.1) (Blier *et al*., 2001, Castellana *et al*., 2011). Most of the mitochondrial-encoded genes are RNA genes, including two ribosomal RNA (rRNA) and 22 transfer RNA (tRNA) genes. Seven of the remaining 13 protein-coding genes (*ND1–ND6* and *ND4L*) encode subunits of the NADH-ubiquinone oxidoreductase complex, the "NADH dehyrdrogenase complex" (Complex I), which is responsible for the oxidation of NADH molecules and the reduction of ubiquinone to ubiquinol. This reaction initiates an electrochemical proton gradient by pumping protons (H^+) from the mitochondrial matrix into the inter-membrane space. The mitochondrial-encoded cytochrome b (Cytb) is a subunit of the ubiquinol-cytochrome c oxidoreductase complex, the "cytochrome *bc* complex" (Complex III). In this part of the pathway, cytochrome c molecules are reduced upon oxidation of ubiquinol molecules. MtDNA also encodes three proteins (Cox1, Cox2 and Cox3) that are subunits of the fourth complex, "cytochrome c oxidase", in which cytochrome c molecules are oxidized and O_2 molecules are reduced to H_2O. Finally, the generated H^+ gradient is utilized to generate the required energy to phosphorylate ADP molecules to ATP. This step is catalysed by a transmembrane complex called ATP synthase, which is composed of two units, F_0 and F_1. The transmembrane F_0 unit pumps the protons across the membrane to the F_1 catalytic unit to generate ATP molecules. Two subunits of F_0, ATP6 and ATP8, are encoded by the mtDNA (da Fonseca *et al*., 2008). In addition to the mtDNA protein-encoding genes, several nuclear-encoded proteins participate in this pathway. For example, complex II (succinate dehydrogenase), which reduces ubiquinone molecules upon succinate oxidation, is solely composed of nuclear-encoded proteins. The ATP synthase complex is also composed of different nuclear-encoded subunits that interact with ATP6 and ATP8 (Scarpulla, 2008).

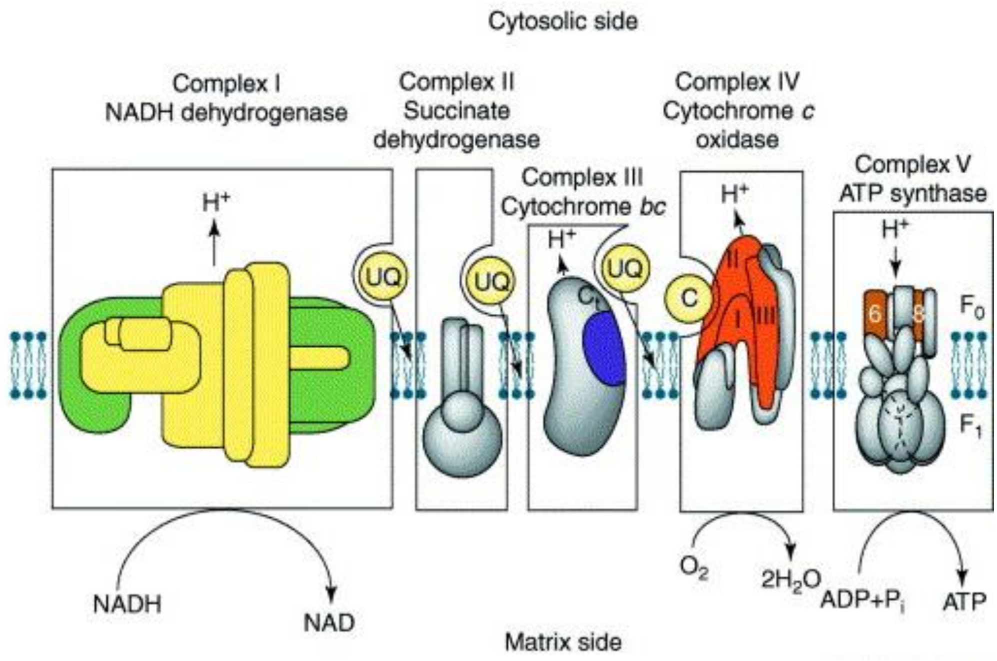

Figure 5.1: A diagram representing the electron transport chain embedded in the inner membrane of mitochondria (Blier *et al*., 2001).

Bovine mtDNA diversity

The uniparental inheritance, non-recombining nature and high substitution rate of mtDNA have made this marker a promising tool for evolutionary studies in different mammalian species, including livestock, to infer the domestication process and subsequent dispersion of domesticates (Bruford *et al*., 2003).

Studies on bovine mtDNA have clarified the differences between the two main types of cattle (taurine and zebu), revealing that they originated from different domestication events (Loftus *et al*., 1994, Bradley *et al*., 1996). This mtDNA dichotomy was further classified upon D-loop sequencing into zebu I haplogroups (I1 and I2) (Baig *et al*., 2005, Lai *et al*., 2006) and taurine T macro-haplogroups composed of the T1, T2, T3 and T4 haplogroups (Troy *et al*., 2001, Mannen *et al*., 2004). The T4 haplogroup is actually nested within the T3 haplogroup (Achilli *et al*., 2008). The T2 and T3 haplogroups are mostly in European taurine cattle, while the T1 haplogroup has been found mainly in African cattle (Troy *et al*., 2001). In Egyptian and Northwest African cattle, T2 and T3 haplogroups were also found in small proportions (Lenstra *et* *al*., 2014). Subsequently, analyses of the complete mtDNA sequences from different European, African and Asian cattle populations has further divided the T macro-haplogroup into a rare T5 haplogroup found in Italian and Iraqi cattle and the T1'2'3 haplogroup, which were shown to have a common ancestry and likely originated in the Near East (Achilli *et al*., 2008). With an increasing number of available sequences, the T1 haplogroup has been sub-divided into six sub-haplogroups (T1a, T1b, T1c, T1d, T1e and T1f), with the T1b, T1c and T1d sub-haplogroups further subdivided into minor haplogroups (T1b1, T1c1 and T1d1) (Bonfiglio *et al*., 2012). Two rare non-T haplogroups have been discovered in some modern Italian taurine cattle breeds (haplogroups Q and R), perhaps the result of local auroch introgression in the case of the R haplogroup (Achilli *et al*., 2009, Bonfiglio *et al*., 2010), while haplogroup P has been found in the European wild auroch *Bos primigenius* (Edwards *et al*., 2010) as well as in modern Korean taurine beef cattle (Achilli *et al*., 2008).

Signatures of selection in mtDNA

Early studies have indicated the unique presence of taurine mtDNA on the African continent, even in African zebu cattle (Loftus *et al*., 1994, Bradley *et al*., 1996). Today, despite a large number of African mtDNA studies (e.g., Bonfiglio *et al*. (2012), Horsburgh *et al*. (2013), Salim *et al*. (2014)), zebu mtDNA has still not been detected on the African continent. These observations are attributed to a predominantly male-mediated Asian zebu introgression into local African taurine cattle (Bradley *et al*., 1996, Hanotte *et al*., 2002), leaving the alternative hypothesis of selection in favour of taurine mtDNA on the African continent thus far unexplored.

Given the critical role of mitochondria as energy providers to eukaryotic cells, selective pressures on mtDNA may be expected. Analyses on mtDNA protein-encoding genes indicate footprints of purifying and adaptive (positive) selection in different mammalian (Mishmar *et al*., 2003, Elson *et al*., 2004, Stewart *et al*., 2008, Shen *et al*., 2010, Melo-Ferreira *et al*., 2014) and non-mammalian species (Castoe *et al*., 2008, Garvin *et al*., 2011, Parmakelis *et al*., 2013). Moreover, the predominant role of purifying selection in shaping the distribution of mammalian mtDNA variations has been supported experimentally using a genetically engineered mouse line (a mtDNA mutator mouse line) (Stewart *et al*., 2008).

An interesting example is the possible link between mtDNA protein-coding polymorphisms and temperature. Mishmar *et al*. (2003) and Ruiz-Pesini *et al*. (2004) revealed that human mtDNA is mainly the target of purifying selection, while the study of Ruiz-Pesini *et al*. (2004) indicated that tropical and sub-tropical African mtDNA haplogroups might have been under greater purifying selection than temperate and arctic Eurasian haplogroups. They also identified signatures of adaptive selection in different mitochondrial genes, e.g., *ND2*, *ND4* and *ATP6*, in human populations living within the arctic zone, as the maintenance of such mtDNA variants may be beneficial in cold environments (Ruiz-Pesini *et al*., 2004). Indeed, similar signatures of positive selection at eight mtDNA genes (including *ATP8*, *CYTB* and *ND4*) were identified in mtDNA lineages of different hare species living in the arctic (Melo-Ferreira *et al*., 2014). Moreover, although not linked to temperature, codons in different mtDNA genes, such as *ND5* in the Pacific salmon (genus *Oncorhynchus*)(Garvin *et al*., 2011) and *Cox 1* and *Cytb* in different land snail species (Parmakelis *et al*., 2013), were also found to be under positive selection.

This chapter focuses on three issues using full mtDNA sequence information from EASZ and worldwide cattle breeds. Firstly, we assess the mtDNA diversity in EASZ in relation to the known cattle mtDNA haplogroups. Secondly, we investigate whether there is any evidence of selection pressure in taurine mtDNA compared to zebu mtDNA. Finally, we investigate the possible presence of signatures of positive selection in African taurine mtDNA protein-coding genes.

Materials and methods

EASZ sample collection and DNA extraction

Blood samples were collected from 13 genetically unrelated EASZ calves from 13 sub-locations in the Busia district in western Kenya (Table S5.1). The calves were sampled as part of a study examining infectious diseases of African livestock (the IDEAL project at the International Livestock Research Institute (ILRI)). Blood was collected from each calf via jugular veni-puncture. Genomic DNA was extracted using a Nucleon DNA extraction kit (GE Healthcare Life Sciences, Buckinghamshire UK) at ILRI laboratories in Nairobi and shipped to the University of Nottingham. Upon arrival, DNA samples were stored at -80^{o}C.

PCR amplification and Sanger sequencing

The full mtDNA of cattle (16,346 bp) was amplified *via* PCR using 11 pairs of primers (see Table S3 in Achilli *et al.*, 2008). PCR amplifications were conducted using a DNA engine Dyad Peltier Thermal Cycler in 20 µl reaction volumes containing 40 ng of genomic DNA, 1x GoTaq Flexi buffer (Promega, Madison, WI, USA), 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 0.5 µM of each primer and 1.25 units of GoTaq Flexi DNA Polymerase (Promega). Thermo-cycling conditions included an initial denaturation step at 95^{o}C for 3 minutes, then 35 cycles of the following: 1 minute at 95^{o}C, 30 seconds at 58^{o}C and 90 seconds at 72^{o}C. This was followed by a final extension phase at 72^{o}C for 5 minutes. Successfully amplified PCR products were purified using a QIAGEN QIAquick PCR Purification Kit (QIAGEN, West Sussex, UK). All purified PCR products were sequenced directly using 32 primers (see Table S4 in Achilli *et al.*, 2008) using an Applied Biosystems BigDye® Terminator v3.1 Cycle Sequencing Chemistry (Applied Biosystems) on an ABI prism 3730xl DNA analyser (Applied Biosystems) following the manufacturer’s recommendations. The 32 sequenced fragments were edited manually using BioEdit v7.0 (Hall, 1999). Fragments were then joined to reconstitute the full mtDNA genome using GAP software (Bonfield *et al.*, 1995). Overlaps containing mismatches were corrected in BioEdit.

Downloaded mtDNA sequences

Full mtDNA sequences of 63 domestic cattle samples were downloaded from the NCBI database. These samples were from six Asian populations (16 samples), 10 European populations (28 samples) and eight African populations (19 samples) (Table S5.1). The full mtDNA of two yak samples *Bos grunniens* (GenBank IDs: GQ464267.1 and GQ464246.1) (Qiu *et al.*, 2012) and a single European auroch *Bos primigenius* (GenBank ID: NC_013996.1) (Edwards *et al.,* 2010) were also included. The UMD3.1 taurine reference mtDNA sequence was downloaded from the Ensembl genome browser (Flicek *et al.*, 2013).

Mitochondrial DNA sequence alignment and phylogeny construction

MEGA 5.05 software (Tamura *et al.*, 2011) was used to align all mtDNA sequences. The alignment was based on the ClustalW (Larkin *et al.*, 2007) function implemented in MEGA 5.05 using the default settings. A neighbour-joining (NJ) tree was constructed for the full mtDNA based on a maximum composite likelihood substitution model (default model). The substitution rates of the codon positions were gamma distributed (gamma parameter = 0.07) to account for variation in the substitution rate between the different codon positions. To estimate the reliability of the internal branches of the tree, 1,000 bootstrap resamplings were performed.

Median-joining network

A median-joining network was constructed using Network 4.6 software (Bandelt *et al.*, 1999) to identify the T1 mtDNA sub-haplogroups (Bonfiglio *et al*. (2012) present in EASZ and N'Dama cattle. The full mtDNA sequences of these samples, in addition to the African cattle samples from different T1 sub-haplogroups (Bonfiglio *et al.*, 2012) were aligned with the UMD3.1 mtDNA sequence using MEGA 5.05. DnaSP 5.1 software (Rozas *et al*., 2003) was used to generate the haplotypes file required for the network calculation.

Taurine-zebu differentiation mtDNA variants

Non-synonymous variants and variants in RNA genes that differentiate between the taurine and zebu haplogroups were identified from the mtDNA sequence alignment. Possible biological effects of the non-synonymous variants were predicted using PolyPhen-2 online tool (http://genetics.bwh.harvard.edu/pph2/) (Adzhubei *et al*., 2010). The inter-species conservation of these non-synonymous variants was explored in 11 different species: *Homo sapiens* (human), *Mus musculus* (mouse), *Canis lupus* (dog), *Oryctolagus cuniculus* (rabbit), *Pan troglodytes* (chimpanzee), *Gorilla gorilla gorilla* (gorilla), *Equus caballus* (horse), *Sus scrofa* (pig), *Ovis aries* (sheep), *Bos primigenius* (auroch) and *Bos grunniens* (yak).

Signatures of selection analysis

The ratio of the number of non-synonymous substitutions per non-synonymous sites (Dn) to the number of synonymous substitutions per synonymous sites (Ds) (the ω ratio) was used to identify signatures of selection in the mtDNA protein-encoding genes. The wild yak *Bos grunniens* mtDNA was used as the reference sequence. The numbers of non-synonymous and synonymous sites and substitutions were estimated using DnaSP 5.1 software (Librado and Rozas, 2009). Both the Dn and Ds ratios were corrected for multiple hits using the Jukes and Cantor model as suggested by Nei and Gojobori (1986). ω ratio analyses were conducted on three datasets: i) taurine and zebu mtDNA together (54 samples), ii) taurine mtDNA (52 samples), and iii) African taurine mtDNA (22 samples). For details of the samples used in these analyses see Table S5.1.

As the calculated ω ratios averaged selective pressure over all coding sites, this method is considered to be highly conservative and stringent for the identification of signatures of positive selection (Crandall *et al*., 1999, Anisimova *et al*., 2001). Therefore, site models, implemented in the CODEML package of Phylogenetic Analysis by Maximum Likelihood (PAML) version 4.6 software, were used to determine ω ratio variation between the codons of the mtDNA protein-coding genes using maximum likelihood estimation (Yang, 1997, Nielsen and Yang, 1998, Yang *et al*., 2000, Yang, 2007). Different models explaining the distribution of ω ratios between codons were implemented in this analysis. Some of them assume only neutrality, while others allow positive selection. The one ratio (M0) model allows a single ω for all codons. The nearly neutral model (M1a) assumes two classes of codons: one with $0 \leq \omega_0 < 1$ and a proportion of codons p_0, while the second class assumes $\omega_1 = 1$ and a proportion of codons $p_1 = 1 - p_0$. Model M2a (positive selection) is an extension of the M1a model that contains a third class that allows $\omega_2 > 1$ and a proportion of codons $p_2 = 1 - p_1 - p_0$. Model M3 (discrete) uses discrete classes to model the heterogeneity of ω between codons. By default, three classes were used in this model. Both the M7 and M8 models assume a beta distribution of ω over codons with two beta function parameters (*p* and *q*). The M7 model does not allow codons under positive selection by constraining ω to be in the interval (0, 1). In contrast, the model M8 allows codons with $\omega_1 > 1$ with a proportion of codons p_1.

After calculating the log likelihood value (L) of each model to fit our data, likelihood ratio tests (LRT) were conducted between the M1a–M2a and M7–M8 models to test for positive selection. A third LRT was conducted between the M0 and M3 models to test for variations in ω ratios between sites. The statistic for each LRT is defined as twice the log likelihood difference between the two models (2ΔL). This statistic was compared to a χ^2 distribution with a degree of freedom (d.f.) equal to the difference in the number of parameters between the two models. In the LTRs between M1a–M2a and M7–M8, the d.f. is 2, while in the third LRT, between M0–M3, it is 4 (Yang *et al*., 2000, Wong *et al*., 2004).

A Bayes Empirical Bayes (BEB) approach was used to identify codons under positive selection if the LRT was significant. This was done by assigning a posterior probability for each codon to be under positive selection and accounting for sampling error to reduce the rate of false positives (Yang *et al.*, 2005). The analyses were conducted in the three cattle datasets described above.

Branch-site models of positive selection, implemented in the CODEML package of PAML version 4.6 software (Zhang *et al.*, 2005), were also tested on the mtDNA genes to detect whether a codon was subjected to positive selection in a specific lineage of the tree (foreground lineage = African taurine), but remained neutral or under purifying selection in the other lineages (background lineages).

Model A assumes positive selection by defining four classes of codons. Class 0 includes codons with $0 \leq \omega_0 < 1$ throughout the tree. Class 1 includes codons with $\omega_1 = 1$ throughout the tree. Class 2a includes codons with $0 \leq \omega_0 < 1$ in the background lineages, and $\omega_2 \geq 1$ in the foreground lineage. Class 2b includes codons with $\omega_1 = 1$ in the background lineages, and $\omega_2 \geq 1$ in the foreground lineages. The null model, which does not assume positive selection, is the same as Model A, but fixes $\omega_2 = 1$. As in the site models, the log likelihood value (L) was calculated for each model to fit our data, and LRTs were conducted between Model A and the null model to test for positive selection. The calculated log likelihood statistic for the LRT was compared to a χ^2 distribution (d.f. = 1) (Yang *et al.*, 2000, Wong *et al.*, 2004). A BEB approach was used to identify codons under positive selection if the LRT was significant.

Nucleotide diversity and inter-population divergence of cattle populations

DnaSP 5.1 software was used to calculate nucleotide diversity (the average number of nucleotide substitutions per site) for the taurine and zebu cattle populations used in the selection analyses for each protein-coding gene. Inter-population nucleotide divergence (the average number of nucleotide substitutions per site between two populations) was also calculated between taurine and zebu cattle, as well as between the taurine populations.

Mitochondrial-related genes in autosomal candidate sweep regions

Nuclear genes related to mitochondria were explored in the autosomal candidate sweep regions identified in Chapters 3 and 4. Variants (SNPs and indels) were investigated within the identified genes (see the Materials and Methods section in Chapter 3).

Results

Phylogenetic analysis

A neighbour-joining tree of the full mtDNA sequences groups the 13 EASZ with the other African cattle, including a single N'Dama sequence, and some South European taurine cattle. The tree shows that the EASZ mtDNA sequences belong to the main taurine cattle branch (T macro-haplogroup) and, more specifically, to the T1 haplogroup (Figure 5.2 and Figure S5.1). All of the EASZ samples carry the diagnostic polymorphisms 16050 (C/T) and 16113 (T/C) that define the T1 haplogroup (Anderson *et al*., 1982). The tree also confirms the separation between the taurine and zebu haplogroups, and within the taurine haplogroups (T, P, Q and R). The separation of the T macro-haplogroup into its haplogroups (T1, T2, T3, T4 and T5) is also visible, but the relationships between the taurine haplogroup were not always supported by high bootstrap values (Figure 5.2).

Median-joining network

The median-joining network divides the EASZ samples between the T1a (n = 6), T1b (n = 1) and T1b1 (n = 6) sub-haplogroups. These samples also carried the corresponding diagnostic polymorphisms 2055+C (T1a), 7542 (G/A) (T1b) and 16022 (G/A) (T1b1). Additionally, the previously unassigned N'Dama sample was affiliated to the T1a sub-haplogroup (Figure 5.3 and Figure S5.2).

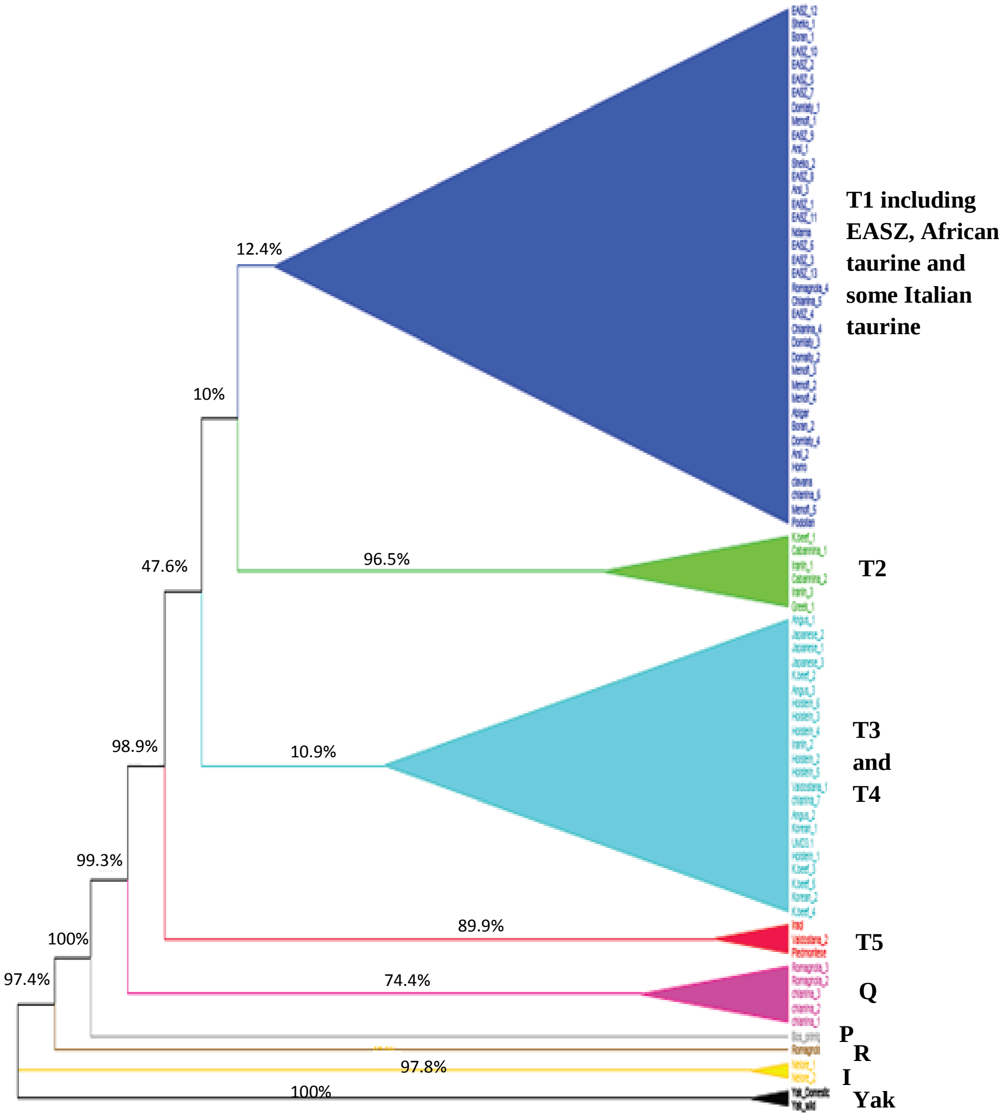

Figure 5.2: Unrooted neighbour-joining (NJ) tree of all mtDNA sequences included in this chapter, with 1,000 bootstrap replications. Dark blue: T1 sub-haplogroup (EASZ, African taurine and some Italian taurine cattle). Green: T2 sub-haplogroup. Light blue: T3 and T4 sub-haplogroups. Red: T5 sub-haplogroup. Pink: Q haplogroup. Grey: P haplogroup. Brown: R haplogroup. Yellow: I haplogroup (Nellore). Black: yak mtDNA. Equal branch lengths are for the purpose of illustration only. Please see also Figure S5.1 for the rooted NJ tree.

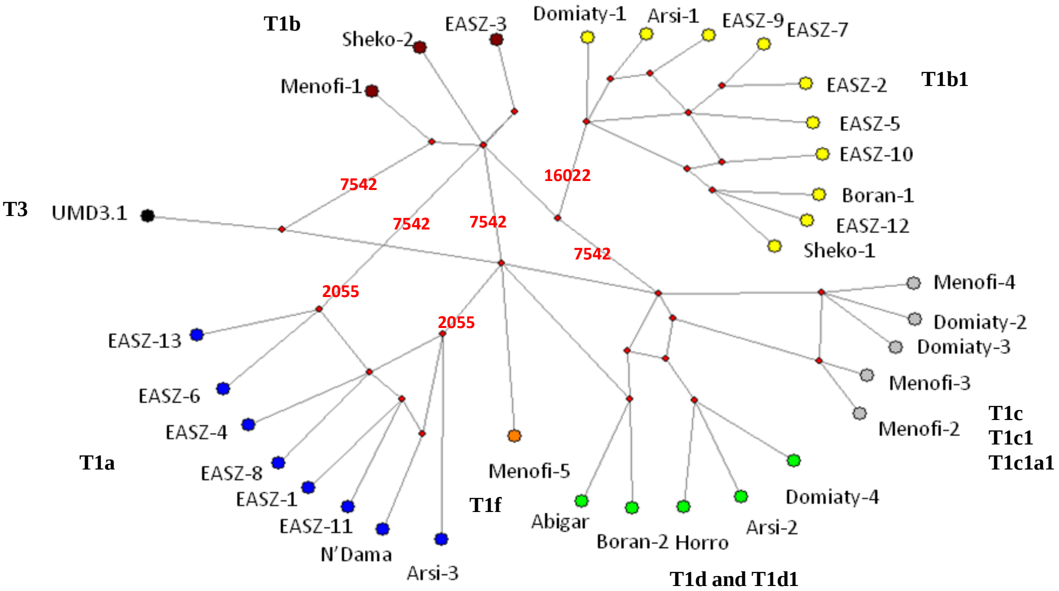

Figure 5.3: Median-joining network analysis including 13 EASZ, 1 N'Dama and 18 African cattle with known mtDNA sub-haplogroup types (refer to Bonfiglio *et al.*, 2012 and Table S5.1). Each node represents one haplotype. Black (T3). Blue (T1a). Green (T1d and T1d1). Grey (T1c, T1c1 and T1c1a1). Yellow (T1b1). Brown (T1b). Orange (T1f). Branch lengths are unscaled. Branches are labelled with the diagnostic polymorphisms of the T1a (2055), T1b (7542) and T1b1 (16022) sub-haplogroups. Positions correspond to the UMD3.1 reference sequence, excluding indels. Please see also Figure S5.2 for full polymorphic sites.

Taurine-zebu diagnostic variants

No unique polymorphisms in African taurine mtDNA were identified. However, 19 non-synonymous polymorphisms and 29 mtDNA RNA polymorphisms, 21 in rRNA and 8 in tRNA genes, which differentiate between taurine and zebu mtDNA (Table 5.1 and 5.2) were found. All of the non-synonymous variants indicated a benign effect on the biological function of the corresponding genes following the analyses conducted on PolyPhen-2 (Adzhubei *et al.*, 2010). Moreover, none of the positions of these polymorphisms were conserved among the mammalian species investigated or were unique to taurine mtDNA (Table 5.3).

Table 5.1: Position (UMD3.1 reference sequence excluding indels) of the taurine-zebu non-synonymous variants

Full mtDNA position	**Gene**	**Nucleotide position (bp)**	**Taurine nucleotide**	**Zebu nucleotide**	**Amino acid residue**	**Taurine amino acid**	**Zebu amino acid**	**Variant sites' annotation in the protein**
3600	*ND1*	500	C	T	167 *	T	I	unknown
8210	*ATP8*	82	G	A	28 ¤	V	I	unknown
8285	*ATP8*	157	A	G	53	T	A	unknown
8308	*ATP6*	19	A	G	7	T	A	unknown
8494	*ATP6*	205	A	G	69	T	A*	unknown
8749	*ATP6*	460	A	G	154	M	V	unknown
9480	*Cox3*	511	G	A	171	V	I*§	subunit III/nuclear encoded subunit *via* interface helix
10066	*ND3*	244	G	A	82	A	T	unknown
10590	*ND4*	62	A	G	21	N	S	NADH-ubiquinone oxidoreductase chain 4, N-terminus
10691	*ND4*	163	C	G	55	L	V*	NADH-ubiquinone oxidoreductase chain 4, N-terminus
12178	*ND5*	70	T	C	24	F	L	not in an active domain
12377	*ND5*	269	T	C	90	I	T	NADH-ubiquinone oxidoreductase, chain 5, N-terminus
12622	*ND5*	514	A	G	172	I	V*	NADH-ubiquinone, various chains
12923	*ND5*	815	A	T	272	Y	F ¥	NADH-ubiquinone , various chains
13433	*ND5*	1325	A	G	442	N	S	NADH dehydrogenase subunit 5, C terminus
13628	*ND5*	1520	T	C	507	M	T	NADH dehydrogenase subunit 5, C terminus
13908	*ND5*	1800	C	A	600	I	M	NADH dehydrogenase subunit 5, C terminus
15579	*CYTB*	1066	G	A	356	V	I	not in an active domain
15629	*CYTB*	1116	C	T	372	I	V	not in an active domain

These zebu diagnostic amino acids were found in one Romagnola (*), one Korean taurine (¤), one Iranian taurine (§) and one European Angus (¥).

Table 5.2: Taurine-zebu diagnostic polymorphisms in mtDNA RNA genes. Polymorphism: (taurine, nucleotide position (UMD3.1 excluding indels), zebu)

Polymorphism	Gene	Polymorphism	Gene
A 518 G	rRNA	T 2637 C	rRNA
C 722 T	rRNA	C 2953 T	rRNA
C 740 T	rRNA	T 2979 C	rRNA
T 761 C	rRNA	C 2988 T	rRNA
A 816 G	rRNA	T 2989 C	rRNA
C 1158 T	rRNA	G 2990 A	rRNA
C 1482 T	rRNA	A 3051 G	tRNA
C 1492 T	rRNA	T 3071 C	tRNA
T 1677 C	rRNA	C 7304 T	tRNA
A 1824 G	rRNA	G 7358 A	tRNA
T 1860 A	rRNA	C 7361 T	tRNA
C 1869 T	rRNA	T 9767 C	tRNA
T 2016 C	rRNA	C 15741 T	tRNA
T 2099 C	rRNA	A 15751 G	tRNA
A 2575 G	rRNA		

Table 5.3: Amino acid status of the taurine-zebu non-synonymous variants in different mammalian species.

Gene	**Amino acid residue**	**Nellore**	**Taurine**	**Human**	**Mouse**	**Rabbit**	**Chimpanzee**	**Gorilla**	**Dog**	**Horse**	**Pig**	**Sheep**	***B. primigenius***	**Wild Yak**
ND1	**167**	I	T	T	T	T	T	M	I	I	T	I	T	I
ATP8	**28**	I	V	M	V	F	M	V	I	I	I	I	V	I
ATP8	**53**	A	T	P	L	T	P	S	N	S	M	T	T	T
ATP6	**7**	A	T	A	A	S	A	A	A	A	A	A	T	A
ATP6	**69**	A	T	S	T	S	S	S	A	T	S	A	T	T
ATP6	**154**	V	M	M	M	M	M	M	M	V	V	V	M	M
Cox3	**171**	I	V	L	L	I	L	L	V	V	V	V	V	V
ND3	**82**	T	A	T	I	N	A	T	N	T	N	A	A	T
ND4	**21**	S	N	H	K	H	R	H	N	N	N	S	N	N
ND4	**55**	V	L	T	M	T	P	S	M	M	T	T	V	M
ND5	**24**	L	F	V	S	T	I	I	T	F	S	F	F	L
ND5	**90**	T	I	I	T	T	I	I	V	V	I	V	T	V
ND5	**172**	V	I	I	I	I	I	I	V	I	I	I	I	V
ND5	**272**	F	Y	L	F	T	L	L	T	T	L	F	Y	H
ND5	**442**	S	N	N	S	V	N	N	L	S	L	N	N	N
ND5	**507**	T	M	L	L	L	L	L	T	M	T	T	M	M
ND5	**600**	M	I	L	I	F	L	L	L	L	T	T	I	I
CYTB	**356**	I	V	V	I	V	V	V	I	I	I	I	V	I
CYTB	**372**	V	I	I	I	I	I	I	I	I	I	I	V	I

Signatures of selection analyses

The average ω ratios for all 13 mtDNA protein-coding genes were less than 1 in the three analysed data sets: i) taurine and zebu mtDNA together (54 samples), ii) taurine mtDNA (52 samples), and iii) African taurine mtDNA (22 samples) (Figure 5.4). These results supported purifying selection as the main type of selection at these genes. Interestingly, the *ATP6* and *ATP8* genes showed lower conservation, i.e., higher ω ratios, than the other genes, whilst the *Cox-1*, Cox-2 and *ND3* genes showed a high degree of conservation (Figure 5.4).

None of the positive selection models (M2a and M8) in PAML fitted any of the analysed data sets significantly better than the neutral models (M1a and M7) for any of the mtDNA protein-coding genes. Different amino acid substitutions were proposed to be under positive selection by the BEB approach, but with a posterior probability < 95% (Tables S5.2). Exceptionally for *Cox-2*, model M8 fitted the data significantly better than M7 (*P*-value = 0.04), and the BEB approach proposed that the non-synonymous variant (D57N) is under positive selection, with a posterior probability > 95%, when the African cattle samples were analysed separately. The outputs of the branch-site tests for positive selection did not indicate any evidence of positive selection acting on the foreground lineage specified (African taurine cattle) (Table S5.3).

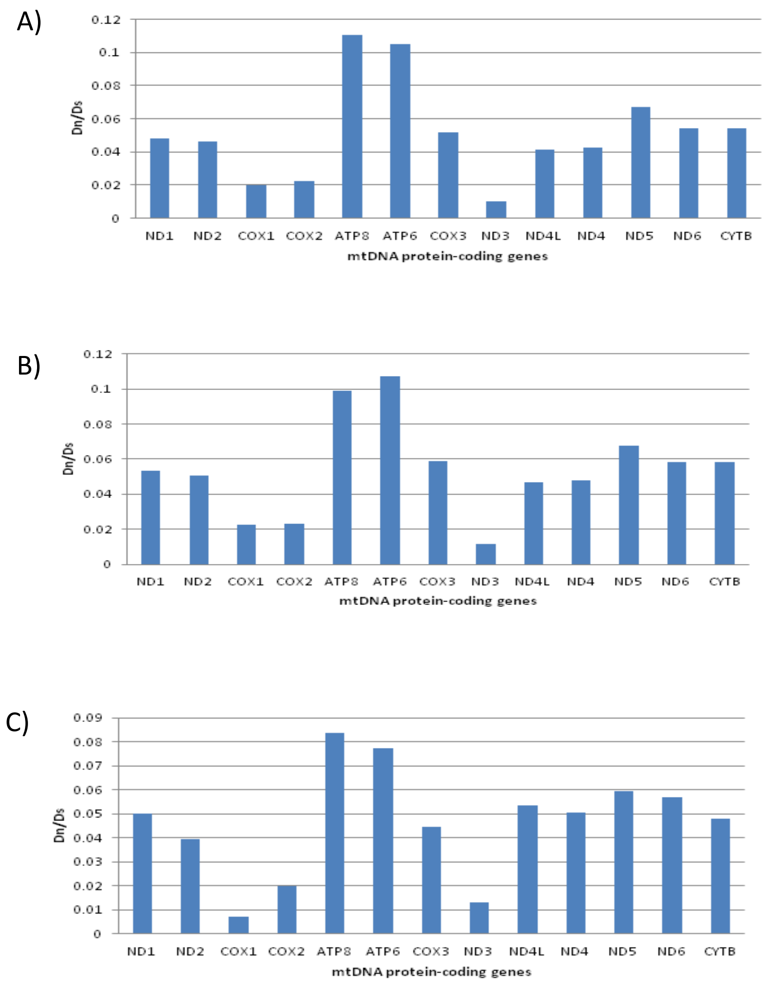

Figure 5.4: Dn/Ds (ω ratio) calculation corrected by the Jukes and Cantor model for the 13 mtDNA protein-coding genes using the wild yak mtDNA sequence as a reference sequence. (A) All cattle. (B) All taurine cattle. (C) African taurine cattle.

Nucleotide diversity and inter-population divergence

The average nucleotide diversities for the taurine and zebu cattle populations over all the protein-coding genes were 0.001 and 0, respectively. Additionally, the mean divergences between the different taurine populations analysed (European, African and Asian) and between the taurine and zebu populations over all mtDNA protein-coding genes were 0.001 and 0.016, respectively (Table S5.4).

Mitochondrial-related genes in autosomal and BTA X candidate sweep regions

A total of 15 mitochondrial-related genes were identified within the candidate sweep regions in Chapters 3 and 4 (Table 5.4). Two genes, *ATP5B* and *GLS2*, were selected as likely candidates to be under positive selection due to their obvious roles in ATP generation (see the Discussion section). Within the 15 genes, 1,109 SNPs (276 EASZ-specific) and 40 indels (10 EASZ-specific) were identified. Ten of these SNPs, located on seven genes, were non-synonymous (three were EASZ-specific in three genes). Within the *ATP5B* and *GLS2* genes, 102 SNPs and seven indels were found (Table S5.5). Interestingly, a single non-synonymous variant was identified in amino acid 53 (G 53 A) in *GLS2*. Based on the 10 EASZ exome sequences (see the Materials and Methods section in Chapter 3), the frequency of the non-reference amino acid of this variant was 35%. PolyPhen-2 predicted a benign effect for this variant.

Table 5.4: A list of mitochondrial-related genes identified within the EASZ autosomal and BTA X candidate sweep regions in Chapters 3 and 4.

BTA	Start (bp)	End (bp)	Gene ID	Gene Name	Gene Description
5	57,119,917	57,125,290	ENSBTAG00000013315	ATP5B	ATP synthase, H+ transporting, mitochondrial F1 complex, beta polypeptide
5	57,241,073	57,253,590	ENSBTAG00000009284	GLS2	glutaminase 2 (liver, mitochondrial)
6	100,006,745	100,011,712	ENSBTAG00000018155	MRPS18C	mitochondrial ribosomal protein S18C
7	33,311,979	33,313,090	ENSBTAG00000005779	FTMT	ferritin mitochondrial (FTMT)
7	62,740,774	62,749,742	ENSBTAG00000019950	GRPEL2	GrpE-like 2, mitochondrial (E. coli)
11	72,315,287	72,324,567	ENSBTAG00000018269	MPV17	MpV17 mitochondrial inner membrane protein
12	35,840,400	35,841,483	ENSBTAG00000003315	MRP63	mitochondrial ribosomal protein 63
13	38,546,389	38,554,838	ENSBTAG00000003425	MGME1	mitochondrial genome maintenance exonuclease 1
13	51,797,279	51,809,475	ENSBTAG00000013545	MAVS	mitochondrial antiviral signaling protein
16	24,976,175	24,998,201	ENSBTAG00000047287	MARC1	mitochondrial amidoxime reducing component 1
16	56,660,856	56,688,617	ENSBTAG00000004358	DARS2	aspartyl-tRNA synthetase 2, mitochondrial
18	36,598,368	36,600,210	ENSBTAG00000001663	PDF	peptide deformylase (mitochondrial)
18	36,674,604	36,696,376	ENSBTAG00000002412	CYB5B	cytochrome b5 type B (outer mitochondrial membrane)
19	27,089,299	27,092,026	ENSBTAG00000004910	SLC25A11	solute carrier family 25 (mitochondrial carrier; oxoglutarate carrier), member 11
X	38,362,070	38,383,556	ENSBTAG00000011648	TMLHE	Trimethyllysine dioxygenase, mitochondrial

Discussion

In this chapter, we explored EASZ mtDNA diversity and affiliated it to the known cattle mtDNA haplogroups. Then, we investigated whether selective pressures might have shaped the diversity of African cattle mtDNA, and more specifically, whether the presence of taurine mtDNA in all types of African cattle (taurine and zebu) presently on the continent is the result of selective pressures, rather than the consequence of a male-mediated zebu introgression. For this purpose, we analysed, for the first time, the full mtDNA of EASZ cattle from Kenya, an indigenous population of cattle from the Horn of Africa, which is the main entry point of Asian zebu cattle (Hanotte *et al*., 2002).

East African Shorthorn Zebu mtDNA diversity

As for other African cattle (Loftus *et al*., 1994, Bradley *et al*., 1996, Bonfiglio *et al*., 2012, Horsburgh *et al*., 2013), only taurine mtDNA was identified in EASZ. More specifically, we identified 13 new taurine T1 haplotypes. Therefore, it seems that the T2 and T3 haplogroups observed in Egyptian and Northwest African cattle (Lenstra *et al*., 2014) might not be present in Kenya, but this will need to be confirmed by analysing a large number of samples and populations.

In agreement with previous studies (Achilli *et al*., 2009, Bonfiglio *et al*., 2010, Bonfiglio *et al*., 2012), we also showed here that sequencing information from the full mtDNA allows the identification of different T sub-haplogroups using diagnostic sites outside the D-loop.

A selective advantage for taurine mtDNA in African zebu cattle?

The presence of only taurine type mtDNA in African cattle is primarily attributed to a male-mediated introgression of Asian zebu cattle into African taurine cattle (Bradley *et al*., 1996, Hanotte *et al*., 2002). Alternatively, taurine mtDNA may have been preferably selected on the continent. To address this issue, we first analysed non-synonymous variants differentiating between taurine and zebu mtDNA using PolyPhen-2. This online tool indicated that they have benign effects. However, PolyPhen-2 does not consider the possible effects from the interaction between the different mutations in a specific gene; thus, it remains possible that these variants may have been selected collectively. Moreover, none of the non-synonymous variants identified were unique to taurine mtDNA when the mtDNAs of other mammalian species were investigated. It might also be possible that the target of selection is not the coding sites, but one or several of the 29 taurine-zebu polymorphisms identified in mtDNA RNA genes (tRNA and rRNA).

Thus far, only one study has investigated the possible selective advantage of taurine/zebu mtDNA in zebu cattle (Guzerat dairy cattle) by comparing production and reproduction traits between the mtDNA types (Paneto *et al*., 2008). Their results did not indicate any apparent selective advantage of any mtDNA type in association with the studied phenotypes.

Signatures of selection in taurine mtDNA

While there is no evidence for positive selection favouring taurine mtDNA over zebu mtDNA, it is still possible that taurine mtDNA is under selection due to the exposure to the African environment. Following the averaged ω-ratio and PAML analyses, we found evidence of purifying selection acting on the mtDNA of the analysed data sets (Figure 5.4 and Tables S5.2 and S5.3). This is consistent with the selection pattern observed in other mammalian, e.g., humans (Mishmar *et al*., 2003, Ruiz-Pesini *et al*., 2004, Elson *et al*., 2004), mice (Stewart *et al*., 2008), whales (Soares *et al*., 2013) and pigs (Soares *et al*., 2013), and non-mammalian mtDNA, e.g., Pacific salmon (Garvin *et al*., 2011) and tits (Zink, 2005).

It is important to note that although the ω ratio categories supporting positive selection have been indicated by PAML in some of the analysed genes (Table S5.2), and the BEB approach demonstrated several of their variants to be under positive selection, these results are not conclusive given that none of the positive selection models significantly fitted the analysed data sets better than the non-positive selection models implemented in PAML.

An exception is the identification of a candidate positively selected non-synonymous variant in *Cox-2* (D 57 N) in African cattle only. It may indicate that this site is under recurrent selection, i.e., it is not fixed yet. This mutated site is the diagnostic marker of the T1b sub-haplogroup (7542 G/A). This sub-haplogroup was the most common one reported by Bonfiglio *et al*. (2012) in African cattle (63.8%), as well among our EASZ cattle (54%). Therefore, T1b and its derived sub-haplogroup (T1b1) may be considered to be an advantageous mtDNA type in African cattle.

We also observed higher averaged ω ratios for *ATP6* and *ATP8* than for the other mtDNA protein-coding genes. Given that these two encoded subunits are parts of a multi-subunit complex (F_0), it is possible that their interactions with the nuclear-encoded subunits of F_0 mask any deleterious mutation from purifying selection. In other words, the nuclear subunits may acquire compensatory mutations that restore the function of the mutated ATP6 and ATP8 subunits (Castellana *et al*., 2011). This is called cyto-nuclear coevolution (Rand *et al*., 2004).

Alternatively, it indicates a possible weak signal of positive selection on these two genes. ATP6 and ATP8 subunits are mainly involved in stabilizing the oligomerization of the ATP synthase complex to efficiently generate ATP (Jonckheere *et al*., 2012). Mutations in these two genes can affect the assembly and the functionality of the overall complex, and, hence, they will be expected to be targeted by positive selection, as has been observed in humans (Ruiz-Pesini *et al*., 2004), hare species from the arctic regions (Melo-Ferreira *et al*., 2014), snakes

(Castoe *et al*., 2008), the parasitic wasp *Nasonia* (Oliveira *et al*., 2008) and bats (Shen *et al*., 2010). Elson *et al*. (2004) have also reported a signal of positive selection, or a relaxation of purifying selection, in *ATP6*, especially in European and Asian human mtDNA sequences.

Signatures of selection in mitochondrial-related nuclear genes

Selection in nuclear-encoded mitochondrial proteins may also be of importance. For example, natural selection may result in the fixation of beneficial variants to optimize the interactions between nuclear-encoded and mitochondrial-encoded proteins, or to maintain variants that improve mitochondrial functionality, e.g., ATP production.

Interestingly, 15 nuclear-encoded genes associated with mitochondria were found within the sweep regions obtained by the high-density SNP genotypes and the EASZ full genome sequence data in Chapters 3 and 4. Although most of these genes have an unclear role in maintaining proper mitochondrial functionality, two of them may be considered to be candidates for selection. Glutaminase 2 (*GLS2*) has a role in regulating energy metabolism by increasing the production of glutamate *via* the hydrolysis of glutamine. This will lead to an increase in ATP production in mitochondria (Hu *et al*., 2010). *ATP5B* has a role in maintaining efficient mitochondrial ATP production. This gene encodes the β subunit of the ATP synthase F1 complex, which is the catalytic unit responsible for ATP production.

Although the non-synonymous variant in *GLS2* exhibited a low non-reference amino acid frequency and a benign effect, genotyping more cattle samples for this variant is highly desirable. Moreover, variants in the other identified genes should also be considered as candidates in future work.

Limitations in detecting positive selection on mtDNA

Several limitations associated with our data may have affected the power of the signatures of selection analyses conducted here, especially the PAML analysis. Firstly, taurine and zebu cattle are subspecies of domestic cattle, showing a low level of divergence in mtDNA protein-coding genes (Table S5.4). This low inter-population divergence and intra-population diversity observed in our data are major factors that can reduce the power of the site-specific ω ratio test (Anisimova *et al*., 2001). This method is, indeed, more suitable to detect recurrent selection, not directional selection that derives an advantageous mutation to fixation (Wong *et al*., 2004). Another problem, present more specifically in *ATP8*, is the size of the sequence analysed (only 66 amino acids). Anisimova *et al*. (2001) have indicated, based on simulated data, that short sequences are associated with lower LRT power to detect adaptive evolution. Moreover, the non-recombining nature of mtDNA is also an obstacle in defining positively selected sites. This is due to the high LD level between variants and, hence, the resulting sweep will mask a favourable site by reducing diversity.

Conclusion

In this chapter, we have shown that our EASZ samples belong to the taurine T1 haplogroup, more specifically to the T1a, T1b and T1b1 sub-haplogroups. Unfortunately, we could not define any selective advantage specific to this type of mtDNA in African zebu cattle, leaving the “male-mediated introgression of Asian zebu” as the more likely explanation. We have also shown that purifying selection is predominantly influencing mtDNA sequence diversity in cattle, with *ATP6* and *ATP8* showing lower conservation than the other mtDNA protein-coding genes. However, a mutated site within the *Cox-2* gene leads us to propose that the T1b, and its derived sub-haplogroup T1b1, might represent an advantageous mtDNA type in African cattle. We also cannot exclude that mtDNA RNA and mitochondrial-related nuclear genes can be considered to be targets of positive selection. Increasing the number of EASZ samples and including more African cattle populations from across the continent will be required to better characterise African cattle mtDNA diversity and the selection pressures by which it has been shaped.

References

ACHILLI, A., OLIVIERI, A., PELLECCHIA, M., UBOLDI, C., COLLI, L., AL-ZAHERY, N., ACCETTURO, M., PALA, M., HOOSHIAR KASHANI, B., PEREGO, U. A., BATTAGLIA, V., FORNARINO, S., KALAMATI, J., HOUSHMAND, M., NEGRINI, R., SEMINO, O., RICHARDS, M., MACAULAY, V., FERRETTI, L., BANDELT, H. J., AJMONE-MARSAN, P. & TORRONI, A. 2008. Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Curr. Biol.,* 18**,** R157-8.

ACHILLI, A., BONFIGLIO, S., OLIVIERI, A., MALUSA, A., PALA, M., HOOSHIAR KASHANI, B., PEREGO, U. A., AJMONE-MARSAN, P., LIOTTA, L., SEMINO, O., BANDELT, H. J., FERRETTI, L. & TORRONI, A. 2009. The multifaceted origin of taurine cattle reflected by the mitochondrial genome. *PLoS One,* 4**,** e5753.

ADZHUBEI, I. A., SCHMIDT, S., PESHKIN, L., RAMENSKY, V. E., GERASIMOVA, A., BORK, P., KONDRASHOV, A. S. & SUNYAEV, S. R. 2010. A method and server for predicting damaging missense mutations. *Nat. Methods,* 7**,** 248-9.

ANDERSON, S., DE BRUIJN, M. H. L., COULSON, A. R., EPERON, I. C., SANGER, F. & YOUNG, I. G. 1982. Complete sequence of bovine mitochondrial DNA conserved features of the mammalian mitochondrial genome. *J. Mol. Biol.,* 156**,** 683-717.

ANISIMOVA, M., BIELAWSKI, J. P. & YANG, Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.,* 18**,** 1585-92.

BAIG, M., BEJA-PEREIRA, A., KULKARNI, K., FARAH, S. & LUIKART, G. 2005. Phylogeography and origin of Indian domestic cattle. *current science* 89**,** 38-40.

BANDELT, H. J., FORSTER, P. & ROHL, A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.,* 16**,** 37-48.

BLIER, P. U., DUFRESNE, F. & BURTON, R. S. 2001. Natural selection and the evolution of mtDNA-encoded peptides: evidence for intergenomic co-adaptation. *Trends Genet.,* 17**,** 400-6.

BONFIELD, J. K., SMITH, K. & STADEN, R. 1995. A new DNA sequence assembly program. *Nucleic Acids Res.,* 23**,** 4992-9.

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.

BONFIGLIO, S., GINJA, C., DE GAETANO, A., ACHILLI, A., OLIVIERI, A., COLLI, L., TESFAYE, K., AGHA, S. H., GAMA, L. T., CATTONARO, F., PENEDO, M. C., AJMONE-MARSAN, P., TORRONI, A. & FERRETTI, L. 2012. Origin and spread of *Bos taurus*: new clues from mitochondrial genomes belonging to haplogroup T1. *PLoS One,* 7**,** e38601.

BRADLEY, D. G., MACHUGH, D. E., CUNNINGHAM, P. & LOFTUS, R. T. 1996. Mitochondrial diversity and the origins of African and European cattle. *PNAS,* 93**,** 5131-5.

BRUFORD, M. W., BRADLEY, D. G. & LUIKART, G. 2003. DNA markers reveal the complexity of livestock domestication. *Nat. Rev. Genet.,* 4**,** 900-10.

CASTELLANA, S., VICARIO, S. & SACCONE, C. 2011. Evolutionary patterns of the mitochondrial genome in Metazoa: exploring the role of mutation and selection in mitochondrial protein coding genes. *Genome. Biol. Evol.*.

CASTOE, T. A., JIANG, Z. J., GU, W., WANG, Z. O. & POLLOCK, D. D. 2008. Adaptive evolution and functional redesign of core metabolic proteins in snakes. *PLoS One,* 3**,** e2201.

CRANDALL, K. A., KELSEY, C. R., IMAMICHI, H., LANE, H. C. & SALZMAN, N. P. 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol Biol Evol,* 16**,** 372-82.

DA FONSECA, R. R., JOHNSON, W. E., O'BRIEN, S. J., RAMOS, M. J. & ANTUNES, A. 2008. The adaptive evolution of the mammalian mitochondrial genome. *BMC Genomics,* 9**,** 119.

EDWARDS, C. J., MAGEE, D. A., PARK, S. D., MCGETTIGAN, P. A., LOHAN, A. J., MURPHY, A., FINLAY, E. K., SHAPIRO, B., CHAMBERLAIN, A. T., RICHARDS, M. B., BRADLEY, D. G., LOFTUS, B. J. & MACHUGH, D. E. 2010. A complete mitochondrial genome sequence from a mesolithic wild aurochs (*Bos primigenius*). *PLoS One,* 5**,** e9255.

ELSON, J. L., TURNBULL, D. M. & HOWELL, N. 2004. Comparative genomics and the evolution of human mitochondrial DNA: assessing the effects of selection. *Am. J. Hum. Genet.,* 74**,** 229-38.

FLICEK, P., AHMED, I., AMODE, M. R., BARRELL, D., BEAL, K., BRENT, S., CARVALHO-SILVA, D., CLAPHAM, P., COATES, G., FAIRLEY, S., FITZGERALD, S., GIL, L., GARCIA-GIRON, C., GORDON, L., HOURLIER, T., HUNT, S., JUETTEMANN, T., KAHARI, A. K., KEENAN, S., KOMOROWSKA, M., KULESHA, E., LONGDEN, I., MAUREL, T., MCLAREN, W. M., MUFFATO, M., NAG, R., OVERDUIN, B., PIGNATELLI, M., PRITCHARD, B., PRITCHARD, E., RIAT, H. S., RITCHIE, G. R., RUFFIER, M., SCHUSTER, M., SHEPPARD, D., SOBRAL, D., TAYLOR, K., THORMANN, A., TREVANION, S., WHITE, S., WILDER, S. P., AKEN, B. L., BIRNEY, E., CUNNINGHAM, F., DUNHAM, I., HARROW, J., HERRERO, J., HUBBARD, T. J., JOHNSON, N., KINSELLA, R., PARKER, A., SPUDICH, G., YATES, A., ZADISSA, A. & SEARLE, S. M. 2013. Ensembl 2013. *Nucleic Acids Res.,* 41**,** D48-55.

GARVIN, M. R., BIELAWSKI, J. P. & GHARRETT, A. J. 2011. Positive Darwinian selection in the piston that powers proton pumps in complex I of the mitochondria of Pacific salmon. *PLoS One,* 6**,** e24127.

HALL, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series***,** 95-98.

HANOTTE, O., BRADLEY, D. G., OCHIENG, J. W., VERJEE, Y., HILL, E. W. & REGE, J. E. 2002. African pastoralism: genetic imprints of origins and migrations. *Science,* 296**,** 336-9.

HORSBURGH, K. A., PROST, S., GOSLING, A., STANTON, J. A., RAND, C. & MATISOO-SMITH, E. A. 2013. The genetic diversity of the Nguni breed of African Cattle (*Bos* spp.): complete mitochondrial genomes of haplogroup T1. *PLoS One,* 8**,** e71956.

HU, W., ZHANG, C., WU, R., SUN, Y., LEVINE, A. & FENG, Z. 2010. Glutaminase 2, a novel p53 target gene regulating energy metabolism and antioxidant function. *PNAS,* 107**,** 7455-60.

JONCKHEERE, A. I., SMEITINK, J. A. & RODENBURG, R. J. 2012. Mitochondrial ATP synthase: architecture, function and pathology. *J. Inherit. Metab. Dis.,* 35**,** 211-25.

LAI, S. J., LIU, Y. P., LIU, Y. X., LI, X. W. & YAO, Y. G. 2006. Genetic diversity and origin of Chinese cattle revealed by mtDNA D-loop sequence variation. *Mol. Phylogenet. Evol.,* 38**,** 146-54.

LARKIN, M. A., BLACKSHIELDS, G., BROWN, N. P., CHENNA, R., MCGETTIGAN, P. A., MCWILLIAM, H., VALENTIN, F., WALLACE, I. M., WILM, A., LOPEZ, R., THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics,* 23**,** 2947-8.

LENSTRA, J., AJMONE-MARSAN, P., BEJA-PEREIRA, A., BOLLONGINO, R., BRADLEY, D., COLLI, L., DE GAETANO, A., EDWARDS, C., FELIUS, M., FERRETTI, L., GINJA, C., HRISTOV, P., KANTANEN, J., LIRÓN, J., MAGEE, D., NEGRINI, R. & RADOSLAVOV, G. 2014. Meta-Analysis of Mitochondrial DNA Reveals Several Population Bottlenecks during Worldwide Migrations of Cattle. *Diversity,* 6**,** 178-187.

LIBRADO, P. & ROZAS, J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25**,** 1451-2.

LOFTUS, R. T., MACHUGH, D. E., BRADLEY, D. G., SHARP, P. M. & CUNNINGHAM, P. 1994. Evidence for 2 independent domestications of cattle. *PNAS,* 91**,** 2757-2761.

MANNEN, H., KOHNO, M., NAGATA, Y., TSUJI, S., BRADLEY, D. G., YEO, J. S., NYAMSAMBA, D., ZAGDSUREN, Y., YOKOHAMA, M., NOMURA, K. & AMANO, T. 2004. Independent mitochondrial origin and historical genetic differentiation in North Eastern Asian cattle. *Mol. Phylogenet. Evol.*, 32**,** 539-44.

MELO-FERREIRA, J., VILELA, J., FONSECA, M. M., DA FONSECA, R. R., BOURSOT, P. & ALVES, P. C. 2014. The elusive nature of adaptive mitochondrial DNA evolution of an arctic lineage prone to frequent introgression. *Genome Biol. Evol.*, 6**,** 886-96.

MISHMAR, D., RUIZ-PESINI, E., GOLIK, P., MACAULAY, V., CLARK, A. G., HOSSEINI, S., BRANDON, M., EASLEY, K., CHEN, E., BROWN, M. D., SUKERNIK, R. I., OLCKERS, A. & WALLACE, D. C. 2003. Natural selection shaped regional mtDNA variation in humans. *PNAS*, 100**,** 171-6.

NEI, M. & GOJOBORI, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, 3**,** 418-26.

NIELSEN, R. & YANG, Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics,* 148**,** 929-36.

OLIVEIRA, D. C., RAYCHOUDHURY, R., LAVROV, D. V. & WERREN, J. H. 2008. Rapidly evolving mitochondrial genome and directional selection in mitochondrial genes in the parasitic wasp *Nasonia* (hymenoptera: pteromalidae). *Mol. Biol. Evol.*, 25**,** 2167-80.

PANETO, J. C., FERRAZ, J. B., BALIEIRO, J. C., BITTAR, J. F., FERREIRA, M. B., LEITE, M. B., MERIGHE, G. K. & MEIRELLES, F. V. 2008. *Bos indicus* or *Bos taurus* mitochondrial DNA - comparison of productive and reproductive breeding values in a Guzerat dairy herd. *Genet Mol Res*, 7**,** 592-602.

PARMAKELIS, A., KOTSAKIOZI, P. & RAND, D. 2013. Animal mitochondria, positive selection and cyto-nuclear coevolution: insights from pulmonates. *PLoS One,* 8**,** e61970.

QIU, Q., ZHANG, G., MA, T., QIAN, W., WANG, J., YE, Z., CAO, C., HU, Q., KIM, J., LARKIN, D. M., AUVIL, L., CAPITANU, B., MA, J., LEWIN, H. A., QIAN, X., LANG, Y., ZHOU, R., WANG, L., WANG, K., XIA, J., LIAO, S., PAN, S., LU, X., HOU, H., WANG, Y., ZANG, X., YIN, Y., MA, H., ZHANG, J., WANG, Z., ZHANG, Y., ZHANG, D., YONEZAWA, T., HASEGAWA, M., ZHONG, Y., LIU, W., HUANG, Z., ZHANG, S., LONG, R., YANG, H., LENSTRA, J. A., COOPER, D. N., WU, Y., SHI, P. & LIU, J. 2012. The yak genome and adaptation to life at high altitude. *Nat. Genet.*, 44**,** 946-9.

RAND, D. M., HANEY, R. A. & FRY, A. J. 2004. Cytonuclear coevolution: the genomics of cooperation. *Trends Ecol. Evol.*, 19**,** 645-53.

ROZAS, J., SANCHEZ-DELBARRIO, J. C., MESSEGUER, X. & ROZAS, R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics,* 19**,** 2496-7.

RUIZ-PESINI, E., MISHMAR, D., BRANDON, M., PROCACCIO, V. & WALLACE, D. C. 2004. Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science,* 303**,** 223-6.

SALIM, B., TAHA, K. M., HANOTTE, O. & MWACHARO, J. M. 2014. Historical demographic profiles and genetic variation of the East African Butana and Kenana indigenous dairy zebu cattle. *Anim. Genet.*, 45**,** 782-90.

SCARPULLA, R. C. 2008. Transcriptional paradigms in mammalian mitochondrial biogenesis and function. *Physiol. Rev.*, 88**,** 611-38.

SHEN, Y. Y., LIANG, L., ZHU, Z. H., ZHOU, W. P., IRWIN, D. M. & ZHANG, Y. P. 2010. Adaptive evolution of energy metabolism genes and the origin of flight in bats. *PNAS,* 107**,** 8666-71.

SOARES, P., ABRANTES, D., RITO, T., THOMSON, N., RADIVOJAC, P., LI, B., MACAULAY, V., SAMUELS, D. C. & PEREIRA, L. 2013. Evaluating purifying selection in the mitochondrial DNA of various mammalian species. *PLoS One*, 8**,** e58993.

STEWART, J. B., FREYER, C., ELSON, J. L., WREDENBERG, A., CANSU, Z., TRIFUNOVIC, A. & LARSSON, N. G. 2008. Strong purifying selection in transmission of mammalian mitochondrial DNA. *PLoS Biol,* 6**,** e10.

TAMURA, K., PETERSON, D., PETERSON, N., STECHER, G., NEI, M. & KUMAR, S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.,* 28**,** 2731-9.

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**,** 1088-91.

WONG, W. S., YANG, Z., GOLDMAN, N. & NIELSEN, R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics,* 168**,** 1041-51.

YANG, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.,* 13**,** 555-6.

YANG, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.,* 24**,** 1586-91.

YANG, Z., NIELSEN, R., GOLDMAN, N. & PEDERSEN, A. M. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics,* 155**,** 431-49.

YANG, Z., WONG, W. S. & NIELSEN, R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.,* 22**,** 1107-18.

ZHANG, J., NIELSEN, R. & YANG, Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.,* 22**,** 2472-9.

ZINK, R. M. 2005. Natural selection on mitochondrial DNA in *Parus* and its relevance for phylogeographic studies. *Proc. Biol. Sci.,* 272**,** 71-8.

Chapter six

General discussion and future prospects

This thesis has investigated the full genome (autosomes, chromosome X and mtDNA) of an ancient, stabilised zebu-taurine admixed African cattle population, the small East African shorthorn zebu (EASZ), with the aim to identify candidate genome regions with signatures of positive selection. These regions might be associated with African environmental challenges, human selection and/or adaptive admixture that lead to reproductive fitness. We cannot confidently separate these different selective forces, as they have likely acted together, leading to the selection of the favourable haplotypes in the EASZ genome. We also aimed to define the origins of the selected regions and to determine the candidate genes/variants under positive selection, both for nuclear and the mtDNA.

The identified regions harbour different candidate genes associated with various biological functions, e.g., innate and adaptive immunity, reproduction and fertility, and physiological development, which are crucial for ensuring the survival of these cattle in their local environment. Several EASZ candidate regions have been confirmed to show signatures of positive selection in separate zebu-taurine admixed African cattle populations from Uganda and Nigeria (data from this thesis), and in other African and non-African tropical-adapted admixed cattle analysed in previous studies (Gautier and Naves, 2011, Flori *et al*., 2014). Although EASZ are not considered to be highly productive breeds, the possibility of past selection for production traits in this indigenous cattle population cannot be excluded, as illustrated by the overlap between production QTL and some candidate regions under positive selection.

Chapter 2 is our starting point to explore the genome of EASZ. However, covering the full genome of EASZ with low-density genotyped SNPs (< 50,000 SNPs after quality control), using the Illumina Bovine SNP50 BeadChip v.1, does not represent a full characterisation of the EASZ genome. Increasing the SNP density to > 300,000 SNPs has been suggested for cattle; in particular, for between breed analyses (Goddard and Hayes, 2009). The target is to have SNPs in high linkage disequilibrium (LD) with the candidate regions under selection in populations with different breeding histories. In addition to the genome coverage issue, SNP arrays will inevitably display an ascertainment bias, as these polymorphic markers are selected in a restricted number of breeds at the array design phase (Matukumalli *et al*., 2009).

In Chapters 3 and 4, we tried to overcome the above limitations using the high-density Illumina BovineHD Genotyping BeadChip and full EASZ genome sequence information. As shown in Table 6.1, not surprisingly, most of the candidate regions identified using the low-density SNP chip (Chapter 2) were also revealed by the high-density SNP chip and/or the full genome sequence analyses. About 42% (10 out of 24 regions) of the candidate regions identified in Chapter 2 were not detected by the high-density SNP chip and/or the full EASZ genome sequence. This might be attributed to the different types of analysis conducted in these chapters, as each test has its own strengths in detecting a specific selection pattern at a specific time-scale (reviewed in Oleksyk *et al*. (2010)). Alternatively, because of the aforementioned limitations associated with the Illumina Bovine SNP50 BeadChip v.1, these 10 regions might have been identified as possible false-positives. In contrast, analyses of the high-density genotyped SNPs and full EASZ genome sequence led to the identification of an additional 272 candidate regions (253 in autosomes and 19 in the BTA X sex chromosome). This illustrates the superiority of the utilised genomic tools in comparison with the low-density SNP array used in Chapter 2. However, we should also emphasise that a *Bos taurus taurus* reference genome was used to align the EASZ sequence reads. Given that EASZ are zebu-taurine admixed cattle with predominantly zebu background, new candidate regions might be identified by using the, yet unavailable, *Bos taurus indicus* genome as a reference.

Table 6.1: Candidate regions identified in Chapter 2 and their overlapping regions in Chapters 3 and 4.

Chapter 2			**Chapter 3 and 4**	
BTA	**Chromosomal Position**	**Identifying Statistic**	**SNP array (*meta-SS*)**	**Sequence (*Hp*)**
2	125,585,810 – 126,058,677	F_{ST}	125,159,084 – 125,994,861	125,640,001 – 126,083,262
3	101,942,771	*Rsb*	_	_
4	47,195,467 – 47,539,595	F_{ST}	_	_
4	51,927,595 – 52,308,430	F_{ST}	_	_
5	57,977,594	*Rsb*	56,651,062 – 57,515,653	_
5	60,556,520	*Rsb*	_	60,610,001 – 60,721,361
5	76,286,670	*iHS*	_	_
7	52,224,595 – 52,720,797	F_{ST}	52,183,599 – 53,001,042	51,360,001 – 53,362,761
11	62,629,106	*Rsb*	61,877,437 – 62,548,419	62,810,001 – 62,971,604
12	27,181,474	*Rsb*	27,050,192 – 29,151,436	_
12	29,217,254	*Rsb*	27,050,192 – 29,151,436	29,110,001 – 29,438,417
12	35,740,174	*Rsb*	35,445,176 – 36,965,854	_
13	46,433,697 – 46,723,493	F_{ST}	_	_
13	57,848,276 – 58,207,174	F_{ST}	58,273,562 – 58,599,491	57,990,001 – 58,122,843
14	24,482,969 – 25,254,540	F_{ST}	_	_
19	27,369,763 – 27,763,447	F_{ST}	26,909,816 – 27,143,239	27,490,001 – 27,674,039
19	42,696,815	*Rsb*	43,023,638 – 43,692,285	42,890,001 – 43,122,753
22	2,314,019 – 2,788,566	F_{ST}	_	2,890,001 – 30,5754,9
23	28,281,915	*iHS*	_	_
24	4,118,163 – 4,474,760	F_{ST}	_	4,660,001 – 4,771,543
29	1,898,171	*iHS*	_	_
X	8,582,093 – 9,248,137	F_{ST}	_	_
X	39,942,044 – 42,024,368	F_{ST}	_	_
X	84,566,018 – 85,993,719	F_{ST}	86,750,029 – 87,174,478*	_

* identified by *iHS*

We also explored the possible presence of signatures of positive selection in the same candidate regions in other cattle populations. Our primary objective was to cross-validate the EASZ candidate regions in other tropically adapted zebu-taurine admixed cattle from Uganda and Nigeria. Assuming these populations had different breeding histories, such an approach should reduce the probability of identifying false-positive candidate regions following genetic drift. It may also help fine mapping the regions of interest. The approach assumes that the same selection pressures will act on cattle populations in different geographic areas. We addressed the latter by comparing crossbreed zebu-taurine populations from three geographic areas along roughly the same latitudes. Importantly, we also identified specific candidate regions for positive selection in the three populations examined. There is definitively room for further analysis, keeping in mind that this thesis primarily focused on EASZ.

Among all the identified candidate genes, several have been recently investigated by different research groups. One example is *RXFP2*. This gene, which is mapped within a candidate region on BTA 12 in EASZ, has been associated with male fertility (Gorlov *et al*., 2002, Agoulnik, 2007, Feng *et al*., 2009) and horn development in sheep (Johnston *et al*., 2011). The available pooled EASZ genome sequence revealed three exonic SNPs, two non-synonymous and one synonymous, in this gene. In addition to these SNPs, other genomic variants, such as mutations in non-coding regions (e.g., transcription factor binding sites) and structural chromosomal variants, may regulate the expression of *RXFP2*. This will require further investigation to accurately define the causative mutations under selection.

It is important to realise that while our bioinformatics tools are increasingly being refined, they are not yet perfect. For example, defining the biological effect of variants on the expression of a candidate gene is an important step to qualify them as candidate mutations under selection. In this thesis, the effects of the non-synonymous variants in *RXFP2* were predicted using the available online tool PolyPhen-2 (Adzhubei *et al*., 2010). However, this tool does not predict the effects of possible interactions between these mutations on gene function. Importantly, any candidate mutation will need to be validated experimentally, using not only model organisms (e.g., mice) and/or experimental cell lines, as in Carneiro *et al*. (2014), but also with the development of targeted mutations in cattle using different biotechnology techniques, such as TALENs (Carlson *et al*., 2012) and CRISPR/Cas9 (Wang *et al*., 2013).

Additionally, several CNV detection algorithms are now available to analyse full genome (e.g., CNVnator (Abyzov *et al*., 2011), CNV-seq (Xie and Tammi, 2009) and cnMOPS (Klambauer *et al*., 2012)) and exome data (e.g., CoNIFER (Krumm *et al*., 2012), ExomeCNV (Sathirapongsasuti *et al*., 2011), ExomeDepth (Plagnol *et al*., 2012) and cnMOPS (Klambauer *et al*., 2012)). Although we followed a customised approach to detect signals of CNV, it will be noteworthy to use these algorithms on our exome data to accurately define CNV regions, i.e., correcting for any sources of bias that might lead to false-positives (e.g., GC-content biased) (Dohm *et al*., 2008, Benjamini and Speed, 2012). The depth of the coverage analysis conducted in Chapters 3 and 4 was confined to the coding regions (exome) of the EASZ genome. This analysis should be widened to include un-pooled full EASZ genome sequence data, as this will allow us to define CNV regions in non-coding regions and to calculate the frequency of the CNVs in EASZ (Abyzov *et al*., 2011, Bickhart *et al*., 2012, Hou *et al*., 2012). A comparative analysis with a non-EASZ genome is also important to define CNVs specific to EASZ. Analysing our genotyped SNPs data for CNV regions is another area that can be further investigated (Hou *et al*., 2011, Hou *et al*., 2012).

The last chapter of this thesis (Chapter 5) has clarified the diversity and phylogeny of EASZ mtDNA, affiliating them with the domestic cattle T1 macro-haplogroup, more specifically, T1a, T1b and T1b1. We only analysed 13 EASZ samples; thus, more data will be required to accurately characterise the diversity of mtDNA in this population. We aimed to address the question of whether positive selection for taurine mtDNA explains its unique presence on the continent. Our results revealed no specific signature of positive selection in taurine *versus* zebu mtDNA. However, a signal of positive selection has been identified in a non-synonymous variant in *Cox-2* in the

T1b/T1b1 sub-haplogroups. Although the available data indicate that T1b and its derived haplotypes are the most common haplotypes on the African continent, this will need to be further investigated *via* the analysis of more animals. It is possible that the unique presence of taurine mtDNA on the African continent is the result of two concomitant factors, a predominantly male-derived zebu introgression (Bradley *et al*., 1996, Hanotte *et al*., 2002), as well as some selection acting on specific taurine variants.

Future prospects

Climate change is a major issue that has direct consequences for livestock health. Given the continuous emission of greenhouse gases, e.g., CO_2, it is estimated that the global mean surface temperature will increase by 3.7^{o}C to 4.8^{o}C in 2100 (IPCC, 2014). In Africa, the Atmospheric Oceanic General Circulation Model (AOGCM) predicts an average increase in temperature of 1.4^{o}C in 2020, 2.6^{o}C in 2060 and 4.4^{o}C in 2100, whilst it has been estimated that there will be an average increase in rainfall of 0.6% in 2020, followed by 4.2% and 12% decreases in 2060 and 2100, respectively (Seo and Mendelsohn, 2008). This will, in turn, affect the distribution of vector-borne diseases, e.g., bluetongue (Wittmann *et al*., 2001), reduce the cattle reproduction rate (Hansen, 2007) and it may decrease milk yield and quality (Bernabucci and Calamari, 1998, Nardone *et al*., 2010).

High attention is now given to conserve genetic resources that are adapted to harsh environmental challenges, e.g., heat stress. Indigenous African cattle populations are mainly selected, naturally and artificially, for survival in the African environment (drought, heat stress and parasitic diseases) (Murray and Trail, 1984, Norval *et al*., 1992, Rege *et al*., 2001). They are used as multi-purpose livestock species for ploughing, transport, milk production, as well as a source of fertilizers, while fulfilling many socio-economic roles. The genome of indigenous African cattle is full of valuable genetic material (Rege *et al*., 2001, Hanotte *et al*., 2010). These adaptive traits are in danger of being eroded due to the continuous introgression of exotic European cattle genomes to increase productivity, e.g., milk yield (Hanotte *et al*., 2010). Although high productivity is a requirement to maintain sufficient income for farmers, it is a short-term, unsustainable solution in the local production system.

The results obtained in this thesis, as well as those in other studies of African cattle populations (Gautier *et al*., 2009, Flori *et al*., 2014), can be considered to be the first step towards conserving this valuable genetic material. Here, we have defined key genome regions that need to be maintained in future composite African cattle populations to sustain survival and productivity in their local environment. To achieve this, informed selective crossbreeding between indigenous and exotic cattle is required to ensure the inheritance and fixation of these genome regions in the resulting composite population. This type of marker-assisted introgression (MAI) crossbreeding involves the introgression of a chromosomal fragment of interest into a different breed or population from which it originates (Lecomte *et al*., 2004). The MAI approach requires identifying genetic markers in high LD with the haplotype of interest. Although no studies on cattle have yet been published, MAI has been successfully used for the introgression of three trypanotolerance QTL in mice (Koudande *et al*., 2005), and two intramuscular fat QTL in pigs (Sato *et al*., 2014). This approach is associated with several drawbacks: i) it is a time-consuming process due to the long generation time (e.g., ~ 6 years in cattle (Keightley and Eyre-Walker, 2000)), ii) the possible introgression of unselected haplotypes and iii) the outcome of this introgression may be dependent on the genetic background of the recipient population (Sato *et al*., 2014, Bahbahani and Hanotte, 2015).

Alternatively, gene modification techniques, such as TALENs (Carlson *et al*., 2012) and CRISPR/Cas9 (Wang *et al*., 2013), may be used in cattle breeding to genetically engineer calves carrying favourable allele(s). These techniques have successfully introduced several SNPs and indels in large (cattle) and small (sheep and goats) ruminants, as well as in pigs (e.g., T 1591 C in the porcine *P65* gene and 11 bp deletion in the bovine and ovine *GDF8* gene) (Carlson *et al*., 2012, Tan *et al*., 2013, Hai *et al*., 2014, Proudfoot *et al*., 2015). However, causative mutations under selection in indigenous African cattle need to be identified first, a prerequisite not needed for MAI.

References

ABYZOV, A., URBAN, A. E., SNYDER, M. & GERSTEIN, M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.,* 21**,** 974-84.

ADZHUBEI, I. A., SCHMIDT, S., PESHKIN, L., RAMENSKY, V. E., GERASIMOVA, A., BORK, P., KONDRASHOV, A. S. & SUNYAEV, S. R. 2010. A method and server for predicting damaging missense mutations. *Nat. Methods,* 7**,** 248-9.

AGOULNIK, A. I. 2007. Relaxin and related peptides in male reproduction. *Adv. Exp. Med. Biol.,* 612**,** 49-64.

BAHBAHANI, H. & HANOTTE, O. 2015. Genetic resistance – tolerance to vector-borne diseases, prospect and challenges of genomics. *OIE Scientific and Technical Review,* 34**,** 185-97.

BENJAMINI, Y. & SPEED, T. P. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.,* 40**,** e72.

BERNABUCCI, U. & CALAMARI, L. 1998. Effects of heat stress on bovine milk yield and composition. *Zoot. Nutriz. Anim.,* 24**,** 247 - 257.

BICKHART, D. M., HOU, Y., SCHROEDER, S. G., ALKAN, C., CARDONE, M. F., MATUKUMALLI, L. K., SONG, J., SCHNABEL, R. D., VENTURA, M., TAYLOR, J. F., GARCIA, J. F., VAN TASSELL, C. P., SONSTEGARD, T. S., EICHLER, E. E. & LIU, G. E. 2012. Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res.,* 22**,** 778-90.

BRADLEY, D. G., MACHUGH, D. E., CUNNINGHAM, P. & LOFTUS, R. T. 1996. Mitochondrial diversity and the origins of African and European cattle. *PNAS,* 93**,** 5131-5.

CARLSON, D. F., TAN, W., LILLICO, S. G., STVERAKOVA, D., PROUDFOOT, C., CHRISTIAN, M., VOYTAS, D. F., LONG, C. R., WHITELAW, C. B. & FAHRENKRUG, S. C. 2012. Efficient TALEN-mediated gene knockout in livestock. *PNAS,* 109**,** 17382-7.

CARNEIRO, M., RUBIN, C. J., DI PALMA, F., ALBERT, F. W., ALFOLDI, J., BARRIO, A. M., PIELBERG, G., RAFATI, N., SAYYAB, S., TURNER-MAIER, J., YOUNIS, S., AFONSO, S., AKEN, B., ALVES, J. M., BARRELL, D., BOLET, G., BOUCHER, S., BURBANO, H. A., CAMPOS, R., CHANG, J. L., DURANTHON, V., FONTANESI, L., GARREAU, H., HEIMAN, D., JOHNSON, J., MAGE, R. G., PENG, Z., QUENEY, G., ROGEL-GAILLARD, C., RUFFIER, M., SEARLE, S., VILLAFUERTE, R., XIONG, A., YOUNG, S., FORSBERG-NILSSON, K., GOOD, J. M., LANDER, E. S., FERRAND, N., LINDBLAD-TOH, K. & ANDERSSON, L. 2014. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science,* 345**,** 1074-9.

DOHM, J. C., LOTTAZ, C., BORODINA, T. & HIMMELBAUER, H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.,* 36**,** e105.

FENG, S., FERLIN, A., TRUONG, A., BATHGATE, R., WADE, J. D., CORBETT, S., HAN, S., TANNOUR-LOUET, M., LAMB, D. J., FORESTA, C. & AGOULNIK, A. I. 2009. INSL3/RXFP2 signaling in testicular descent. *Ann. N. Y. Acad. Sci.,* 1160**,** 197-204.

FLORI, L., THEVENON, S., DAYO, G. K., SENOU, M., SYLLA, S., BERTHIER, D., MOAZAMI-GOUDARZI, K. & GAUTIER, M. 2014. Adaptive admixture in the West African bovine hybrid zone: insight from the Borgou population. *Mol. Ecol.,* 23**,** 3241-57.

GAUTIER, M., FLORI, L., RIEBLER, A., JAFFREZIC, F., LALOE, D., GUT, I., MOAZAMI-GOUDARZI, K. & FOULLEY, J. L. 2009. A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. *BMC Genomics,* 10**,** 550.

GAUTIER, M. & NAVES, M. 2011. Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Mol. Ecol.,* 20**,** 3128-43.

GODDARD, M. E. & HAYES, B. J. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.,* 10**,** 381-91.

GORLOV, I. P., KAMAT, A., BOGATCHEVA, N. V., JONES, E., LAMB, D. J., TRUONG, A., BISHOP, C. E., MCELREAVEY, K. & AGOULNIK, A. I. 2002. Mutations of the *GREAT* gene cause cryptorchidism. *Hum. Mol. Genet.,* 11**,** 2309-18.

HAI, T., TENG, F., GUO, R., LI, W. & ZHOU, Q. 2014. One-step generation of knockout pigs by zygote injection of CRISPR/Cas system. *Cell Res.,* 24**,** 372-5.

HANOTTE, O., BRADLEY, D. G., OCHIENG, J. W., VERJEE, Y., HILL, E. W. & REGE, J. E. 2002. African pastoralism: genetic imprints of origins and migrations. *Science,* 296**,** 336-9.

HANOTTE, O., DESSIE, T. & KEMP, S. 2010. Ecology. Time to tap Africa's livestock genomes. *Science,* 328**,** 1640-1.

HANSEN, P. J. 2007. Exploitation of genetic and physiological determinants of embryonic resistance to elevated temperature to improve embryonic survival in dairy cattle during heat stress. *Theriogenology,* 68 Suppl 1**,** S242-9.

HOU, Y., BICKHART, D. M., HVINDEN, M. L., LI, C., SONG, J., BOICHARD, D. A., FRITZ, S., EGGEN, A., DENISE, S., WIGGANS, G. R., SONSTEGARD, T. S., VAN TASSELL, C. P. & LIU, G. E. 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.

IPCC, 2014: Summary for Policymakers. In: Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Edenhofer, O., R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlomer, C. von Stechow, T. Zwickel and J.C. Minx (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

JOHNSTON, S., MCEWAN, J., PICKERING, N., KIJAS, J., BERALDI, D., PILKINGTON, J., PEMBERTON, J. & SLATE, J. 2011. Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. *Mol. Ecol.,* 20**,** 2555 - 2566.

KEIGHTLEY, P. D. & EYRE-WALKER, A. 2000. Deleterious mutations and the evolution of sex. *Science,* 290**,** 331-3.

KLAMBAUER, G., SCHWARZBAUER, K., MAYR, A., CLEVERT, D. A., MITTERECKER, A., BODENHOFER, U. & HOCHREITER, S. 2012. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.,* 40**,** e69.

KOUDANDE, O. D., VAN ARENDONK, J. A. & IRAQI, F. 2005. Marker-assisted introgression of trypanotolerance QTL in mice. *Mamm. Genome,* 16**,** 112-9.

KRUMM, N., SUDMANT, P. H., KO, A., O'ROAK, B. J., MALIG, M., COE, B. P., QUINLAN, A. R., NICKERSON, D. A. & EICHLER, E. E. 2012. Copy number variation detection and genotyping from exome sequence data. *Genome Res.,* 22**,** 1525-32.

LECOMTE, L., DUFFE, P., BURET, M., SERVIN, B., HOSPITAL, F. & CAUSSE, M. 2004. Marker-assisted introgression of five QTLs controlling fruit quality traits into three tomato lines revealed interactions between QTLs and genetic backgrounds. *Theor. Appl. Genet.,* 109**,** 658-68.

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., SONSTEGARD, T. S. & VAN TASSELL, C. P. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One,* 4**,** e5350.

MURRAY, M. & TRAIL, J. C. M. 1984. Genetic resistance to animal trypanosomiasis in Africa. *Preventive Veterinary Medicine,* 2**,** 541-551.

NARDONE, A., RONCHI, B., LACETERA, N., RANIERI, M. S. & BERNABUCCI, U. 2010. Effects of climate changes on animal production and sustainability of livestock systems. *Livestock Science,* 130**,** 57-69.

NORVAL, R. A. I., PERRY, B. D. & A.S., Y. 1992. *The Epidemiology of Theileriosis in Africa,* London, Academic Press.

OLEKSYK, T. K., SMITH, M. W. & O'BRIEN, S. J. 2010. Genome-wide scans for footprints of natural selection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.,* 365**,** 185-205.

PLAGNOL, V., CURTIS, J., EPSTEIN, M., MOK, K. Y., STEBBINGS, E., GRIGORIADOU, S., WOOD, N. W., HAMBLETON, S., BURNS, S. O., THRASHER, A. J., KUMARARATNE, D., DOFFINGER, R. & NEJENTSEV, S. 2012. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics,* 28**,** 2747-54.

PROUDFOOT, C., CARLSON, D. F., HUDDART, R., LONG, C. R., PRYOR, J. H., KING, T. J., LILLICO, S. G., MILEHAM, A. J., MCLAREN, D. G., WHITELAW, C. B. & FAHRENKRUG, S. C. 2015. Genome edited sheep and cattle. *Transgenic Res.,* 24**,** 147-53.

REGE, J. E. O., KAHI, A., M., O.-A., MWACHARO, J. & HANOTTE, O. 2001. *Zebu cattle of Kenya: Uses, performance, farmer preferences and measures of genetic diversity.* Nairobi, Kenya, International Livestock Reaserch Institute.

SATHIRAPONGSASUTI, J. F., LEE, H., HORST, B. A., BRUNNER, G., COCHRAN, A. J., BINDER, S., QUACKENBUSH, J. & NELSON, S. F. 2011. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics,* 27**,** 2648-54.

SATO, S., OHNISHI, C., KIKUCHI, T., KOHIRA, K., EGAWA, S., TERAI, S., NAKAMURA, T., ARATA, S., KOMATSUDA, A. & UEMOTO, Y. 2014. Evaluation of quantitative trait loci affecting intramuscular fat and reproductive traits in pigs using marker-assisted introgression. *Anim. Genet.,* 45**,** 799-807.

SEO, S. N. & MENDELSOHN, R. 2008. Measuring impacts and adaptations to climate change: a structural Ricardian model of African livestock management1. *Agricultural Economics,* 38**,** 151-165.

TAN, W., CARLSON, D. F., LANCTO, C. A., GARBE, J. R., WEBSTER, D. A., HACKETT, P. B. & FAHRENKRUG, S. C. 2013. Efficient nonmeiotic allele introgression in livestock using custom endonucleases. *PNAS,* 110**,** 16526-31.

WANG, H., YANG, H., SHIVALILA, C. S., DAWLATY, M. M., CHENG, A. W., ZHANG, F. & JAENISCH, R. 2013. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell,* 153**,** 910-8.

WITTMANN, E. J., MELLOR, P. S. & BAYLIS, M. 2001. Using climate data to map the potential distribution of Culicoides imicola (Diptera: Ceratopogonidae) in Europe. *Rev. Sci. Tech.,* 20**,** 731-40.

XIE, C. & TAMMI, M. T. 2009. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics,* 10**,** 80.

Supplementary Tables

Chapter Two

Table S2.1: Candidate regions identified by pairwise and combined reference *Rsb* analyses and pairwise F_{ST} analyses.

Rsb	**EASZ vs European taurine**	**EASZ vs N'Dama**	**EASZ vs Nellore**	**EASZ vs combined references**
BTA	**Chromosomal position (bp)**			
3	_	_	_	101,942,771
5	_	_	_	57,138,952
5	_	60,513,092	_	60,556,520
5	_	113,737,833	_	_
11	62,629,106	62,629,106*	_	62,629,106
12	27,181,474	_	_	27,181,474
12	29,217,254*	29,217,254	_	29,217,254
12	35,740,174*	_	_	35,740,174
18	_	13,238,432	_	_
19	42,252,751*	_	_	42,696,851
F_{ST}				
BTA	**Chromosomal position (bp)**			
2	_	125,585,810 - 126,058,677		
4	_	_	47,195,467 - 47,539,595	
4	51,927,595 - 52,308,430	_	_	
7	52,224,595 - 52,720,797	52,224,595 - 52,720,797	_	
13	46,433,697 - 46,723,493	46,433,697 - 46,723,493	_	
13	57,848,276 - 58,207,174	_	_	
14	24,482,969-25,254,540	_	_	
19	_	27,369,763 - 27,763,447	_	
22	_	_	2,314,019 - 2,788,566	
24	_	_	4,118,163 - 4,474,760	
X	_	_	8,582,093 - 9,248,137	
X	40,319,976 - 43,999,854	39,942,044 - 42,024,368	_	
X	84,566,018 - 85,993,719	84,566,018 - 85,993,719	_	

* Single SNP passed significant threshold.

Table S2.2: Genes within the candidate regions intervals

The file is an electronic version

Table S2.3: Functional term clusters of genes within the identified candidate region intervals.

The file is an electronic version

Table S2.4: Bovine QTL spans the identified candidate regions intervals.

The file is an electronic version

Table S2.5: Trypanotolerance QTL identified by (Hanotte *et al*., 2003) spanning the candidate regions intervals

BTA	**start**	**stop**	**QTL_ID**
7	38,332,974	59,135,155	PCVF minus PCVM QTL (10516)
7	38,332,974	59,135,155	Body weight (mean) QTL (10517)
7	38,332,974	59,135,155	Percentage decrease in body weight up to day 150 after challenge QTL (10518)
7	38,332,974	59,135,155	Parasites natural logarithm of mean number QTL (10519)
7	38,332,974	59,135,155	Parasite detection rate QTL (10520)
13	29,549,346	71,827,292	Percentage decrease in PCV up to day 150 after challenge QTL (10524)
13	29,549,346	71,827,292	Percentage decrease in PCV up to day 100 after challenge QTL (10525)
13	29,549,346	71,827,292	Parasite detection rate QTL (10526)
29	1,568,804	25,983,377	BWF scaled by BWI QTL (10559)
29	1,568,804	25,983,377	Parasites natural logarithm of mean number QTL (10560)

Chapter Three

Table S3.1: The EASZ samples of the full genome sequence pool, and the ten exome sequences. Their sublocations, sex and alive/dead status after one year from their date of birth.

Full genome sequence			
Sample ID (DEDG number)	**Sublocation**	**Sex**	**Alive/ Dead**
354	Namboboto	male	**Alive**
90	Bukati	male	**Alive**
500	Bujwanga	female	**Alive**
341	Kidera	female	**Alive**
364	Otimong	female	**Alive**
352	Magombe East	male	**Dead**
378	Bumala	male	**Dead**
70	Mabusi	male	**Dead**
194	Kokare	female	**Dead**
18	Bujwanga	female	**Dead**
Exome sequence			
Sample ID (DEDG number)	**Sublocation**	**Sex**	**Alive/ Dead**
289	Kidera	female	**Dead**
1693	Otimong	male	**Alive**
524	MagombeEast	female	**Dead**
915	SimurEast	female	**Alive**
1148	Bujwanga	female	**Dead**
2025	Bujwanga	male	**Alive**
2063	OjwandoB	male	**Dead**
1401	OjwandoB	male	**Dead**
923	Ikonzo	female	**Dead**
2183	Ikonzo	male	**Alive**

Table S3.2: A summary of the exome sequence reads mapped to the UMD3.1 bovine reference genome for each sequenced EASZ sample.

Sample ID (DEDG number)	**289**	**524**	**915**	**923**	**1148**	**1401**	**1693**	**2025**	**2063**	**2183**
Total Raw Reads	60,521,890	61,886,159	76,115,987	71,041,457	54,406,488	70,583,687	75,825,024	97,926,144	56,892,642	81,334,699
Mapped Reads (MAPQ0)	58,123,736	59,259,487	72,520,518	67,565,041	52,179,782	67,757,017	72,357,186	93,087,864	54,589,349	77,623,887
Total MAPQ30 Aligned Reads	52,530,808	53,401,659	65,074,820	60,547,213	47,286,573	61,153,497	65,090,879	84,299,461	49,335,902	69,825,532
Reads In Targets:	46,060,296	47,152,655	56,371,737	52,981,396	41,868,685	54,020,004	57,191,616	75,303,053	43,725,770	61,038,375
Reads Off Targets:	6,470,512	6,249,004	8,703,083	7,565,817	5,417,888	7,133,493	7,899,263	8,996,408	5,610,132	8,787,157
Percent of Target Bases Not Covered	8.58%	8.68%	7.77%	7.56%	9.23%	8.30%	7.90%	6.31%	8.83%	7.26%
Percent of Target Bases Covered	91.42%	91.32%	92.23%	92.44%	90.77%	91.70%	92.10%	93.69%	91.17%	92.74%

Table S3.3: Candidate regions from each genome-wide SNP analysis (*iHS*, *Rsb*, ΔAF and *meta-SS*) conducted on EASZ with the combined reference populations (Holstein-Friesian, Jersey, N’Dama, Muturu, Nellore and Gir).

iHS	***Rsb***	ΔAF	***meta-SS***
BTA 7:52824870-52929978	BTA 3:120604441-121240490	BTA 5:48603538-49209163	BTA 1:21156402-22526511
	BTA 5:43269121-44220056	BTA 7:52183599-52265269	BTA 1:54859494-55507566
	BTA 5:48585608-49161124	BTA 13:47508061-48142997	BTA 1:149241884-149992523
	BTA 5:62467698-62644905	BTA 13:48796718-48911554	BTA 1:150577829-151624225
	BTA 7:32972540-33427191	BTA 13:49433476-49762965	BTA 2:70314631-71161113
	BTA 8:23344221-23663852	BTA 13:50524278-50890135	BTA 2:125159084-125994861
	BTA 8:75877341-76036999		BTA 2:129113592-129923930
	BTA 9:70621641-70687691		BTA 3:12494224-12881786
	BTA 11:62482731-62765277		BTA 3:34007860-34727876
	BTA 12:35689908-35743087		BTA 3:76084701-76781970
	BTA 13:40486503-40909800		BTA 3:84969287-85118197
	BTA 14:28205941-28430215		BTA 3:98862402-99422213
	BTA 16:24578091-25533972		BTA 3:117829869-118619968
	BTA 19:2568979-2765065		BTA 3:119780513-121238836
	BTA 19:40668949-42066750		BTA 4:31365784-31765553
	BTA 19:46361364-46602836		BTA 4:63641280-63815686
	BTA 22:45134706-46239607		BTA 4:66591996-67082312
	BTA 24:61972128-62488062		BTA 5:23652016-24473786
	BTA 27:4653655-4975928		BTA 5:43230619-44574214
			BTA 5:48477903-49268610
			BTA 5:56651062-57515653
			BTA 5:58604207-58841085
			BTA 5:62272683-62659987
			BTA 5:109303999-110096347
			BTA 7:31748136-33875610
			BTA 7:50281923-50809190
			BTA 7:52183599-53001042
			BTA 7:61232987-61658196
			BTA 7:62415406-63117931
			BTA 7:65078295-65614779
			BTA 7:72421285-72514475
			BTA 7:84899824-85607137
			BTA 8:22905793-24288281
			BTA 8:44479612-44734659
			BTA 8:54926810-55060853
			BTA 8:64248747-64461984
			BTA 8:65301113-65634601
			BTA 8:70046025-70695580

			BTA 8:75162283-76369374
			BTA 9:69198185-69475040
			BTA 9:70454678-70753715
			BTA 9:73280867-74185868
			BTA 9:76289561-76853587
			BTA 9:85550547-85655401
			BTA 9:87580569-87599592
			BTA 9:94121197-95380876
			BTA 9:104535674-105668863
			BTA 10:80515703-80833218
			BTA 10:103294561-104290655
			BTA 11:38402190-39743107
			BTA 11:61877437-62548419
			BTA 11:71346519-72547901
			BTA 11:73807783-75446726
			BTA 12:20989169-21254061
			BTA 12:24843013-25658768
			BTA 12:27050192-29151436
			BTA 12:33250917-34546381
			BTA 12:35445176-36965854
			BTA 13:18130223-18421481
			BTA 13:38040277-38561478
			BTA 13:39430011-41356847
			BTA 13:42225984-43136553
			BTA 13:45350218-45772147
			BTA 13:50616630-50837529
			BTA 13:55428804-55542599
			BTA 13:58273562-58599491
			BTA 13:81265605-82376453
			BTA 14:28186226-28430215
			BTA 15:42536074-43195359
			BTA 16:24517859-25540339
			BTA 16:26979772-27160301
			BTA 16:40523561-41395850
			BTA 16:46869577-47614377
			BTA 16:50610769-50762363
			BTA 16:56620669-57041236
			BTA 18:13483509-14050131

			BTA 18:19446425-20061183
			BTA 19:2568979-2765065
			BTA 19:3337282-3823638
			BTA 19:9515063-10250080
			BTA 19:20622520-21225583
			BTA 19:26909816-27143239
			BTA 19:30826413-31463726
			BTA 19:39330233-39519992
			BTA 19:40045779-42066750
			BTA 19:43023638-43692285
			BTA 19:44788419-45414418
			BTA 19:46031543-46786391
			BTA 21:33590777-33696403
			BTA 21:60026698-60449172
			BTA 22:29533544-30366810
			BTA 22:42705202-44541377
			BTA 22:45102551-46400273
			BTA 24:61008938-62530799
			BTA 25:41769025-42283544
			BTA 27:4494286-5052515
			BTA 27:6938210-7107679
			BTA 28:4635419-5123022

Table S3.4: Pearson correlation coefficients (r) between the *P*-values of the three genome-wide SNP analyses (*iHS*, *Rsb* and Δ*AF*) conducted on EASZ and the zebu cattle populations from Uganda and Nigeria.

EASZ	***Rsb***	***iHs***	**Δ*AF***
Rsb	1	0.228	0.098
iHs	0.228	1	0.133
Δ*AF*	0.098	0.133	1

Uganda	***Rsb***	***iHs***	**Δ*AF***
Rsb	1	0.22	0.08
iHs	0.22	1	0.116
Δ*AF*	0.08	0.116	1

Nigeria	***Rsb***	***iHs***	**Δ*AF***
Rsb	1	0.22	0.077
iHs	0.22	1	0.122
Δ*AF*	0.077	0.122	1

Table S3.5: Candidate regions identified by the separate *meta-SS* analyses of EASZ with the combined reference cattle populations (Holstein-Friesian, Jersey, Guinean N'Dama, Muturu, Nellore and Gir), European taurine (Holstein-Friesian and Jersey), African taurine (N'Dama and Muturu) and Asian zebu (Nellore and Gir).

EASZ *vs.* combined reference	**EASZ *vs.* European taurine**	**EASZ *vs.* African taurine**	**EASZ *vs.* Asian zebu**
BTA 1:21156402-22526511	BTA 1:43720640-43985755	BTA 1:21273311-22733168	BTA 1:21858918-22046473
BTA 1:54859494-55507566	BTA 1:54859494-55507566	BTA 1:149241884-149992523	BTA 1:119051755-119624353
BTA 1:149241884-149992523	BTA 1:149187317-149992523	BTA 1:150701102-151801139	BTA 1:131228416-131797328
BTA 1:150577829-151624225	BTA 1:150701102-151546627	BTA 2:63595920-63786435	BTA 2:38756568-38949731
BTA 2:70314631-71161113	BTA 2:70314631-71185300	BTA 2:70314631-71185300	BTA 2:89312957-89889861
BTA 2:125159084-125994861	BTA 2:125435695-125994861	BTA 2:120552241-120684646	BTA 2:124919532-125201684
BTA 2:129113592-129923930	BTA 3:12387203-12856963	BTA 2:125159084-125994861	BTA 2:128258499-129103667
BTA 3:12494224-12881786	BTA 3:57647858-57821484	BTA 2:127257140-128368124	BTA 3:15357241-15448231
BTA 3:34007860-34727876	BTA 3:66396281-66830763	BTA 3:12427792-12856963	BTA 3:118235551-118619968
BTA 3:76084701-76781970	BTA 3:75903912-76781970	BTA 3:57647858-57955899	BTA 3:120514654-120933064
BTA 3:84969287-85118197	BTA 3:84969287-85118197	BTA 3:66396281-66684875	BTA 5:101216012-101377388
BTA 3:98862402-99422213	BTA 3:94387038-94651204	BTA 3:75879604-76825243	BTA 6:4876731-5213570
BTA 3:117829869-118619968	BTA 3:98862402-99422213	BTA 3:84984361-85159851	BTA 7:72421285-72514475
BTA 3:119780513-121238836	BTA 3:120239456-121336001	BTA 3:93932193-94651204	BTA 8:23393589-23705690
BTA 4:31365784-31765553	BTA 4:31365784-31901772	BTA 3:98862402-99422213	BTA 8:43890714-44734659
BTA 4:63641280-63815686	BTA 4:66591996-66981450	BTA 3:102077605-102677147	BTA 8:45522131-46530833
BTA 4:66591996-67082312	BTA 5:23652016-24473786	BTA 3:117382593-117829869	BTA 8:48413463-48996862
BTA 5:23652016-24473786	BTA 5:43377141-44220056	BTA 3:120604441-120933064	BTA 8:75323318-76140060
BTA 5:43230619-44574214	BTA 5:47583595-47889976	BTA 4:63641280-63815686	BTA 9:75528881-75644797
BTA 5:48477903-49268610	BTA 5:48477903-49268610	BTA 4:66717413-67074925	BTA 9:77427348-77608380
BTA 5:56651062-57515653	BTA 5:56716286-57504991	BTA 4:70466113-70980770	BTA 9:87513863-88558239
BTA 5:58604207-58841085	BTA 5:62483017-62659987	BTA 5:23652016-24338695	BTA 9:94121197-94202901
BTA 5:62272683-62659987	BTA 5:66466104-66839402	BTA 5:25696803-25956523	BTA 13:19230559-19283084
BTA 5:109303999-110096347	BTA 5:109303999-109917484	BTA 5:43269121-43928536	BTA 13:34447404-34673102
BTA 7:31748136-33875610	BTA 5:112892862-113026417	BTA 5:48409250-49268610	BTA 15:44861498-44957064
BTA 7:50281923-50809190	BTA 6:107869349-108180993	BTA 5:58604207-58841085	BTA 15:47056199-47426249
BTA 7:52183599-53001042	BTA 7:33085021-33427191	BTA 5:62483017-62659987	BTA 19:2473530-3880072
BTA 7:61232987-61658196	BTA 7:52183599-53104028	BTA 5:109457734-109948881	BTA 19:30788891-31093921
BTA 7:62415406-63117931	BTA 7:62551178-63113709	BTA 5:112892862-113062786	BTA 19:40342087-41622554
BTA 7:65078295-65614779	BTA 7:84899824-85607137	BTA 6:107869349-108053301	BTA 20:14483323-14769267
BTA 7:72421285-72514475	BTA 8:22898801-23663852	BTA 7:6194849-6324756	BTA 20:39393926-39441079
BTA 7:84899824-85607137	BTA 8:25186672-25251326	BTA 7:32640500-33478056	BTA 24:4549374-4741755

BTA 8:22905793-24288281	BTA 8:54926810-54967023	BTA 7:52183599-53001042	BTA 28:4636092-4743614
BTA 8:44479612-44734659	BTA 9:69198185-69434969	BTA 7:62415406-63113709	
BTA 8:54926810-55060853	BTA 9:70454678-70753715	BTA 7:84821734-84942326	
BTA 8:64248747-64461984	BTA 9:104535674-105573011	BTA 7:92566598-93032329	
BTA 8:65301113-65634601	BTA 10:80571881-80833218	BTA 8:7045806-7220044	
BTA 8:70046025-70695580	BTA 10:103224277-104290655	BTA 8:22898801-23703357	
BTA 8:75162283-76369374	BTA 11:61842250-62548419	BTA 8:54926810-54967023	
BTA 9:69198185-69475040	BTA 11:71321015-71959918	BTA 8:96511836-96607609	
BTA 9:70454678-70753715	BTA 11:74590729-75446726	BTA 9:61515334-61560160	
BTA 9:73280867-74185868	BTA 12:20903217-21254061	BTA 9:69083365-69406467	
BTA 9:76289561-76853587	BTA 12:24843013-26501234	BTA 9:70454678-70996872	
BTA 9:85550547-85655401	BTA 12:27050192-27481578	BTA 9:95243509-95398119	
BTA 9:87580569-87599592	BTA 12:27998214-29572256	BTA 9:104535674-105668863	
BTA 9:94121197-95380876	BTA 12:35445176-35920785	BTA 10:79699073-79969610	
BTA 9:104535674-105668863	BTA 13:18087066-18401050	BTA 10:80571881-80833218	
BTA 10:80515703-80833218	BTA 13:38077663-38561478	BTA 10:103224277-104290655	
BTA 10:103294561-104290655	BTA 13:40486503-41356847	BTA 11:62198651-62410583	
BTA 11:38402190-39743107	BTA 14:28186226-28371763	BTA 11:64392003-65111327	
BTA 11:61877437-62548419	BTA 16:24517859-25561046	BTA 11:73897030-75446726	
BTA 11:71346519-72547901	BTA 16:47310187-47946214	BTA 12:21105889-21364886	
BTA 11:73807783-75446726	BTA 16:50589449-50936489	BTA 12:25020864-25753544	
BTA 12:20989169-21254061	BTA 16:56614124-57126487	BTA 12:27159435-27294344	
BTA 12:24843013-25658768	BTA 18:13570777-14050131	BTA 12:27998214-29572256	
BTA 12:27050192-29151436	BTA 19:2631478-2748282	BTA 12:31767663-31856909	
BTA 12:33250917-34546381	BTA 19:9515063-9866038	BTA 12:35711647-36163614	
BTA 12:35445176-36965854	BTA 19:24020434-24240860	BTA 13:18130223-18401050	
BTA 13:18130223-18421481	BTA 19:26909816-27113818	BTA 13:40260771-41328969	
BTA 13:38040277-38561478	BTA 19:31260078-31463726	BTA 13:50578262-51056217	
BTA 13:39430011-41356847	BTA 19:40277088-41570503	BTA 13:57900251-58599491	
BTA 13:42225984-43136553	BTA 19:44701807-45606680	BTA 13:81326857-82765411	
BTA 13:45350218-45772147	BTA 19:46427077-46786391	BTA 16:24517859-25616669	
BTA 13:50616630-50837529	BTA 21:60108089-60449172	BTA 16:27045486-27160301	
BTA 13:55428804-55542599	BTA 21:67053720-67771364	BTA 16:40938961-41888513	
BTA 13:58273562-58599491	BTA 22:29992570-30416885	BTA 16:47516943-47949171	
BTA 13:81265605-82376453	BTA 22:42705202-44541377	BTA 16:50589449-50762363	
BTA 14:28186226-28430215	BTA 22:45102551-46400273	BTA 16:56606500-56887107	

BTA 15:42536074-43195359	BTA 23:18307738-18594576	BTA 18:13411086-14050131	
BTA 16:24517859-25540339	BTA 24:43869361-43952239	BTA 18:41782002-41883323	
BTA 16:26979772-27160301	BTA 24:61008938-62172371	BTA 19:9515063-9866038	
BTA 16:40523561-41395850	BTA 25:41715836-42452921	BTA 19:23766690-24240860	
BTA 16:46869577-47614377		BTA 19:26909816-27131960	
BTA 16:50610769-50762363		BTA 19:31242346-31463726	
BTA 16:56620669-57041236		BTA 19:39342925-41700252	
BTA 18:13483509-14050131		BTA 19:42808886-43406339	
BTA 18:19446425-20061183		BTA 19:44836581-45134496	
BTA 19:2568979-2765065		BTA 19:46487440-46786391	
BTA 19:3337282-3823638		BTA 20:71168947-71977818	
BTA 19:9515063-10250080		BTA 21:67360105-67771364	
BTA 19:20622520-21225583		BTA 22:29654676-30237259	
BTA 19:26909816-27143239		BTA 22:43706894-44292266	
BTA 19:30826413-31463726		BTA 22:45102551-46400273	
BTA 19:39330233-39519992		BTA 23:18307738-18594576	
BTA 19:40045779-42066750		BTA 24:42318083-43053802	
BTA 19:43023638-43692285		BTA 24:43641560-43952239	
BTA 19:44788419-45414418		BTA 25:37361829-37759374	
BTA 19:46031543-46786391		BTA 25:41769025-42061192	
BTA 21:33590777-33696403		BTA 26:22075366-22274137	
BTA 21:60026698-60449172		BTA 26:30849958-30917539	
BTA 22:29533544-30366810		BTA 28:28503464-28838715	
BTA 22:42705202-44541377		BTA 29:43412008-44067968	
BTA 22:45102551-46400273			
BTA 24:61008938-62530799			
BTA 25:41769025-42283544			
BTA 27:4494286-5052515			
BTA 27:6938210-7107679			
BTA 28:4635419-5123022			

Table S3.6: Candidate regions from each genome-wide SNP analysis (*iHS*, *Rsb*, Δ*AF* and *meta-SS*) performed on zebu cattle populations from Uganda with the combined reference populations (Holstein-Friesian, Jersey, N'Dama, Muturu, Nellore and Gir).

iHS	*Rsb*	Δ*AF*	*meta-SS*
BTA 1:67180060-67313350	BTA 3:75775876-76364422	BTA 13:48796718-48911554	BTA 1:58355682-58612631
BTA 3:120559807-121238836	BTA 3:120604441-121247679	BTA 13:49433476-49762965	BTA 1:65792291-66240276
BTA 5:47121805-47437350	BTA 5:43829015-44574214	BTA 13:50524278-50742202	BTA 1:66748211-67328094
BTA 6:4964608-5141190	BTA 5:48149086-49174693		BTA 1:149241884-149960460
BTA 13:40360159-41102866	BTA 5:60390408-60627450		BTA 2:70314631-71185300
BTA 19:46606725-46706542	BTA 5:62457573-62644905		BTA 2:125159084-125994861
	BTA 5:66758715-66864425		BTA 3:33458232-33531571
	BTA 6:4837150-4876731		BTA 3:34254043-34727876
	BTA 6:40303148-40934693		BTA 3:67940307-68128218
	BTA 7:32640500-33427191		BTA 3:75391246-76825243
	BTA 8:45762230-46406194		BTA 3:98563369-99283161
	BTA 9:70649533-70687691		BTA 3:120601191-121300696
	BTA 11:62482731-62765277		BTA 5:23652016-24338695
	BTA 12:25490082-25676204		BTA 5:43834751-44574214
	BTA 12:26945295-27085108		BTA 5:47143913-49212943
	BTA 12:35689908-35886159		BTA 5:49857890-49993761
	BTA 13:31691092-31787896		BTA 5:50952073-51113065
	BTA 13:39579929-41328969		BTA 5:60390408-60699053
	BTA 14:28164594-28430215		BTA 5:62045272-62587423
	BTA 15:44849752-45106327		BTA 5:63555403-64334857
	BTA 15:48585759-48841386		BTA 5:65268314-66966194
	BTA 19:2473530-3913240		BTA 5:109303999-109688098
	BTA 20:3206730-3568975		BTA 6:4876731-5065951
	BTA 22:29533544-30218613		BTA 6:16514162-16839617
	BTA 24:61968531-62570516		BTA 7:680008-814340
			BTA 7:31130800-35059248
			BTA 7:50029472-50809190
			BTA 7:61232987-61396966
			BTA 7:62551178-62782874
			BTA 7:100845937-102014448
			BTA 8:23344221-23663852
			BTA 8:45755851-46530833
			BTA 8:60315332-60591587

			BTA 8:65373897-65634601
			BTA 9:69198185-69406467
			BTA 9:72738853-74253900
			BTA 9:75053662-76853587
			BTA 9:88390838-88469911
			BTA 9:94121197-94242831
			BTA 10:25708993-25852549
			BTA 10:76573284-77211431
			BTA 10:80515703-80796559
			BTA 11:6784799-6924499
			BTA 11:37968468-39743107
			BTA 11:61868285-62784490
			BTA 11:71387248-72221099
			BTA 11:94562083-95050103
			BTA 12:21086969-21306618
			BTA 12:24813715-26180618
			BTA 12:28949354-29572256
			BTA 12:35689908-36746504
			BTA 12:90591899-90828989
			BTA 13:18130223-18421481
			BTA 13:31639422-32688842
			BTA 13:39579929-41363364
			BTA 13:47532424-49197290
			BTA 13:50490155-50837529
			BTA 13:51339718-51522874
			BTA 13:58270096-58599491
			BTA 14:27987402-28430215
			BTA 15:44843650-45012875
			BTA 15:48555721-48841386
			BTA 15:52859008-52964463
			BTA 15:63854687-64454641
			BTA 16:25241257-25540339
			BTA 16:26807748-27160301
			BTA 16:46869577-47614377
			BTA 16:50138923-50762363
			BTA 19:1831783-3880072
			BTA 19:9515063-9780078

			BTA 19:26704580-27154113
			BTA 19:39330233-40808559
			BTA 19:44438549-47446995
			BTA 19:50168911-50624183
			BTA 20:3206730-4121521
			BTA 20:39393926-39537515
			BTA 21:33590777-33696403
			BTA 21:60026698-60449172
			BTA 21:64366520-64391303
			BTA 21:70668775-71102609
			BTA 22:29533544-30366810
			BTA 22:31696964-32034690
			BTA 22:45231901-46400273
			BTA 23:8124702-8664162
			BTA 24:61972128-62573437
			BTA 26:39451175-39523027

Table S3.7: Candidate regions from each genome-wide SNP analysis (*iHS*, *Rsb*, Δ*AF* and *meta-SS*) performed on zebu cattle populations from Nigeria with the combined reference populations (Holstein-Friesian, Jersey, N’Dama, Muturu, Nellore and Gir).

iHS	***Rsb***	**Δ*AF***	***meta-SS***
BTA 5:47241942-48399194	BTA 2:572692-997590	BTA 5:47143913-49209163	BTA 1:33254585-33443559
BTA 5:48930543-49018256	BTA 2:70314631-71209253	BTA 7:51259460-53502939	BTA 1:36856368-37557262
BTA 5:76084996-76200813	BTA 5:27274794-27335812	BTA 7:54290110-55300310	BTA 1:43720640-43985755
BTA 6:4964608-5090033	BTA 5:47112945-49253605	BTA 12:29227227-29572256	BTA 1:70234565-70425655
	BTA 5:56716286-57930050		BTA 1:147319412-147434939
	BTA 5:58516607-58841085		BTA 1:149547998-150039543
	BTA 5:60374628-60444745		BTA 1:150701102-151399165
	BTA 5:62045272-62379266		BTA 2:307743-1038299
	BTA 7:773076-814340		BTA 2:70296102-71292148
	BTA 7:54054385-54422398		BTA 3:14111392-14659428
	BTA 11:18685487-18979192		BTA 3:75903912-76413468
	BTA 12:27816136-29615976		BTA 3:84683280-85159851
	BTA 13:57490777-57902524		BTA 3:98750825-99483458
	BTA 15:46020475-46333987		BTA 3:101308813-102471446
	BTA 18:13404310-13925544		BTA 4:49962548-50223466
	BTA 18:15396378-15517277		BTA 4:53809476-54112087
	BTA 19:2506362-2748282		BTA 4:55085910-57531310
	BTA 19:51832293-51964299		BTA 4:63669888-63915653
	BTA 20:11164837-11338515		BTA 4:66698519-67082312
	BTA 22:44480638-44553450		BTA 5:15612090-16031807
	BTA 24:43936586-44060913		BTA 5:17855007-18400249
	BTA 26:45869170-45878388		BTA 5:19585877-20210838
			BTA 5:43230619-44574214
			BTA 5:46938597-49993761
			BTA 5:56006965-57930050
			BTA 5:58604207-58874619
			BTA 5:60250896-60610616
			BTA 5:62146394-63255485
			BTA 5:71095194-71137918
			BTA 5:72195598-72290257
			BTA 5:77640688-78095336
			BTA 5:120826563-121179132
			BTA 6:4782814-4876731
			BTA 6:36964005-37577335
			BTA 6:48105787-49139974
			BTA 6:78353035-78490912

			BTA 6:79803715-81241492
			BTA 6:83529875-84200299
			BTA 6:95134468-95192633
			BTA 7:680008-814340
			BTA 7:20376230-20479996
			BTA 7:32640500-33093884
			BTA 7:50029472-50670070
			BTA 7:51252285-52178639
			BTA 7:52865974-54497796
			BTA 7:56112647-56311087
			BTA 7:62415406-62941787
			BTA 8:21615512-21733252
			BTA 8:45755851-46131855
			BTA 8:53262721-53965241
			BTA 8:54724271-55231867
			BTA 8:58080010-58601842
			BTA 8:60004919-60608370
			BTA 8:96337832-96616451
			BTA 10:20524518-20813721
			BTA 10:26735446-27036997
			BTA 10:57626493-57815351
			BTA 10:93813158-93886902
			BTA 10:101246708-101611132
			BTA 10:103171282-103846097
			BTA 11:18497556-19272280
			BTA 11:47991402-49439232
			BTA 11:62343547-62919865
			BTA 12:27998214-29572256
			BTA 12:89754107-91041467
			BTA 13:18132557-18320265
			BTA 13:55428804-55616661
			BTA 13:56893876-58090472
			BTA 14:13392060-14313552
			BTA 14:30997027-31287089
			BTA 14:49149502-49374173
			BTA 15:41920165-42568640
			BTA 15:63026020-63383174
			BTA 16:25389029-25889912
			BTA 16:50138923-50936489
			BTA 18:13878200-14454853

			BTA 19:2555043-2765065
			BTA 19:27004483-27435200
			BTA 19:34613439-35140272
			BTA 19:44788419-44924467
			BTA 19:46580102-46673984
			BTA 20:3435820-3704975
			BTA 20:9913289-11170115
			BTA 20:12988915-13274192
			BTA 20:70125037-71797830
			BTA 21:33590777-33696403
			BTA 21:42271864-42565638
			BTA 21:67360105-67771364
			BTA 21:68524346-69408187
			BTA 21:70885995-71196532
			BTA 22:43706894-46126149
			BTA 24:42979783-44160038
			BTA 24:53024005-53605996
			BTA 24:61259888-62530799
			BTA 25:42193923-42561781
			BTA 26:21612594-22894146
			BTA 26:45869170-45929906

Table S3.8: the genome coordinates (UMD3.1) of the overlapping genome-wide SNP candidate regions of EASZ and zebu cattle populations from Nigeria.

BTA	**start**	**stop**
1	150,701,102	151,399,165
3	84,969,287	85,118,197
4	63,669,888	63,815,686
4	66,698,519	67,082,312
5	56,651,062	57,515,653
5	58,604,207	58,841,085
7	52,865,974	53,001,042
8	54,926,810	55,060,853
10	103,294,561	103,846,097
13	55,428,804	55,542,599
15	42,536,074	42,568,640
18	13,878,200	14,050,131
22	43,706,894	44,541,377
25	42,193,923	42,283,544

Table S3.9: Gene desert candidate regions identified by EASZ genome-wide SNP analyses and *Hp* analysis.

Genome-wide SNP analyses	***Hp* analysis**
BTA 1:54,859,494-55,507,566	BTA 1:11,070,001-11,251,624
BTA 3:76,084,701-76,781,970	BTA 1:54,880,001-55,141,728
BTA 9:87,580,569-87,599,592	BTA 1:55,150,001-55,253,859
BTA 13:50,616,630-50,837,529	BTA 2:70,570,001-70,811,366
BTA 19:2,568,979-2,765,065	BTA 5:78,060,001-78,184,292
BTA 21:60,026,698-60,449,172	BTA 5:89,980,001-90,101,800
BTA 13:48,796,718-48,911,554	BTA 5:90,160,001-90,264,128
	BTA 6:51,030,001-51,137,831
	BTA 6:52,270,001-52,381,282
	BTA 6:52,810,001-53,302,557
	BTA 6:53,400,001-53,537,934
	BTA 9:57,340,001-57,441,269
	BTA 9:71,980,001-72,171,785
	BTA 9:72,890,001-72,991,383
	BTA 9:90,910,001-91,013,212
	BTA 11:39,240,001-39,530,799
	BTA 11:39,550,001-39,683,044
	BTA 12:70,760,001-71,243,560
	BTA 12:74,990,001-75,109,791
	BTA 12:82,290,001-82,473,041
	BTA 13:49,590,001-49,844,283
	BTA 13:82,010,001-82,111,606
	BTA 14:46,100,001-46,252,557
	BTA 16:38,780,001-38,881,005
	BTA 16:41,730,001-41,885,996
	BTA 17:63,070,001-63,185,472
	BTA 18:37,400,001-37,541,995
	BTA 20:49,370,001-49,599,936
	BTA 21:40,100,001-40,201,502
	BTA 23:18,980,001-19,091,451
	BTA 24:4,660,001-4,771,543
	BTA 29:34,280,001-34,380,825

Table S3.10: Genes within genome-wide SNP candidate regions: EASZ, East African zebu-sharing candidate regions, East and West African zebu-sharing candidate regions and EASZ-Nigerian zebu-sharing candidate regions.

The file is an electronic version

Table S3.11: Functional term clusters of the genes within the different types of candidate regions identified.

The file is an electronic version

Table S3.12: The coordinates of the autosomal 100 kb candidate sweep windows, number of uncovered bases within window, number of SNPs within window, window's *Hp* and ZHp values.

The file is an electronic version

Table S3.13: Genome coordinates (UMD3.1), size and mean ZHp value of candidate sweep regions in EASZ autosomes identified by the *Hp* analysis.

BTA	**start**	**stop**	**Other studies***	**size (bp)**	mean ZHp
1	11,070,001	11,251,624		181,623	-4.21
1	14,910,001	15,065,189		155,188	-4.54
1	42,100,001	42,221,164		121,163	-4.20
1	42,520,001	42,621,530		101,529	-4.10
1	50,440,001	50,622,427		182,426	-5.81
1	54,880,001	55,141,728		261,727	-6.56
1	55,150,001	55,253,859		103,858	-4.12
1	56,220,001	56,392,319		172,318	-4.53
1	66,910,001	67,064,443		154,442	-4.58
1	67,310,001	67,710,526		400,525	-6.98
1	71,180,001	71,283,103		103,102	-4.11
1	80,510,001	80,660,409		150,408	-4.37
1	127,270,001	127,400,871		130,870	-4.71
2	70,570,001	70,811,366	Liao *et al*., 2013; Kemper *et al*., 2014; Gautier *et al*., 2009	241,365	-6.68
2	70,990,001	71,191,313	Liao *et al*., 2013; Kemper *et al*., 2014; Gautier *et al*., 2009	201,312	-5.27
2	106,240,001	106,412,107		172,106	-5.22
2	111,340,001	111,461,100		121,099	-4.46
2	125,300,001	125,620,820	Gautier *et al*., 2009	320,819	-8.12
2	125,640,001	126,083,262	Gautier *et al*., 2009	443,261	-7.66
3	57,450,001	57,872,320	Kemper *et al*., 2014	422,319	-7.01
3	65,990,001	66,121,297		131,296	-4.30
4	64,060,001	64,231,634		171,633	-4.91
4	75,050,001	75,152,011		102,010	-4.02
5	24,980,001	25,100,643		120,642	-4.54
5	47,540,001	48,083,606	Liao *et al*., 2013; Kemper *et al*., 2014; Xu *et al*., 2015, Perez O'Brien *et al*., 2014	543,605	-8.12

5	48,610,001	49,021,113	Liao *et al*., 2013; Kemper *et al*., 2014; Xu *et al*., 2015, Perez O'Brien *et al*., 2014	411,112	-5.01
5	49,120,001	49,241,076	Liao *et al*., 2013; Kemper *et al*., 2014; Perez O'Brien *et al*., 2014	121,075	-4.47
5	59,510,001	59,637,265	Liao *et al*., 2013; Kemper *et al*., 2014; Gautier *et al*., 2009; Chan *et al*., 2010	127,264	-4.59
5	60,610,001	60,721,361	Liao *et al*., 2013; Kemper *et al*., 2014; Flori *et al*., 2014; Gautier *et al*., 2009; Gautier and Navas, 2011; Chan *et al*, 2010	111,360	-4.33
5	66,500,001	67,010,316	Chan *et al*., 2010	510,315	-6.20
5	68,800,001	68,992,941		192,940	-5.32
5	78,060,001	78,184,292	Larkin *et al*., 2012	124,291	-4.43
5	89,610,001	89,721,081		111,080	-4.31
5	89,980,001	90,101,800		121,799	-4.22
5	90,160,001	90,264,128		104,127	-4.02
5	114,370,001	114,544,924	Liao *et al*., 2013; Perez O'Brien *et al*., 2014	174,923	-4.97
5	120,800,001	121,199,019	Liao *et al*., 2013	399,018	-8.21
6	5,470,001	5,842,010		372,009	-4.70
6	5,930,001	6,706,107		776,106	-7.02
6	12,760,001	12,881,667		121,666	-4.27
6	51,030,001	51,137,831		107,830	-4.15
6	52,270,001	52,381,282		111,281	-4.30
6	52,810,001	53,302,557		492,556	-10.34
6	53,400,001	53,537,934		137,933	-6.01
6	60,120,001	60,321,420		201,419	-6.34
6	60,480,001	60,622,934		142,933	-4.30
6	81,600,001	81,925,350	Perez O'Brien *et al*., 2014	325,349	-5.70
6	94,070,001	94,177,765		107,764	-4.01
6	99,890,001	100,008,555		118,554	-4.17
7	440,001	561,436		121,435	-4.34
7	22,910,001	23,012,434		102,433	-4.02
7	31,740,001	31,897,059	Flori *et al*., 2014	157,058	-4.54
7	33,100,001	33,293,306		193,305	-4.43
7	44,180,001	44,474,576	Liao *et al*., 2013; Kemper *et al*., 2014	294,575	-13.65
7	51,360,001	53,362,761	Liao *et al*., 2013; Qanbari *et al*., 2014; Porto Neto *et al*., 2013; Gautier *et al*., 2009	2,002,760	-10.79

7	53,720,001	54,521,446	Gautier *et al*., 2009; Perez O'Brien *et al*., 2014; Gautier and Navas, 2011	801,445	-6.95
7	54,550,001	54,831,282	Gautier *et al*., 2009; Gautier and Navas, 2011	281,281	-5.70
7	89,130,001	89,324,444		194,443	-10.42
7	93,190,001	93,311,412		121,411	-4.11
7	93,400,001	93,502,562		102,561	-4.12
8	59,900,001	60,020,644		120,643	-4.15
9	57,340,001	57,441,269		101,268	-4.21
9	57,570,001	57,761,240		191,239	-4.48
9	71,170,001	71,419,135		249,134	-5.98
9	71,980,001	72,171,785		191,784	-6.20
9	72,890,001	72,991,383		101,382	-4.06
9	73,890,001	74,081,863		191,862	-5.45
9	76,600,001	76,876,188		276,187	-6.12
9	90,910,001	91,013,212	Larkin *et al*., 2012	103,211	-4.55
10	17,100,001	17,251,035		151,034	-4.25
10	59,110,001	59,210,489	Liao *et al*., 2014	100,488	-4.12
10	76,650,001	76,840,921		190,920	-5.41
11	7,110,001	7,261,196		151,195	-5.24
11	37,550,001	37,720,588	Kemper *et al*., 2014; Gautier *et al*., 2009	170,587	-4.57
11	39,240,001	39,530,799	Kemper *et al*., 2014; Gautier *et al*., 2009	290,798	-5.93
11	39,550,001	39,683,044	Kemper *et al*., 2014; Gautier *et al*., 2009	133,043	-4.38
11	51,480,001	51,661,703		181,702	-4.30
11	62,810,001	62,971,604		161,603	-4.89
11	64,380,001	64,590,513	Gautier and Navas, 2011; Flori *et al*., 2014	210,512	-4.69
11	75,230,001	75,441,012		211,011	-6.40
11	107,180,001	107,323,262		143,261	-12.65
12	20,870,001	21,021,506	Gautier *et al*., 2009	151,505	-5.64
12	21,130,001	21,320,859	Gautier *et al*., 2009	190,858	-4.63
12	29,110,001	29,438,417	Liao *et al*., 2013; Porto Neto *et al*., 2013; Flori *et al*., 2014; Gautier *et al*., 2009; Gautier and Navas, 2011	328,416	-5.83
12	70,760,001	71,243,560		483,559	-6.58
12	72,510,001	72,644,371		134,370	-4.40
12	74,730,001	74,973,069		243,068	-7.40
12	74,990,001	75,109,791		119,790	-4.17
12	82,290,001	82,473,041		183,040	-4.69
13	350,001	497,233		147,232	-5.50

13	5,280,001	5,574,912	Liao *et al*., 2013	294,911	-10.45
13	17,480,001	17,590,581		110,580	-4.37
13	24,480,001	24,623,450	Gautier and Navas, 2011	143,449	-4.71
13	47,980,001	48,164,495		184,494	-4.99
13	48,330,001	48,460,640		130,639	-4.96
13	48,650,001	49,056,444	Porto Neto *et al*., 2013	406,443	-10.07
13	49,340,001	49,551,378	Porto Neto *et al*., 2013	211,377	-5.80
13	49,590,001	49,844,283	Porto Neto *et al*., 2013	254,282	-6.78
13	50,240,001	50,852,056		612,055	-7.32
13	51,310,001	51,521,330		211,329	-4.73
13	51,600,001	51,800,737	Liao *et al*., 2013	200,736	-6.18
13	52,040,001	52,141,150		101,149	-4.04
13	55,510,001	55,623,671		113,670	-4.48
13	55,710,001	55,833,923		123,922	-4.43
13	57,990,001	58,122,843	Kemper *et al*., 2014; Flori *et al*., 2014	132,842	-4.89
13	82,010,001	82,111,606		101,605	-4.06
14	660,001	1,085,829		425,828	-4.43
14	2,190,001	2,405,393		215,392	-6.33
14	41,390,001	41,563,443	Flori *et al*., 2014	173,442	-5.30
14	46,100,001	46,252,557	Gautier *et al*., 2009	152,556	-4.51
14	83,120,001	83,233,061		113,060	-4.56
14	83,380,001	83,483,495		103,494	-4.10
15	59,500,001	59,741,560		241,559	-5.22
15	85,090,001	85,272,412		182,411	-5.36
16	33,120,001	33,222,179		102,178	-4.24
16	38,780,001	38,881,005	Kemper *et al*., 2014; Chan *et al*., 2010	101,004	-4.92
16	41,730,001	41,885,996	Kemper *et al*., 2014; Chan *et al*., 2010	155,995	-4.24
16	44,660,001	44,861,757	Liao *et al*., 2013; Kemper *et al*., 2014; Chan *et al*., 2010	201,756	-7.59
17	35,600,001	35,745,362		145,361	-5.19
17	35,760,001	36,031,948	Liao *et al*., 2013	271,947	-7.68
17	49,010,001	49,134,618	Perez O'Brien *et al*., 2014	124,617	-4.12
17	50,750,001	51,039,324		289,323	-7.79
17	51,130,001	51,676,138		546,137	-6.30
17	51,840,001	52,101,083		261,082	-6.02
17	63,070,001	63,185,472		115,471	-4.46
18	23,180,001	23,634,981		454,980	-8.03
18	36,500,001	36,700,717		200,716	-4.64
18	37,010,001	37,390,731		380,730	-4.40
18	37,400,001	37,541,995		141,994	-4.43
18	50,540,001	50,663,542	Gautier *et al*., 2009	123,541	-4.66
19	9,500,001	9,631,079		131,078	-4.72

19	12,460,001	12,671,997		211,996	-4.80
19	26,890,001	27,154,002	Gautier *et al*., 2009	264,001	-7.60
19	27,490,001	27,674,039	Gautier *et al*., 2009; Liao *et al*., 2013	184,038	-4.46
19	27,970,001	28,092,385	Gautier *et al*., 2009; Liao *et al*., 2013	122,384	-4.39
19	39,270,001	39,422,844		152,843	-4.42
19	40,490,001	40,714,976		224,975	-4.60
19	40,960,001	41,450,870		490,869	-5.98
19	42,890,001	43,122,753	Chan *et al*., 2010	232,752	-6.09
19	43,140,001	43,341,262	Chan *et al*., 2010	201,261	-6.53
20	2,790,001	2,895,925	Kemper *et al*., 2014	105,924	-4.02
20	48,720,001	49,031,220		311,219	-5.35
20	49,370,001	49,599,936		229,935	-6.83
21	1,830,001	1,955,744		125,743	-4.31
21	40,100,001	40,201,502		101,501	-4.09
22	2,890,001	3,057,549		167,548	-5.84
22	30,030,001	30,260,687		230,686	-6.05
22	39,720,001	40,077,687	Liao *et al*., 2013	357,686	-12.27
22	41,820,001	41,970,661	Gautier *et al*., 2009; Chan *et al*., 2010	150,660	-5.09
22	45,220,001	45,370,457	Gautier *et al*., 2009; Chan *et al*., 2010; Flori *et al*., 2014	150,456	-4.17
23	350,001	644,988	Perez O'Brien *et al*., 2014	294,987	-4.55
23	18,520,001	18,691,851		171,850	-4.36
23	18,790,001	18,890,398		100,397	-4.28
23	18,980,001	19,091,451		111,450	-4.12
23	24,460,001	24,580,718		120,717	-4.85
24	4,660,001	4,771,543		111,542	-4.24
25	39,250,001	39,406,618		156,617	-4.10
26	16,070,001	16,291,986		221,985	-5.72
26	21,070,001	21,194,173		124,172	-4.24
26	39,260,001	39,481,524		221,523	-4.35
27	26,310,001	26,453,084		143,083	-4.85
29	90,001	300,385		210,384	-5.95
29	17,780,001	17,880,857		100,856	-4.06
29	34,280,001	34,380,825		100,824	-4.07
29	45,320,001	45,462,334		142,333	-4.69

*The candidate sweep regions were cross-referenced with the ones obtained previously on tropical-adapted cattle and commercial breeds.

Table S3.14: Genes within EASZ candidate *Hp* sweep regions.

The file is an electronic version

Table S3.15: Genes within the genome-wide SNP analyses and *Hp* overlapping candidate regions for EASZ, East African zebu (EASZ and Uganda), and East and West African zebu (EASZ, Uganda and Nigeria)**.**

The file is an electronic version

Table S3.16: Annotation of variants (SNPs and indels) within the 12 candidate genes mapped on EASZ genome-wide SNP analyses and *Hp* analysis overlapping candidate regions.

The file is an electronic version

Table S3.17: Bovine QTL spanning the candidate *Hp* sweep regions, EASZ genome-wide SNP analyses and *Hp* analysis overlapping candidate regions, East African zebu-sharing, and East and West African zebu-sharing candidate regions.

The file is an electronic version

Table S3.18: Trypanotolerance QTL spanning the candidate *Hp* sweep regions.

BTA	start	stop	QTL_ID
7*	38,332,974	59,135,155	PCVF minus PCVM QTL (10516)
7*	38,332,974	59,135,155	Body weight (mean) QTL (10517)
7*	38,332,974	59,135,155	Percentage decrease in body weight up to day 150 after challenge QTL (10518)
7*	38,332,974	59,135,155	Parasites natural logarithm of mean number QTL (10519)
7*	38,332,974	59,135,155	Parasite detection rate QTL (10520)
13*	29,549,346	71,827,292	Percentage decrease in PCV up to day 150 after challenge QTL (10524)
13*	29,549,346	71,827,292	Percentage decrease in PCV up to day 100 after challenge QTL (10525)
13*	29,549,346	71,827,292	Parasite detection rate QTL (10526)
23	12,487,278	18,536,147	Parasite detection rate QTL (10543)
26	10,723,763	25,449,045	Percentage decrease in PCV up to day 150 after challenge QTL (10548)
26	10,723,763	25,449,045	Percentage decrease in PCV up to day 100 after challenge QTL (10549)
26	10,723,763	25,449,045	BWF scaled by BWI QTL (10550)
26	10,723,763	25,449,045	Body weight (mean) QTL (10551)
26	10,723,763	25,449,045	Percentage decrease in body weight up to day 150 after challenge QTL (10552)
27	21,255,822	32,724,672	PCVI minus PCVM QTL (10553)
27	21,255,822	32,724,672	PCV variance QTL (10554)
27	21,255,822	32,724,672	Percentage decrease in PCV up to day 150 after challenge QTL (10555)
27	21,255,822	32,724,672	Percentage decrease in PCV up to day 100 after challenge QTL (10556)
29	1,568,804	25,983,377	BWF scaled by BWI QTL (10559)
29	1,568,804	25,983,377	Parasites natural logarithm of mean number QTL (10560)

*span EASZ genome-wide SNP analyses candidate regions. The BTA 13 QTL also cover candidate region identified in zebu cattle from Uganda

Table S3.19: The median and standard deviation of the average depth of coverage and the normalised depth of coverage (reads/bp) for the ten EASZ autosomal exome sequences.

	EASZ samples									
Average coverage	**289**	**524**	**915**	**923**	**1148**	**1401**	**1693**	**2025**	**2063**	**2183**
median	40.67	41.67	49.16	47	36.33	48.26	51.1	69.31	39.46	55.09
Standard deviation	56.1	57.6	68.5	63.7	52.37	69.21	71.4	93.99	55	75.47
Normalised coverage	**289**	**524**	**915**	**923**	**1148**	**1401**	**1693**	**2025**	**2063**	**2183**
median	0.88	0.88	0.87	0.88	0.87	0.89	0.89	0.92	0.9	0.9
Standard deviation	1.21	1.22	1.2	1.2	1.25	1.28	1.25	1.25	1.26	1.24

Table S3.20: Signals of multiple copies (represented as standardised depth of coverage (SDOC)), identified on the ten EASZ exome sequences, within EASZ genome-wide SNP analyses and *Hp* analysis overlapping candidate regions. SDOC value = 3 set as a threshold.

Exome regions														
BTA	**start (BP)**	**stop (BP)**	**Genes**	**Candidate regions**	**289**	**524**	**915**	**923**	**1148**	**1401**	**1693**	**2025**	**2063**	**2183**
9	**74,028,868**	**74,029,044**	***HBS1L***	**BTA 9:73890001-74081863**	6.671368	5.621976	5.526219	4.779615	5.654543	6.145115	5.494708	5.996021	5.451463	5.873202
5	**48,775,936**	**48,776,118**	***Man1***	**BTA 5:48610001-49021113**	5.915964	5.431604	5.702516	6.507744	6.209973	5.13598	5.984377	5.168023	5.88899	5.338395
11	**39,664,144**	**39,664,359**	**_**	**BTA 11:39550001-39683044**	4.738119	4.398525	4.332418	4.623824	4.385016	4.778073	4.405201	4.427803	4.478111	4.531415
5	**48,773,871**	**48,774,151**	***Man1***	**BTA 5:48610001-49021113**	4.431358	4.212838	4.404367	4.130095	4.585497	4.692389	4.328629	4.763599	4.173041	4.227306
13	**48,777,935**	**48,778,137**	**_**	**BTA 13:48650001-49056444**	4.17825	3.818036	3.834466	2.756535	4.244678	3.837716	3.851559	3.955284	4.431415	4.257783
13	**48,097,307**	**48,097,529**	***GPCPD1***	**BTA 13:47980001-48164495**	4.137967	4.306549	4.01193	3.201785	4.338236	3.56217	4.12747	3.862079	3.799471	3.895503
12	**29,273,838**	**29,274,045**	***RXFP2***	**BTA 12:29110001-29438417**	4.071303	3.868016	4.335337	4.899948	4.359239	4.55252	4.114592	3.723334	4.055665	4.019532
2	**125,417,834**	**125,417,954**	***TAF12***	**BTA 2:125300001-125620820**	3.958295	4.643736	3.902474	4.228308	3.778988	3.150366	4.005963	5.819505	4.137247	4.415602
22	**45,223,999**	**45,224,176**	***ERC2***	**BTA 22:45220001-45370457**	3.852952	3.877213	3.907144	4.230661	3.612492	3.893635	4.046559	4.046575	4.123256	3.520635
19	**27,050,600**	**27,050,799**	***INCA1***	**BTA 19:26890001-27154002**	3.829424	3.388353	3.790683	4.962076	3.728199	2.823669	3.522733	3.950284	3.479866	3.360431
19	**41,305,767**	**41,305,990**	***TOP2A***	**BTA 19:40960001-41450870**	3.768107	3.665669	3.850665	4.725331	3.75493	3.421001	4.049358	3.797282	3.585796	3.802217
9	**74,011,926**	**74,012,131**	***HBS1L***	**BTA 9:73890001-74081863**	3.701265	3.743935	3.398538	3.201	3.596072	3.526769	3.55353	3.617361	2.922782	3.263965
2	**125,705,828**	**125,706,057**	***SESN2***	**BTA 2:125640001-126083262**	3.620877	3.27295	3.328341	3.844087	3.301268	2.393949	3.008847	3.056318	2.701293	3.26211
11	**39,635,629**	**39,635,758**	**_**	**BTA 11:39550001-39683044**	3.501631	3.787493	2.930796	3.038464	3.160549	3.683977	2.721177	3.521495	3.929203	2.9368
7	**33,171,236**	**33,171,457**	**_**	**BTA 7:33100001-33293306**	3.499135	3.40241	3.298715	2.541283	3.686193	3.528647	3.178649	3.286459	3.825636	2.947931
7	**52,347,974**	**52,348,181**	***ECSCR***	**BTA 7:51360001-53362761**	3.443701	2.971686	3.390658	3.947163	3.151766	2.794048	2.854863	3.538625	3.180247	3.08468
13	**49,655,060**	**49,655,283**	**_**	**BTA 13:49590001-49844283**	3.43158	3.272255	3.253473	2.959078	3.151002	3.472006	3.196147	3.205489	3.477141	3.184989
19	**27,083,610**	**27,083,739**	***PFN1***	**BTA 19:26890001-27154002**	3.386306	3.370305	2.968303	3.847696	3.221839	2.654902	3.286578	3.847928	2.989465	3.466306
2	**126,036,792**	**126,036,912**	***RPA2***	**BTA 2:125640001-126083262**	3.385058	2.890643	3.22122	3.995171	3.309096	3.169873	3.17291	3.083556	3.082676	3.431059
7	**52,239,216**	**52,239,336**	***MATR3***	**BTA 7:51360001-53362761**	3.379354	2.766215	3.005956	4.051337	3.749965	3.620544	3.240103	2.974071	3.322516	3.744178

12	**20,891,040**	**20,891,245**	***INTS6***	**BTA 12:20870001-21021506**	3.305739	3.778643	3.40423	3.746659	3.230622	3.627914	3.367769	3.846332	3.25438	3.236403
1	**55,150,444**	**55,150,620**	**_**	**BTA 1:55150001-55253859**	3.192731	4.169106	2.641686	2.070147	4.32869	4.423777	3.120415	3.162185	2.916423	2.979998
12	**20,889,198**	**20,889,406**	***INTS6***	**BTA 12:20870001-21021506**	3.151378	2.701311	2.931818	2.548971	2.986225	2.65548	2.750994	3.010353	2.801045	2.478716
5	**48,781,810**	**48,781,936**	***Man1***	**BTA 5:48610001-49021113**	3.1512	3.753133	3.786597	2.713704	3.956557	3.643952	3.74181	5.083967	4.458124	3.845017
13	**48,129,311**	**48,129,526**	***GPCPD1***	**BTA 13:47980001-48164495**	3.06279	2.565257	2.738737	2.09164	3.148902	2.569074	3.215465	2.740844	2.690391	3.722844
7	**53,252,865**	**53,253,052**	***SLC4A9***	**BTA 7:51360001-53362761**	2.977411	2.854373	2.937655	3.27615	2.55643	2.473853	2.761213	2.968325	2.842109	3.086005
13	**48,093,369**	**48,093,600**	***GPCPD1***	**BTA 13:47980001-48164495**	2.957804	3.110517	3.076446	2.43507	3.339837	2.738419	3.308976	3.221875	3.069775	2.742144
13	**48,107,101**	**48,107,276**	***GPCPD1***	**BTA 13:47980001-48164495**	2.935167	2.589379	2.675836	1.836539	2.707842	2.627449	2.998488	2.768188	2.607901	2.653495
7	**52,449,674**	**52,449,803**	***UBE2D2***	**BTA 7:51360001-53362761**	2.537858	3.358852	3.187653	3.242262	3.595117	3.460447	3.308556	2.974709	3.67083	2.868425
11	**39,477,126**	**39,477,254**	**_**	**BTA 11:39240001-39530799**	2.464064	2.724913	2.594109	1.832617	2.140001	2.20004	2.59379	2.649447	3.139365	2.063565
7	**33,151,917**	**33,152,132**	***SRFBP1***	**BTA 7:33100001-33293306**	2.435545	3.182015	2.942617	2.533596	3.029185	2.78509	3.313315	3.02078	3.532377	2.906058

Chapter Four

Table S4.1: Candidate regions on BTA X of EASZ, African zebu cattle from Uganda and Nigeria defined by the *Rsb* and *iHS* analyses, and the sharing candidate regions for East African zebu cattle.

EASZ			**Uganda**			**Nigeria**			**East African zebu-sharing**		
BTA	**start**	**stop**	**BTA**	**start**	**stop**	**BTA**	**start**	**stop**	**BTA**	**start**	**stop**
Rsb			***Rsb***			***Rsb***					
X	52,724,365	53,286,404	X	64,636,360	64,803,846	X	26,640,020	26,726,577	X	64,636,360	64,803,846
X	57,753,489	57,862,192	X	66,692,636	67,237,062	X*	68,409,400	68,493,120	X	86,752,071	87,096,922
X	58,699,877	59,561,015				X**	100,112,923	100,285,092			
X	64,636,360	64,803,846				X	101,717,122	102,147,368			
X	68,399,979	68,541,416				X	103,073,067	103,499,632			
X	90,841,068	91,235,638									
iHS			***iHS***			***iHS***					
X	52,724,365	52,831,886	X	15,898,277	16,419,921						
X	86,750,029	87,174,478	X	26,637,789	26,730,011						
			X	86,752,071	87,096,922						

*Overlap with EASZ candidate region identified by *Rsb* analysis

** Overlap with EASZ candidate region identified by *Hp* analyses.

Table S4.2: The coordinates of the BTA X 100 kb windows, number of uncovered bases, number of SNPs within window, *Hp* and ZHp values.

The file is an electronic version

Table S4.3: Gene desert BTA X candidate genome regions identified by EASZ *Hp* analyses.

BTA	Start (bp)	End (bp)
X	27,900,001	28,042,935
X	58,030,001	58,239,841
X	142,500,001	142,633,019
X	144,260,001	144,424,871

Table S4.4: Genes mapped on the EASZ (*Rsb*, *iHS* and *Hp* analyses), East African zebu-sharing, and EASZ-Nigerian zebu cattle-sharing candidate regions.

The file is an electronic version

Table S4.5: Functional term clusters of the genes mapped within (A) EASZ autosomal and BTA X candidate regions combined (B) EASZ BTA X and autosomal SNPs and *Hp* analyses overlapping candidate regions.

The file is an electronic version

Table S4.6: Annotation of variants (SNPs and indels) within the genes in the EASZ BTA X candidate regions.

The file is an electronic version

Table S4.7: The median and standard deviation of the average depth of coverage and the normalised depth of coverage (reads/bp) for the ten EASZ BTA X exome sequences

	Female					**Male**				
Average coverage	**289**	**524**	**915**	**923**	**1148**	**1401**	**1693**	**2025**	**2063**	**2183**
median	50.5	52.5	59.8	54.9	44.45	32.25	32.85	43.63	25.9	35.19
Standard deviation	79	80.4	97	89.1	69.29	51.58	53.57	68.17	38.3	56.35
Normalized coverage	**289**	**524**	**915**	**923**	**1148**	**1401**	**1693**	**2025**	**2063**	**2183**
median	1.1	1.1	1.1	1	1.1	0.59	0.57	0.58	0.59	0.58
Standard deviation	1.7	1.7	1.7	1.7	1.7	0.95	0.94	0.91	0.87	0.92

Table S4.8: Signals of multiple copies (represented as standardised depth of coverage (SDOC)), identified on the ten EASZ exome sequences, within EASZ BTA X candidate regions. SDOC value = 3 set as a threshold.

Exome regions					**Female**					**Male**				
BTA	**start**	**stop**	**Genes**	**Candidate regions**	**289**	**524**	**915**	**923**	**1148**	**1401**	**1693**	**2025**	**2063**	**2183**
X	**89,907,781**	**89,908,637**	**uncharacterised gene**	**BTA X:89,870,001-89,983,990**	67.6275	60.2594	67.7323	68.6099	57.9021	32.6209	36.1586	34.3549	27.3356	36.7127
X	**97,403,427**	**97,403,604**	***FGD1***	**BTA X:97,300,001-97,720,659**	3.4653	3.3669	4.1077	4.1824	3.2495	1.4498	1.3097	1.4163	1.5885	1.3635

Chapter Five

Table S5.1: List of all domestic cattle included in the study and the corresponding references (if downloaded from NCBI database).

Sample ID	Genebank ID	Breed	Source	Reference	Haplogroup	used in selection (ω ratio) analyses
Angus-1	AY676873.1	Angus mix	Scotland	unpublished	unknow	*
Angus-2	AY676872.1	Angus mix	Scotland	unpublished	unknow	*
Angus-3	AY676871.1	Angus mix	Scotland	unpublished	unknow	*
Iranian-1	EU177859.1	Iranian	Iran	Achilli *et al*., 2008	T2	*
Iranian-2	EU177838.1	Iranian	Iran	Achilli *et al*., 2008	T3	*
Iranian-3	EU177860.1	Iranian	Iran	Achilli *et al*., 2008	T2	*
Chianina-1	HQ184030.1	Chianina	Italy	Bonfiglio *et al*., 2010	Q2	*
Chianina-2	HQ184031.1	Chianina	Italy	Bonfiglio *et al*., 2010	Q2	*
Chianina-3	HQ184032.1	Chianina	Italy	Bonfiglio *et al*., 2010	Q2	*
Chianina-4	JN817313	Chianina	Italy	Bonfiglio *et al*., 2012	T1a	
Chianina-5	JN817316	Chianina	Italy	Bonfiglio *et al*., 2012	T1a	
Chianina-6	EU177841	Chianina	Italy	Achilli *et al*., 2008	T1e	*
Chianina-7	EU177816	Chianina	Italy	Achilli *et al*., 2009	T3	
K.beef-1	DQ124396.1	Beef cattle	Korea	unpublished	T2	*
K.beef-2	DQ124397.1	Beef cattle	Korea	unpublished	T3	*
K.beef-3	DQ124392	Beef cattle	Korea	unpublished	T4	
K.beef-4	DQ124400	Beef cattle	Korea	unpublished	T4	
K.beef-5	DQ124401	Beef cattle	Korea	unpublished	T4	
Holstein-1	DQ124412.1	Holstein-Friesian	England	unpublished	T4	*
Holstein-2	DQ124413.1	Holstein-Friesian	England	unpublished	T3	*

Holstein-3	DQ124406.1	Holstein-Friesian	England	unpublished	T3	*
Holstein-4	DQ124407.1	Holstein-Friesian	England	unpublished	T3	*
Holstein-5	DQ124408.1	Holstein-Friesian	England	unpublished	T3	*
Holstein-6	DQ124409.1	Holstein-Friesian	England	unpublished	T3	*
Japan-1	AB074968.1	Japanese Black	Japan	unpublished	unknow	*
Japan-2	AB074967.1	Japanese Black	Japan	unpublished	unknow	*
Japan-3	AB074966.1	Japanese Black	Japan	unpublished	unknow	*
Korean-1	DQ124386.1	Korean cattle	Korea	unpublished	T3	*
Korean-2	DQ124372.1	Korean cattle	Korea	unpublished	T3	*
Nelore-1	AY126697.1	Nellore	Asia	unpublished	unknow	*
Nelore-2	NC_005971.1	Nellore	Asia	unpublished	I1	*
Romagnola-1	HQ184041.1	Romagnola	Italy	Bonfiglio *et al.*, 2010	R1	*
Romagnola-2	HQ184033.1	Romagnola	Italy	Bonfiglio *et al.*, 2010	Q2	*
Romagnola-3	HQ184034.1	Romagnola	Italy	Bonfiglio *et al.*, 2010	Q1a	*
Romagnola-4	JN817347	Romagnola	Italy	Bonfiglio *et al.*, 2012	T1a	*
EASZ-1 (S352)	not submitted	EASZ	Kenya (Magombe East)	This study	unknow	*
EASZ-2 (R90)	not submitted	EASZ	Kenya (Bukati)	This study	unknow	*
EASZ-3 (S194)	not submitted	EASZ	Kenya (Kokare)	This study	unknow	*
EASZ-4 (S378)	not submitted	EASZ	Kenya (BumalaA)	This study	unknow	*
EASZ-5 (S479)	not submitted	EASZ	Kenya (Simur East)	This study	unknow	*
EASZ-6 (S666)	not submitted	EASZ	Kenya (Luanda)	This study	unknow	*
EASZ-7 (R341)	not submitted	EASZ	Kenya (Kidera)	This study	unknow	*
EASZ-8 (R354)	not submitted	EASZ	Kenya (Namboboto)	This study	unknow	*
EASZ-9 (R364)	not submitted	EASZ	Kenya (Otimong)	This study	unknow	*

EASZ-10 (R944)	not submitted	EASZ	Kenya (Ojwando B)	This study	unknow	*
EASZ-11 (R500)	not submitted	EASZ	Kenya (Bujwanda)	This study	unknow	*
EASZ-12 (S70)	not submitted	EASZ	Kenya (Mabusi)	This study	unknow	*
EASZ-13 (S294)	not submitted	EASZ	Kenya (Kamunuoit)	This study	unknow	*
N'Dama	ERR022967	N'Dama	Gambia	Roslin Institute	unknow	*
Arsi-1	JN817302	Arsi	Ethiopia	Bonfiglio *et al.*, 2012	T1b1	
Arsi-2	JN817304	Arsi	Ethiopia	Bonfiglio *et al.*, 2012	T1d1	
Arsi-3	JN817303	Arsi	Ethiopia	Bonfiglio *et al.*, 2012	T1a	*
Sheko-1	JN817349	Sheko	Ethiopia	Bonfiglio *et al.*, 2012	T1b1	*
Sheko-2	JN817348	Sheko	Ethiopia	Bonfiglio *et al.*, 2012	T1b	
Boran-1	JN817305	Boran	Ethiopia	Bonfiglio *et al.*, 2012	T1b1	
Boran-2	JN817299	Boran	Ethiopia	Bonfiglio *et al.*, 2012	T1d	
Domiaty-1	JN817324	Domiaty	Egypt	Bonfiglio *et al.*, 2012	T1b1	*
Domiaty-2	JN817323	Domiaty	Egypt	Bonfiglio *et al.*, 2012	T1c1	*
Domiaty-3	JN817322	Domiaty	Egypt	Bonfiglio *et al.*, 2012	T1c1a1	
Domiaty-4	JN817321	Domiaty	Egypt	Bonfiglio *et al.*, 2012	T1d1	
Menofi-1	JN817327	Menofi	Egypt	Bonfiglio *et al.*, 2012	T1b	*
Menofi-2	JN817326	Menofi	Egypt	Bonfiglio *et al.*, 2012	T1c	*
Menofi-3	JN817235	Menofi	Egypt	Bonfiglio *et al.*, 2012	T1c	*
Menofi-4	JN817238	Menofi	Egypt	Bonfiglio *et al.*, 2012	T1c	
Menofi-5	JN817329	Menofi	Egypt	Bonfiglio *et al.*, 2012	T1f	
Horro	JN817330	Horro	Ethiopia	Bonfiglio *et al.*, 2012	T1d1	
Abigar	JN817298	Abigar	Ethiopia	Bonfiglio *et al.*, 2012	T1d	*
Podolian	JN817343	Italian_Podolian	Italy	Bonfiglio *et al.*, 2012	T1f	*

Clavana	JN817306	clavana	Italy	Bonfiglio *et al*., 2012	T1e	*
Valdostana-1	EU177817	Valdostana	Italy	Achilli *et al*., 2008	T3	
Valdostana-2	EU177862	Valdostana	Italy	Achilli *et al*., 2008	T5	
Greek-1	EU177849	Greek	Greece	Achilli *et al*., 2008	T2	
Cabannina-1	EU177850	Cabannina	Italy	Achilli *et al*., 2008	T2	
Cabannina-2	EU177851	Cabannina	Italy	Achilli *et al*., 2008	T2	
Piedmontese	EU177863	Piedmontese	Italy	Achilli *et al*., 2008	T5	
Iraqi	EU177864	Iraqi	Iraq	Achilli *et al*., 2008	T5	

* Samples used in signatures of selection (ω ratio) analyses.

Table S5.2: Likelihood values and parameter estimates of the different site models implemented in CODEML package for the 13 mtDNA protein-coding genes. (A) taurine and zebu cattle. (B) taurine. (C) African taurine. Sites considered as positively selected using Bayes Empirical Bayes (BEB) approach. * posterior probability > 95%.

The file is an electronic version

Table S5.3: Branch-site tests for positive selection on the 13 mtDNA protein-coding genes. African taurine cattle set as foreground lineage. Negative 2ΔL values indicate higher log likelihood for the null model. Sites considered as positively selected using Bayes Empirical Bayes (BEB) approach.

The file is an electronic version

Table S5.4: mtDNA protein-coding gene intra-population nucleotide diversity and inter-population nucleotide divergence for the taurine and zebu cattle used in the selection analyses.

	ND1	**ND2**	**Cox1**	**Cox2**	**ATP8**	**ATP6**	**Cox3**	**ND3**	**ND4L**	**ND4**	**ND5**	**ND6**	**CYTB**	**mean**
Taurine diversity	0.00108	0.00055	0.00088	0.00123	0.00183	0.0014	0.00097	0.00044	0.00039	0.0019	0.00155	0.00139	0.00167	0.0011754
European-Asian taurine	0.00131	0.00045	0.00126	0.00089	0.00224	0.00184	0.0012	0.00087	0.00034	0.0016	0.00182	0.0017	0.00182	0.0013338
European-African taurine	0.00134	0.00058	0.00081	0.00156	0.00222	0.00166	0.00092	0.00029	0.00049	0.00139	0.00193	0.00167	0.00196	0.0012938
African- Asian taurine	0.00045	0.00048	0.00085	0.0012	0.00097	0.00071	0.00098	0.00058	0.00015	0.00067	0.00064	0.00072	0.001	0.0007231
mean taurine divergence	0.001033	0.000503	0.000973	0.001217	0.00181	0.001403	0.001033	0.00058	0.000327	0.00122	0.001463	0.001363	0.001593	0.0011169
taurine-zebu	0.01927	0.00987	0.01053	0.00792	0.03042	0.01209	0.01188	0.02323	0.01353	0.01123	0.02031	0.01734	0.01623	0.0156808
zebu diversity	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table S5.5: Annotation of variants (SNPs and indels) within the mitochondrial-related nuclear genes mapped on the specified candidate sweep regions in chapter 3 and 4.

The file is an electronic version

References

CHAN, E. K., NAGARAJ, S. H. & REVERTER, A. 2010. The evolution of tropical adaptation: comparing taurine and zebu cattle. *Anim. Genet.,* 41**,** 467-77.

FLORI, L., THEVENON, S., DAYO, G. K., SENOU, M., SYLLA, S., BERTHIER, D., MOAZAMI-GOUDARZI, K. & GAUTIER, M. 2014. Adaptive admixture in the West African bovine hybrid zone: insight from the Borgou population. *Mol. Ecol.,* 23**,** 3241-57.

GAUTIER, M., FLORI, L., RIEBLER, A., JAFFREZIC, F., LALOE, D., GUT, I., MOAZAMI-GOUDARZI, K. & FOULLEY, J. L. 2009. A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. *BMC Genomics,* 10**,** 550.

GAUTIER, M. & NAVES, M. 2011. Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Mol. Ecol.,* 20**,** 3128-43.

HANOTTE, O., RONIN, Y., AGABA, M., NILSSON, P., GELHAUS, A., HORSTMANN, R., SUGIMOTO, Y., KEMP, S., GIBSON, J., KOROL, A., SOLLER, M. & TEALE, A. 2003. Mapping of quantitative trait loci controlling trypanotolerance in a cross of tolerant West African N'Dama and susceptible East African Boran cattle. *PNAS,* 100**,** 7443-8.

KEMPER, K. E., SAXTON, S. J., BOLORMAA, S., HAYES, B. J. & GODDARD, M. E. 2014. Selection for complex traits leaves little or no classic signatures of selection. *BMC Genomics,* 15**,** 246.

LARKIN, D. M., DAETWYLER, H. D., HERNANDEZ, A. G., WRIGHT, C. L., HETRICK, L. A., BOUCEK, L., BACHMAN, S. L., BAND, M. R., AKRAIKO, T. V., COHEN-ZINDER, M., THIMMAPURAM, J., MACLEOD, I. M., HARKINS, T. T., MCCAGUE, J. E., GODDARD, M. E., HAYES, B. J. & LEWIN, H. A. 2012. Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. *PNAS,* 109**,** 7693-8.

LIAO, X., PENG, F., FORNI, S., MCLAREN, D., PLASTOW, G. & STOTHARD, P. 2013. Whole genome sequencing of Gir cattle for identifying polymorphisms and loci under selection. *Genome,* 56**,** 592-8.

PEREZ O'BRIEN, A. M., UTSUNOMIYA, Y. T., MESZAROS, G., BICKHART, D. M., LIU, G. E., VAN TASSELL, C. P., SONSTEGARD, T. S., DA SILVA, M. V., GARCIA, J. F. & SOLKNER, J. 2014. Assessing signatures of selection through variation in linkage disequilibrium between taurine and indicine cattle. *Genet. Sel. Evol.,* 46**,** 19.

QANBARI, S., PAUSCH, H., JANSEN, S., SOMEL, M., STROM, T. M., FRIES, R., NIELSEN, R. & SIMIANER, H. 2014. Classic selective sweeps revealed by massive sequencing in cattle. *PLoS Genet.,* 10**,** e1004148.

QANBARI, S., PAUSCH, H., JANSEN, S., SOMEL, M., STROM, T. M., FRIES, R., NIELSEN, R. & SIMIANER, H. 2014. Classic selective sweeps revealed by massive sequencing in cattle. *PLoS Genet.,* 10**,** e1004148.

XU, L., BICKHART, D. M., COLE, J. B., SCHROEDER, S. G., SONG, J., VAN TASSELL, C. P., SONSTEGARD, T. S. & LIU, G. E. 2015. Genomic signatures reveal new evidences for selection of important traits in domestic cattle. *Mol. Biol. Evol.,* 32**,** 711-25.

Supplementary Figures

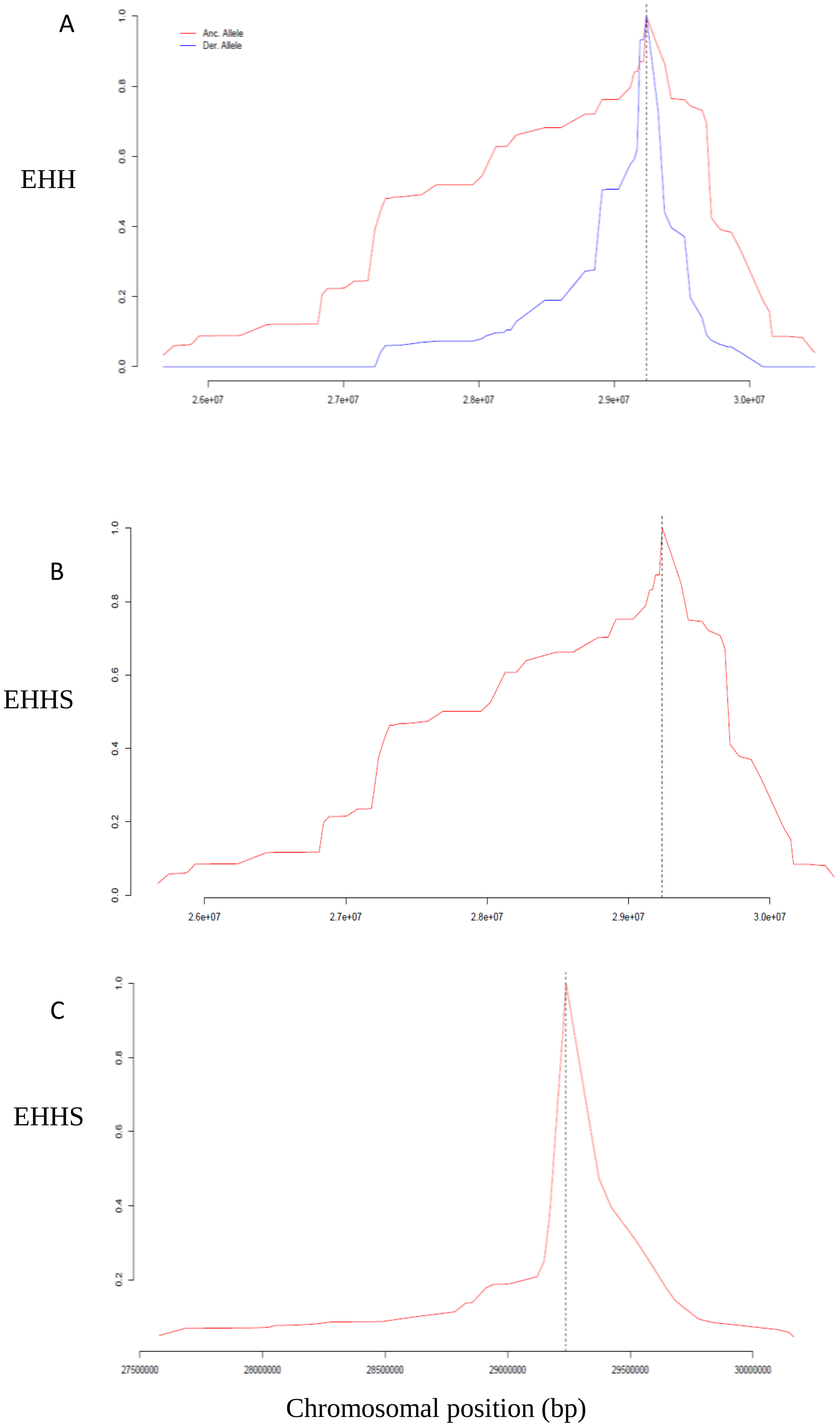

Figure S2.1: Example of an EHH decay plot with calculation for the ancestral allele in red and for the derived allele in blue; and example of an EHHS decay plot in EASZ (B) and in the combined reference populations (C). Both calculated for a selected core SNP "Hapmap23766.BTA.152495" on chromosome 12 at chromosome position 29,217,254 bp (indicated by the dashed line). In both plots, EHH and EHHS are calculated until they decay below the arbitrary value of 0.05.

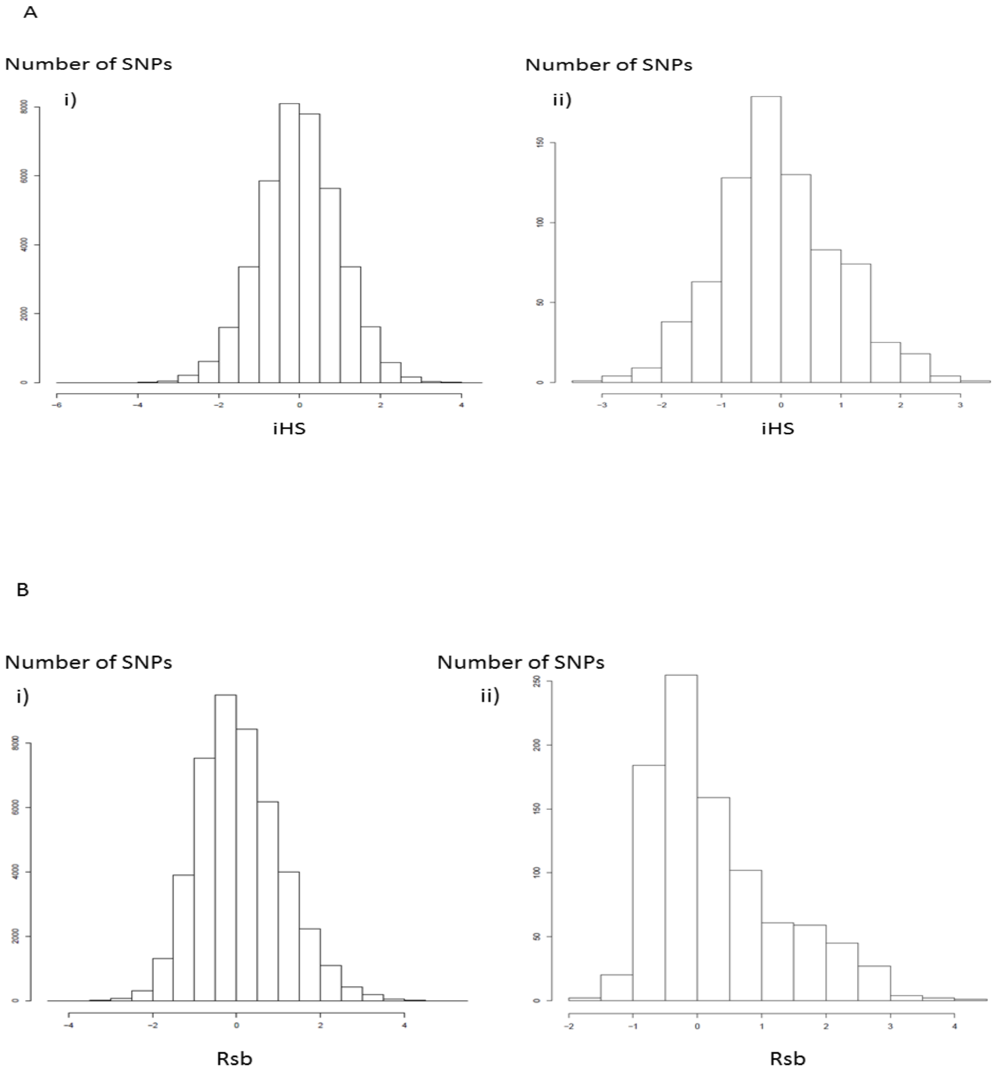

Figure S2.2: Histogram plots for the genome-wide (i) autosomal and (ii) BTA X (A) *iHS* values in EASZ, and (B) *Rsb* values between EASZ and the combined reference populations (Holstein-Friesian, Jersey, N'Dama and Nellore).

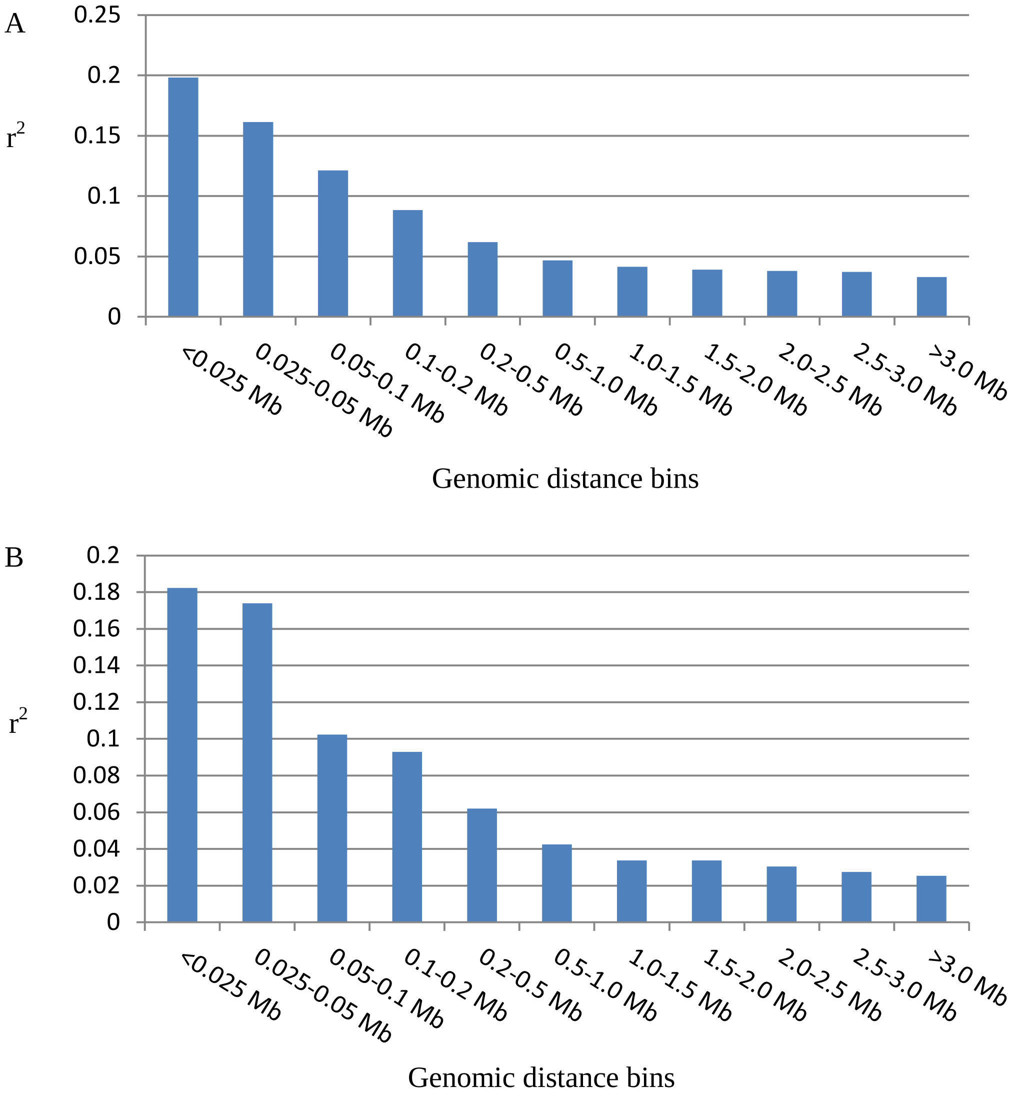

Figure S2.3: Mean r^2 values over increasing distances across (A) EASZ autosomes and (B) EASZ BTA X. In (A) values averaged across all the autosomes for each bin size.

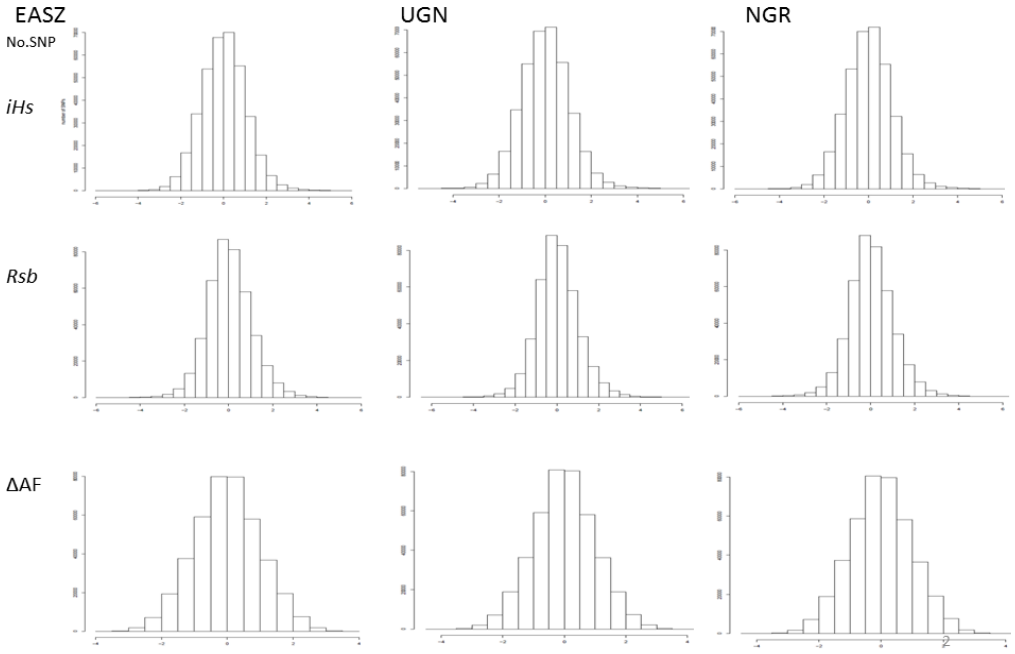

Figure S3.1: Histogram plots of *Rsb*, *iHS* and Δ*AF* standardized values performed on the autosomal SNPs indicating normal distribution. EASZ: East African shorthorn zebu. UGN: zebu cattle from Uganda. NGR: zebu cattle from Nigeria.

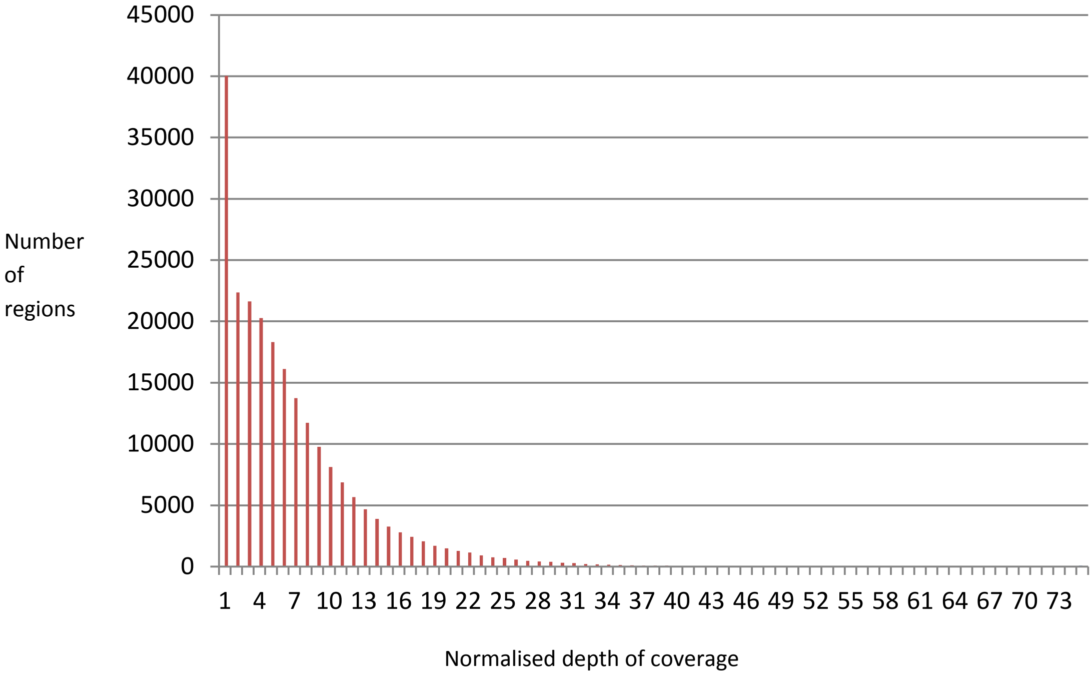

Figure S3.2: Distribution of the normalized depth of coverage values for the captured target exome regions. This distribution is for sample 289 which resembles the distribution obtained in the other samples.

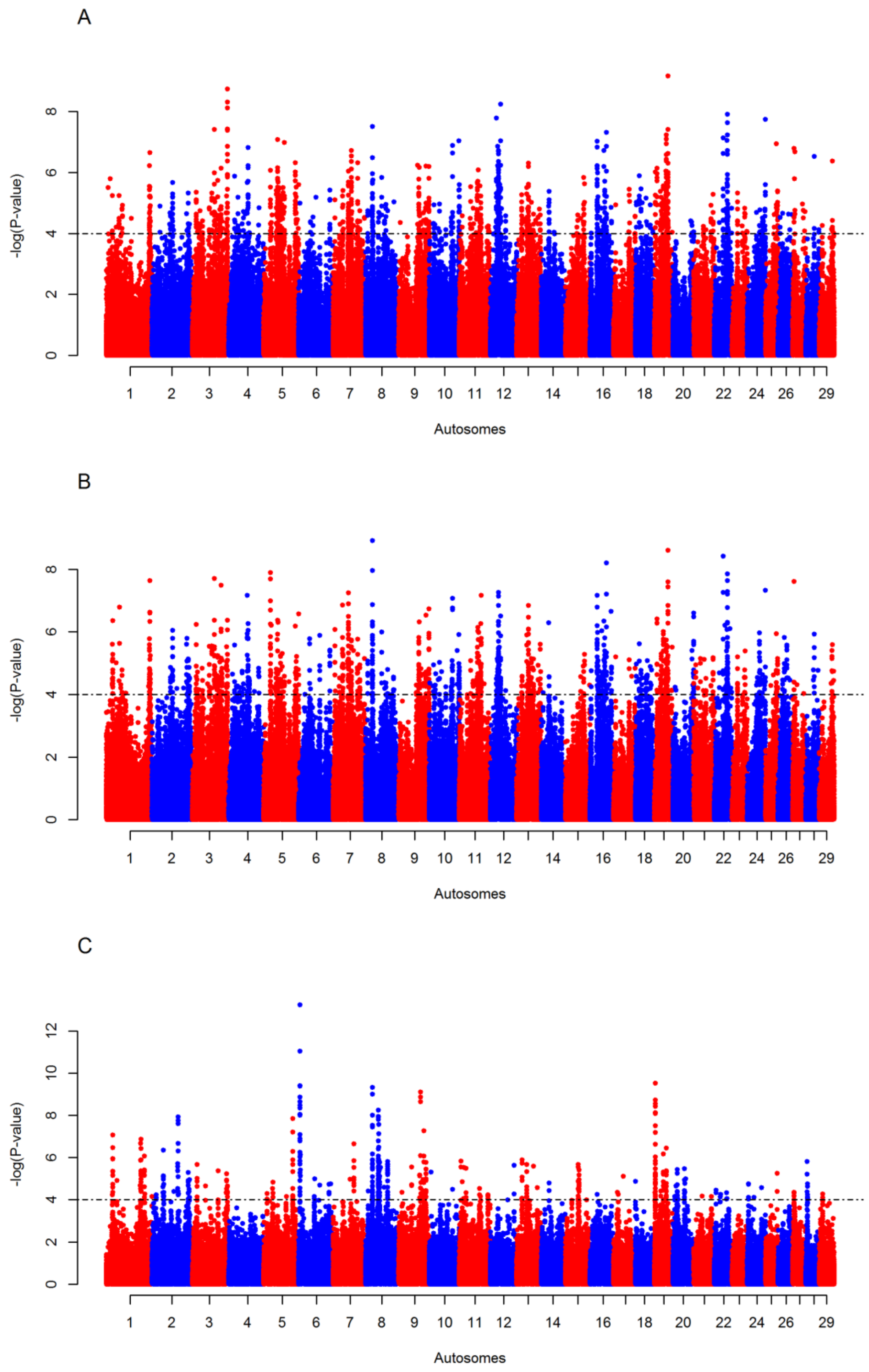

Figure S3.3: Manhattan plots of the genome-wide autosomal *meta-SS* analyses between EASZ and (A) European taurine (Holstein-Friesian and Jersey), (B) African taurine (N'Dama and Muturu) and (C) Asian zebu (Nellore and Gir). Threshold set as – $\log_{10}$ *P*-value = 4.

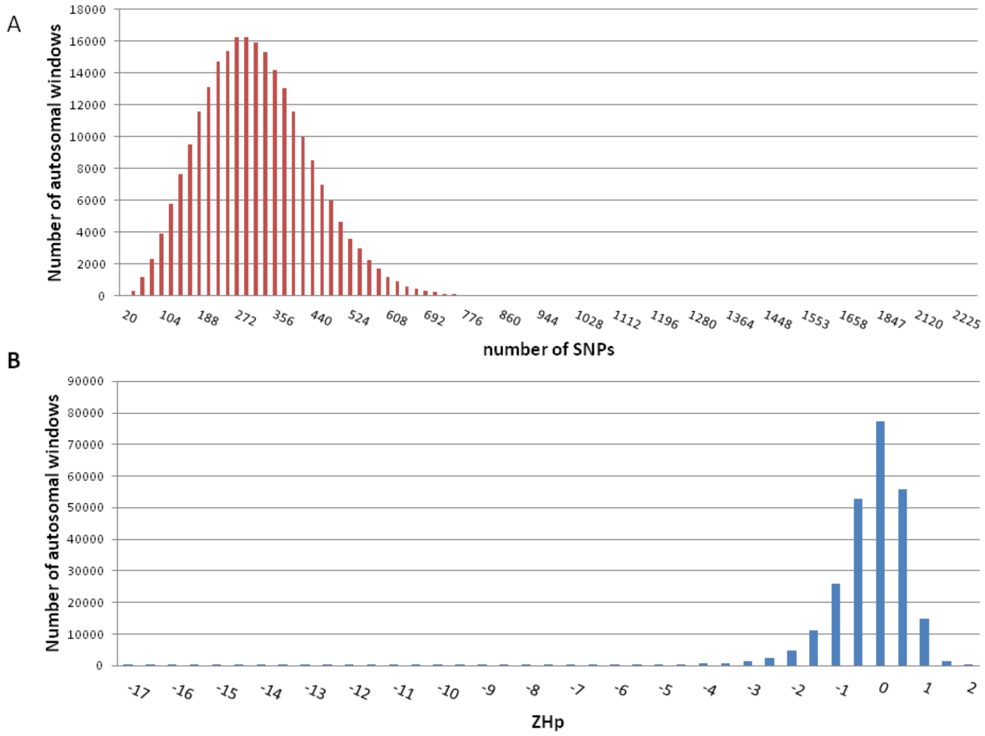

Figure S3.4: Distribution of the (A) SNPs and (B) ZHp values in the 100 kb autosomal windows on EASZ.

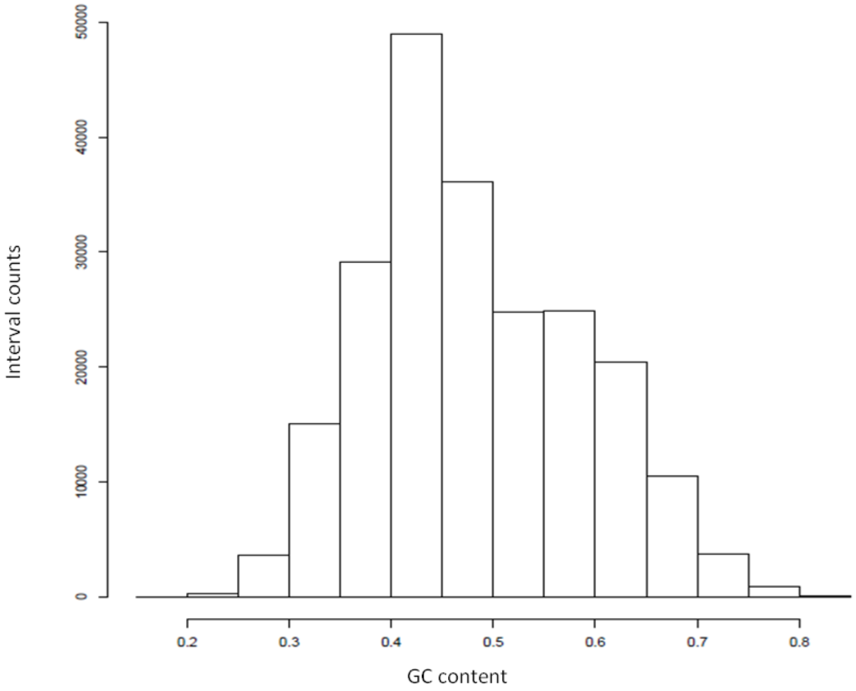

Figure S3.5: GC content histogram plot of the captured autosomal exome by Agilent SureSelect XT target enrichment system.

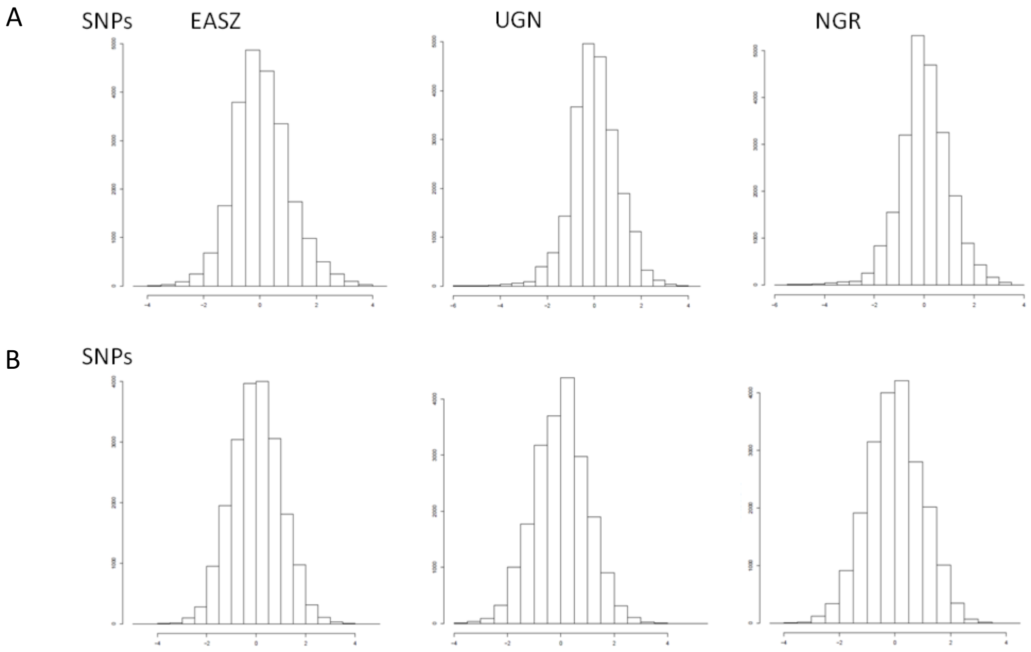

Figure S4.1: Histogram plots of (A) *Rsb* and (B) *iHS* standardized values performed on the BTA X SNPs indicating normal distribution.

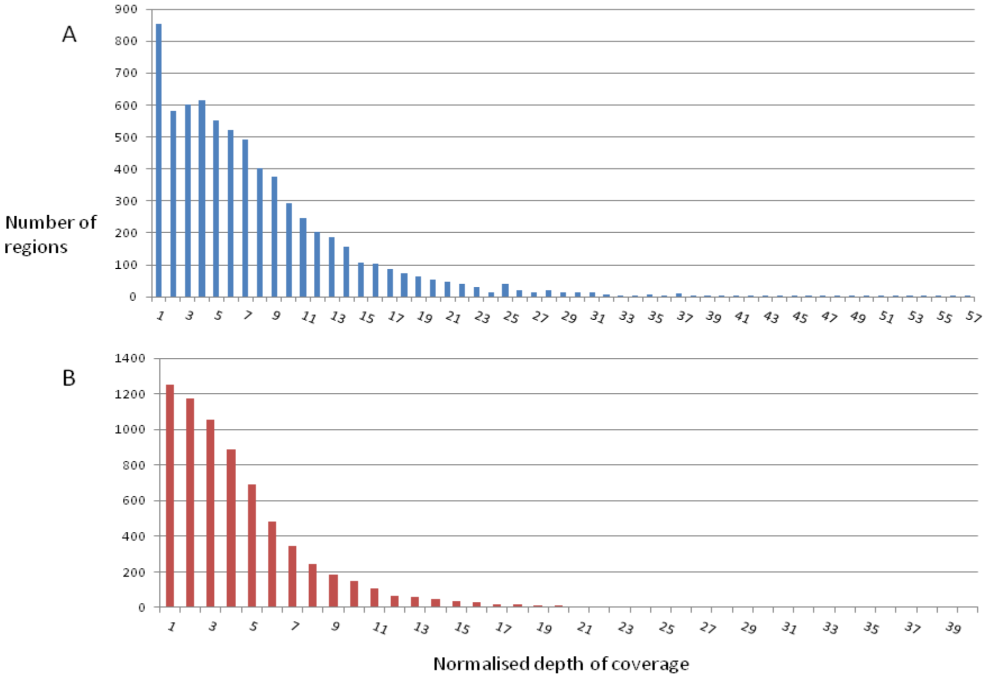

Figure S4.2: Distribution of the normalised depth of coverage values for the captured target exome on BTA X. (A) Sample 289 (female). (B) Sample 2063 (male).

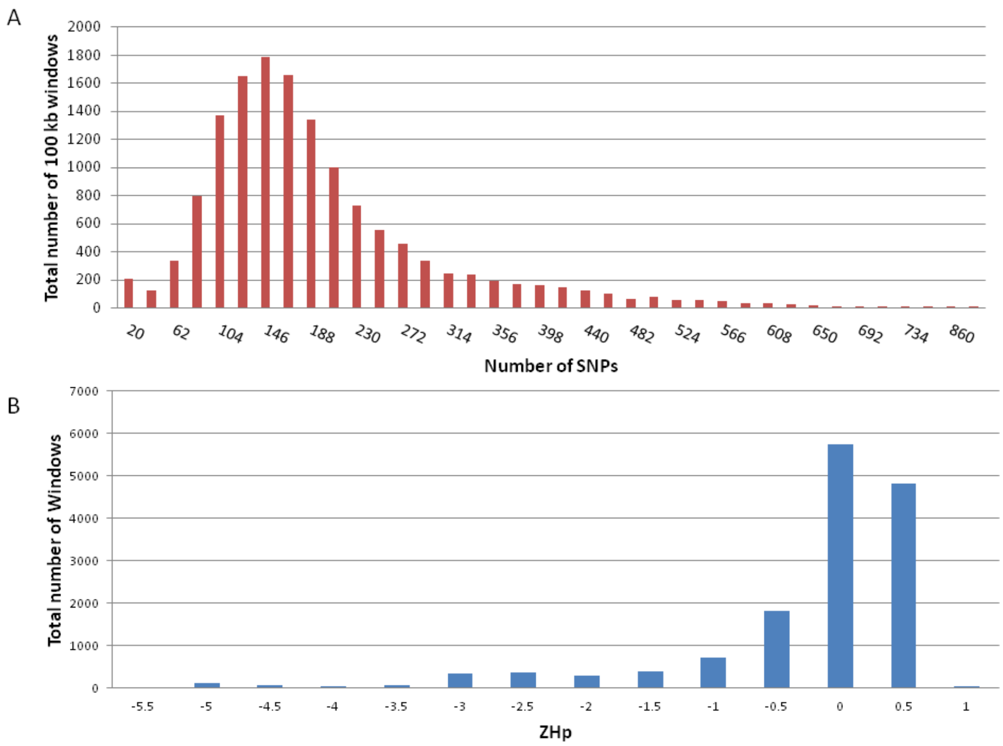

Figure S4.3: Distribution of the (A) SNPs and (B) ZHp values in the 100 kb BTA X windows on EASZ.

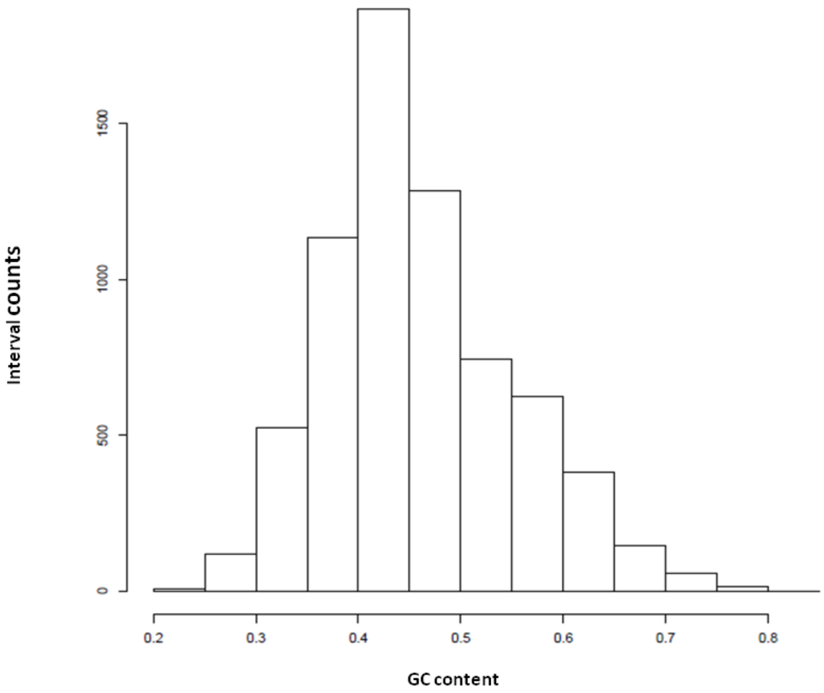

Figure S4.4: : The GC content histogram plot of the captured BTA X exome regions by the Agilent SureSelect XT target enrichment system.

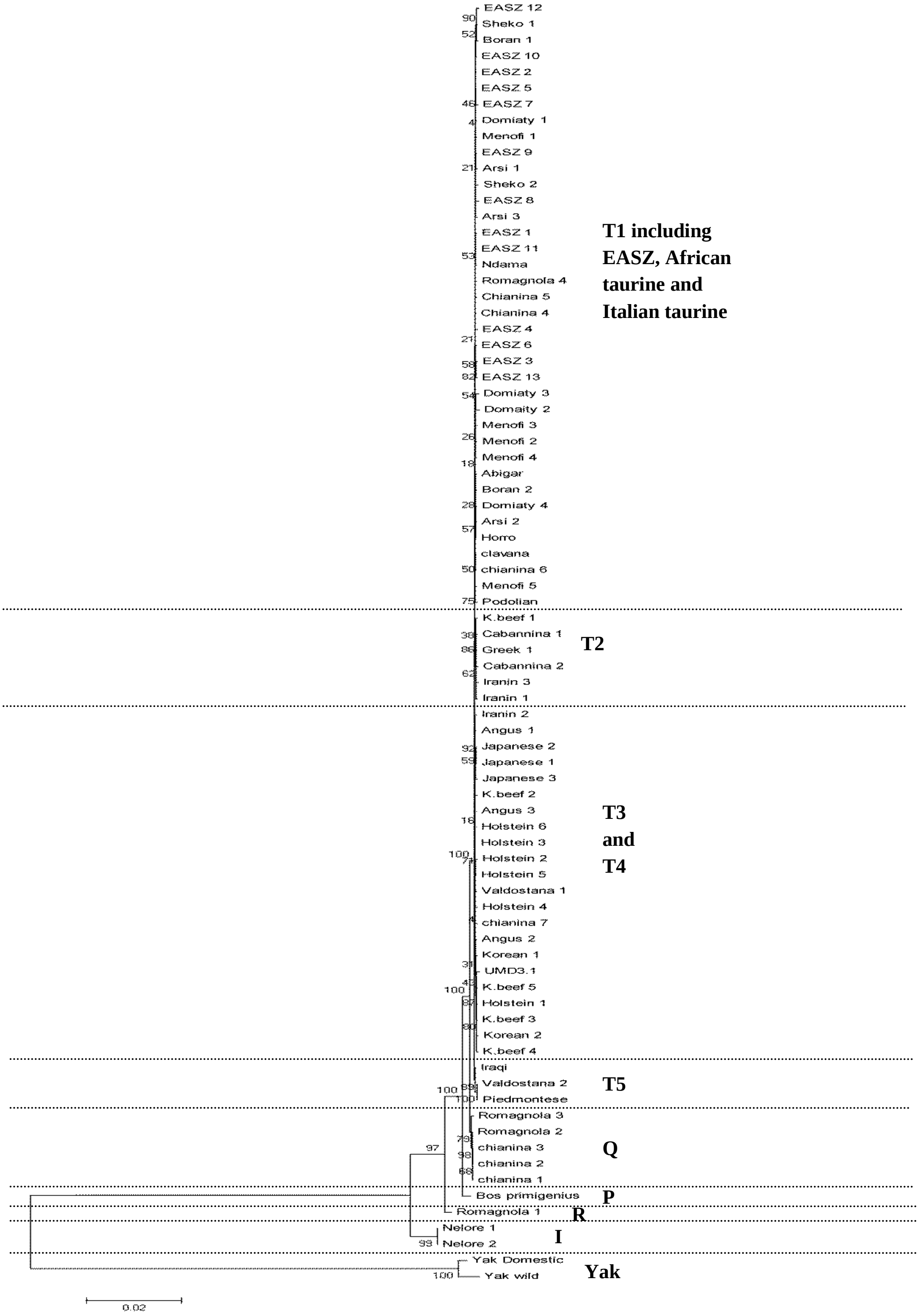

Figure S5.1: Rooted neighbour-joining (NJ) tree of all mtDNA included in Chapter 5 with 1000 bootstrap replication. Yak mtDNA was used as outgroup to root the tree. A scale bar (divergence of 0.02) is shown.

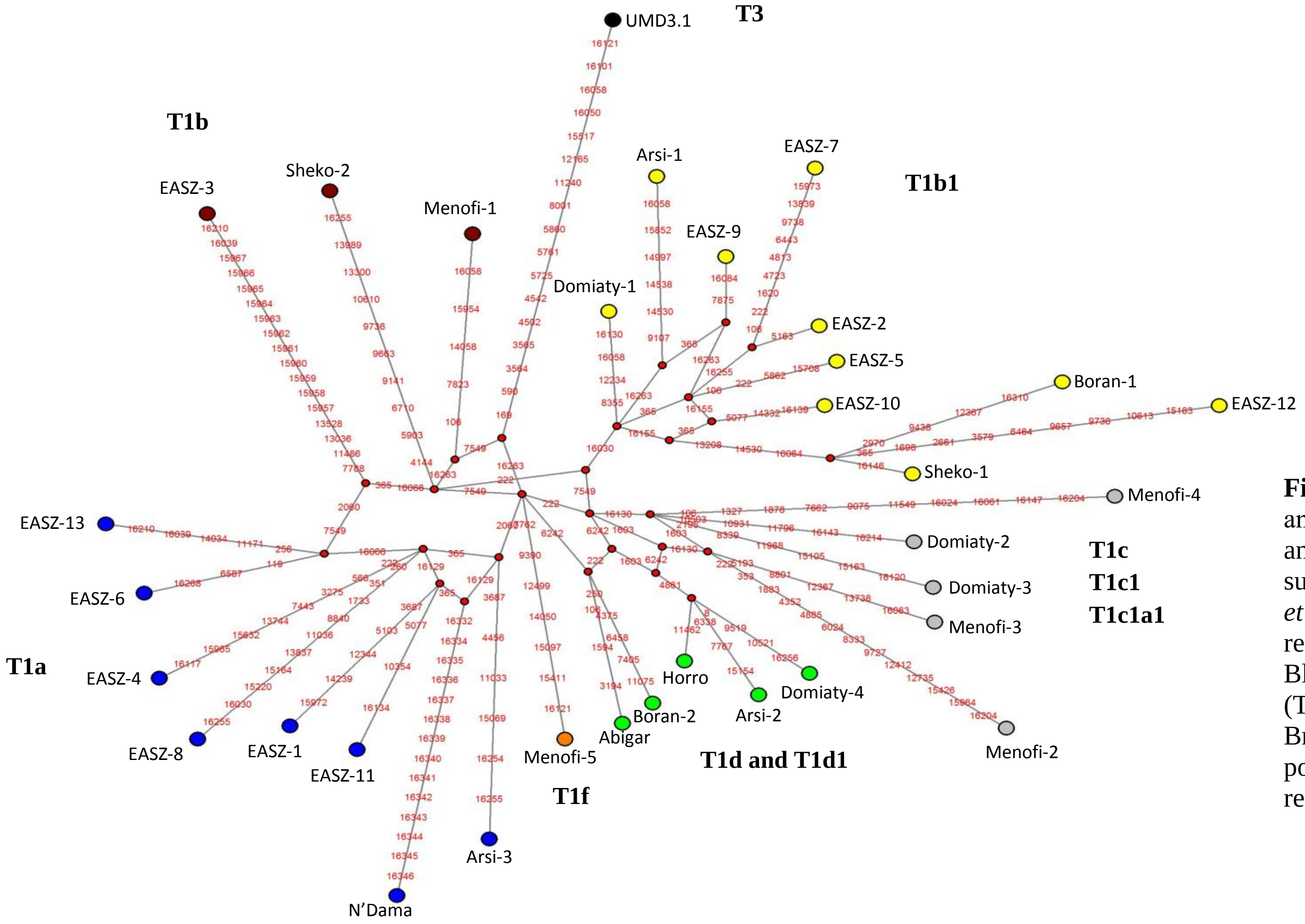

Figure S5.2: Median-joining network analysis including 13 EASZ, 1 N'Dama and 18 African cattle with known mtDNA sub-haplogroup types (refer to Bonfiglio *et al.*, 2012 and Table S5.1). Each node represents one haplotype. Black (T3). Blue (T1a). Green (T1d and T1d1). Grey (T1c, T1c1 and T1c1a1). Yellow (T1b1). Brown (T1b). Orange (T1f). Polymorphic positions correspond to UMD3.1 reference sequence including indels.