UNIVERSITÀ CATTOLICA DEL SACRO CUORE

Sede di Piacenza

Scuola di Dottorato per il Sistema Agro-alimentare

Doctoral School on the Agro-Food System

cycle XXIX

S.S.D: AGR/17 BIO/07 VET/06

Exploring livestock evolutionary history, diversity, adaptation and conservation through landscape genomics and ecological modelling

Candidate: **Elia Vajana**

Matr. n.: **4212128**

Academic Year 2015/2016

Scuola di Dottorato per il Sistema Agro-alimentare

Doctoral School on the Agro-Food System

cycle XXIX

S.S.D: AGR/17 BIO/07 VET/06

Exploring livestock evolutionary history, diversity, adaptation and conservation through landscape genomics and ecological modelling

Coordinator: Ch.mo Prof. Marco Trevisan

Candidate: Elia Vajana

Matriculation n.: 4212128

Tutor: Prof. Paolo Ajmone-Marsan

Dr.ssa Licia Colli

Academic Year 2015/2016

Contents

Ai miei amati nonni

To my beloved grandparents

Ringraziamenti

Vorrei innanzitutto esprimere la mia profonda gratitudine ai miei supervisori di tesi, il Prof. Paolo Ajmone Marsan, per la grande opportunità, il sostegno offerto lungo questo duro e formativo percorso, e la Dott.ssa Licia Colli, per il sostegno, la pazienza, le discussion scientifihe e non, l'amicizia dimostrata nei momenti più difficili.

Ringrazio di cuore i Professori Stéphane Joost e Michael W. Bruford per avermi osptitato presso i rispettivi laboratori di ricerca, offrendomi un'opportunità di crescita professionale ed umana unica ed irripetibile.

Non può essere quantificato nella parola "grazie" la riconoscenza ai ragazzi dell'*Animal Genetics Lab* di Piacenza: Riccardo, Marco, Elisa, Stefano, Lorenzo e successivamente Mario (Barbato), Marcello, Mario (Di Guardo), Chiara e Roberta. Prima di tutto, per la vicinanza, il supporto e l'amicizia che mi avete costantemente offerto; secondariamente, per l'aiuto nel lavoro quotidiano e la conoscenza che sempre avete condiviso e messo a disposizione. In particolare, un grazie di cuore a Marco, per la sua immensa disponibilità, esempio, la vicinanza umana e professionale; a Stefano, per l'altrettanto immenso aiuto offerto, e la nostra amicizia nata in cammino; a Mario, il cui fondamentale ed imprescindibile supporto umano, scientifico durante il nostro incontro a Cardiff—prima—e a Piacenza—poi—ha rappresentato una certezza e sprono fondamentale per il raggiungimento di questo traguardo.

Un grazie di cuore ai miei amici e colleghi di dottorato del XXIX ciclo Agrisystem per la compagnia, il reciproco sostegno e conforto; grazie ad André e Letizia per le nostre chiacchierate e l'aiuto che ci siamo offerti.

Ringrazio per l'accoglienza e l'ospitalità il gruppo di amici e colleghi del *Laboratory of geographic information systems* a Lausanne: Romain, per la sincera amicizia che hai saputo offrirmi fin dal mio arrivo (grazie!), Matthew per preziosi suggerimenti, Estelle per l'amicizia, l'aiuto offerto insieme alle formative discussioni scientifiche, Solange, Kevin.

Un enorme grazie per la loro allegria, vera amicizia a "*Los Cardiffianos*", coloro che hanno costituito la mia "famiglia" durante l'avventura Gallese: un grazie infinito a Natalia per le nostre serate, e per avermi fatto sentire a "casa" nonostante le centinaia di chilometri, Juanma e Isabel, Dani y Noemi per aver condiviso questo intenso tratto di cammino con un sorriso e una battuta sempre pronti: *¡Muchas gracias amigos!*

Un doverso grazie al Dr. Pablo Orozco-terWengel per la costante disponibililtà e aiuto dimostrati, insieme ad un orecchio sempre pronto all'ascolto, *gracias Pablo*…

Grazie di cuore al gruppo di amici microbiologi dell'Università Cattolica di Piacenza: Vania, Alessandro, Francesco, Elisa, Giulia, Sotos. Un grazie particolare a Vania, per l'amicizia e per le nostre lunghe (e fruttuose!!!) chiacchierate su modelli e comunità ecologiche in microbiologia, a Francesco per la tua amicizia e vicinanza anche nei momenti di lontananza, e ad Elisa, per la nostra bella amicizia.

Grazie ad Andrea per il tuo supporto ed ascolto, per i quali sempre ti sarò grato.

Di nuovo, la parola “grazie” non racchiude la gratitudine che provo nei confronti di ciascuno di voi, cari amici “di sempre”: medicina per tutte le mie solitudini, siete sempre stati il mio porto franco... Vi voglio bene ragazzi, semplicemente, GRAZIE per esserci…

Niente di tutto questo sarebbe stato possibile senza il costante ed incondizionato supporto della mia famiglia. Questo lavoro è dedicato a voi tutti, che sempre avete accettato e compreso i miei sacrifici, la mia lontananza, i miei momenti di difficoltà: a mia madre, per la quale dovrei scrivere un libro di ringraziamenti a parte, ai miei fratelli, Iaia e Ricky, ai miei nonni, mie sicure colonne… Grazie a mio padre, in particolare, per la sua positiva ispirazione e messaggio… Grazie a Giorgio, Nando, Teja, a nonna Gegia e zii.

Grazie a voi tutti per aver condiviso, al mio fianco, questa intensa parte di vita; ve ne sarò sempre grato.

1. General introduction

1.1 A general definition for biodiversity

The term ‘biodiversity’ was introduced by the entomologist Edward Osborne Wilson in 1986 as a fusion of the expression ‘biological diversity’, to indicate the “*variability among living organisms from […] terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part*”, or rather the “*diversity within species, between species and of ecosystems*” (*Secretariat of the Convention on Biological Diversity. Handbook of the Convention on Biological Diversity Including its Cartagena Protocol on Biosafety* 2005). Therefore, biodiversity can be conveniently described at different levels of biological complexity, starting from the genes carried by the populations composing a species, the species belonging to a particular biological community, and the ecosystems harboured in a defined region of the biosphere.

1.2 Evolution of livestock biodiversity

Livestock biodiversity is rather limited at the species level, counting approximately 30 mammalian and avian species, but extremely diversified at the genetic level (Simianer 2005). Domestication, i.e. the process of genetically adapting wild animals and plants to the human ends (Bruford *et al.* 2003; Driscoll *et al.* 2009), represents a fundamental turning point in the evolution of both human societies and modern-day livestock. On the one hand, it prompted agricultural development enabling the establishment of permanent settlements of farmers and crucial social rearrangements (Ajmone-Marsan *et al.* 2010); on the other hand, it substantially

contributed to shape the genetic makeup of the early tamed populations through initial genetic bottlenecks and subsequent selection[1] (Bruford *et al.* 2003).

Three explanations have been suggested to describe the first stages of domestication (Larson & Fuller 2014): (i) following the 'commensal pathway', some wild species populations (e.g. wolves) were attracted by the human niche, evolved 'synanthropic ecotypes', underwent habituation and commensalism to the anthropic habitat, and were finally domesticated; (ii) following the 'prey pathway', wild populations of large herbivorous (e.g. cattle and water buffalo) were firstly targeted by intense human hunting and then subjected to herd and breeding management in order to optimize food availability; (iii) a 'directed pathway' took place more recently (starting ~6,000 years before present) to domesticate specific species (e.g. horses, donkeys and Old World camels) for specific tasks (e.g. transportation).

Genetic information provided by mitochondrial and nuclear markers like microsatellites and Single Nucleotide Polymorphisms (SNPs) contributed to shed light on the complexity of domestication processes in most of the modern-day domestic species (see e.g. (MacHugh *et al.* 1997; Tapio 2006; Decker *et al.* 2014). For example, molecular evidence suggested the occurrence of two independent domestication events in as many geographic centres for cattle (*Bos taurus* and *Bos indicus*), water buffalo (*Bubalus bubalis*), and dogs (*Canis lupus familiaris*) (Kumar *et al.* 2007a; Ajmone-Marsan *et al.* 2010; Frantz *et al.* 2016), and an even more intricate scenario was suggested for pig (*Sus scrofa domesticus*) (Larson *et al.* 2005;

[1] During and after domestication process, farmers started to consciously select the most convenient phenotypic characteristics among those offered by the initial variability of the early tamed populations (Diamond 2002). For this reason, similar patterns of morphological and, in the case of animals, behavioural change appeared in different species after domestication: typically, domestic ruminant species (e.g. cattle and sheep) tended to show reduced or completely absent horns compared to their wild relatives, together with a contemporaneous reduction in body size (Ajmone-Marsan *et al.* 2010); at the same time, animals were selected for tameness, with a consequent reduction of senses acuteness and brain size. Indeed, these traits ceased to be adaptive under a strict human management (Diamond 2002).

Frantz *et al.* 2015).

Despite the complexity of each species history, recognizable patterns were described for several livestock species and for the evolutionary events following domestication (Bruford *et al.* 2003):

1) Most species were domesticated between 11,500 and 8,000 Years Before Present (YBP) (Bruford *et al.* 2003; Driscoll *et al.* 2009), in a precise set of areas generally located along an East-West axis, and often at similar latitudes. In particular, cattle, goats, sheep and pigs were most likely domesticated in two macro-areas, one encompassing the Fertile Crescent (along the Tigris and Euphrates basin), and another in Asia, spanning from the Indus Valley to some vast regions of modern-day China (Luikart *et al.* 2001; Larson *et al.* 2005). Similarly, recent findings based on both mtDNA and Y-chromosomal variation would suggest water buffalo ecotypes[2] ('river' and 'swamp') to derive from independent domestication events possibly occurred in the North-West of India and in a wide region encompassing China and South-eastern Asia, respectively (Kumar *et al.* 2006, 2007a; Yindee *et al.* 2010).
2) Domestication was generally followed by human-driven migrations out of the centres of origin[3] (Diamond 2002; Larson *et al.* 2014). Newly established populations generally suffered a gradual decrease in genetic diversity, especially as a consequence of subsequent founder effects not counteracted by gene flow over large distances (Bruford *et al.* 2003; Ajmone-Marsan *et al.* 2010). This trend is evident in both hardly transportable livestock species like cattle and sheep (Ajmone-Marsan *et*

[2] *Ecotype*: genetically distinct group of individuals within a species, which are adapted to specific environmental, conditions and inhabit a given geographical area.
[3] *Centre of origin*: geographical location where a taxon, either wild or domestic, firstly evolved: generally, centres of origin corresponds to hotspot of genetic diversity.

al. 2010), and in the more movable goats when evaluated with autosomal microsatellite markers (Cañón *et al.* 2006) (but see Luikart *et al.* 2001 for contrasting results based on mtDNA). Domesticated populations that were transported to new sites interbred with indigenous wild populations in several cases, giving rise to the so-called 'introgressive capture' (Larson *et al.* 2014).

3) The colonization wave was gradual in time and space during the thousands of years that followed domestication. Within such time span livestock populations settled in heterogeneous habitats became locally adapted[4] to specific environmental pressures. The traditional use of sustainable rearing techniques further facilitated the local adaptation process (Taberlet *et al.* 2008; Ajmone-Marsan & The GLOBALDIV Consortium 2010).

4) The introduction of the concept of 'breed[5]' around 200 years ago. At that time, farmers began to apply more systematic mating practices, crossing individuals with similar phenotypes to favour desirable traits (e.g. productivity or robustness), while avoiding interbreeding with groups showing different characteristics. Thus, domestic species experienced artificial fragmentation for the first time, which eventually increased within-breed undesirable effects of genetic drift (Taberlet *et al.* 2008).

5) The 'creation' and massive commercialization of industrial transboundary breeds[6] in the last decades to address an increasing food demand. Such an 'industrial revolution' in livestock was boosted by technological advances in quantitative

[4] Refer to section 1.4 for a detailed discussion on the process of local adaptation.
[5] *Breed*: a culturally accepted sub-specific group of domestic animals which share similar external characteristics and derive from a common geographic area and, possibly, genetic isolation (Scherf 2000; Blasco 2008; Hoffmann 2010a).
[6] *Transboundary breed*: breed which occurs in more than one country (Food and Agriculture Organization of the United Nations 2012).

genetics methods, leading to at least two implications of fundamental importance for the management and conservation of Animal Genetic Resources[7] (AnGR): (i) genetic diversity within industrial breeds was remarkably reduced, by causing effective population size[8] (N_e) to decay under the 'danger' threshold of 50 in several cases[9] (Taberlet *et al.* 2008); (ii) the evolutionary heritage represented by locally adapted[10] and indigenous breeds[11] started being eroded by genetic introgression and replacement with the more productive—and genetically homogeneous—industrial breeds.

6) Genetic erosion is particularly affecting local breeds in developing countries, with the actual risk of losing unique adaptations towards endemic diseases, environment and alternative farming systems (Ajmone-Marsan & The GLOBALDIV Consortium 2010).

1.3 The biodiversity crisis

The rapid decline in the amount of biodiversity, referred to as 'biodiversity crisis', has been affecting natural and agricultural landscapes during the last two centuries (Singh 2002; Koh *et*

[7] *Animal Genetic Resources* (*AnGR*): genetic diversity found in animals and microbes which already are (or might potentially prove) useful for human needs. Such a diversity can be already characterized or still uncharacterized, and does not necessarily refer to the sole domesticated animals.

[8] *Effective population size*: Size of the idealized Wright-Fisher population which would show the genetic properties observed in the population under study (Wang 2005). An idealized Wright-Fisher population is assumed to have constant size, non-overlapping generations, random mating among individuals and genotype frequencies in Hardy-Weinberg equilibrium in the case of sexual diploids.

[9] An effective population size of ~50 is generally suggested to avoid inbreeding depression in the short term (in the next five generations; Kristensen *et al.* 2015); $N_e \geq 500$ is deemed to preserve long-term evolutionary potential (Franklin & Frankham 1998).

[10] *Locally adapted breed*: breed residing in a single country for a sufficient time to be genetically adapted to one or more traditional production systems or local environments (Food and Agriculture Organization of the United Nations 2012).

[11] *Indigenous breed* (*alias* "*autochthonous*" or "*native breeds*"): breed adapted to and utilized in a single, particular geographical region; indigenous breeds constitute a subset within locally adapted breeds (Food and Agriculture Organization of the United Nations 2012).

al. 2004): species extinction in the wild is estimated to occur around 1,000 times faster than the inferred background rates (De Vos *et al.* 2015), 1-2% of the total amount of domestic breeds is reported to disappear each year (Simianer 2005), 17% to be either "endangered" or "critically" maintained" (FAO 2015), and up to 60% to present a still unknown risk status (FAO 2015).

Biodiversity crisis endangers ecosystem functioning and basic services (Gamfeldt *et al.* 2008; Mace *et al.* 2012), erodes the adaptive potential of natural and domestic populations towards environment challenges or new market demands (Kotschi 2007; Bellard *et al.* 2012), undermines food security (Frison *et al.* 2011) and ultimately threatens human well-being (Ceballos *et al.* 2015). Anthropogenic impact on the biosphere (Vitousek *et al.* 1997), together with economical choices favouring short-term agricultural productivity in spite of variability preservation (Taberlet *et al.* 2008), are both suggested as the main causes of such decline (Galaz *et al.* 2015).

1.3.1 The "Noah's ark" problem

The Convention on Biological Diversity (CBD) formally acknowledged the central role of biodiversity in providing "the goods and services that sustain our lives", and states the urgency of conserving the evolutionary heritage in order to attenuate human foot-print and favour a sustainable exploitation of the biological resources[12] (*Secretariat of the Convention on Biological Diversity. Handbook of the Convention on Biological Diversity Including its Cartagena Protocol on Biosafety* 2005).

[12] *Biological resources*: include genetic resources, organisms, populations and any biotic component of ecosystems with "actual or potential use or value for humanity" (*Secretariat of the Convention on Biological Diversity. Handbook of the Convention on Biological Diversity Including its Cartagena Protocol on Biosafety* 2005).

However, the achievement of CBD's goal is hindered by the limited amount of economic resources available for biodiversity conservation. In the case of livestock, the resources available overall are insufficient to grant protection to all existing breeds (Bennewitz *et al.* 2007); analogously, resources for wildlife conservation are inadequate in the majority of developing countries where a high amount of biodiversity and elevated threats to ecosystems are typically concomitant (Brooks *et al.* 2006). Here the fundamental question conveying the "Noah's ark" problem in conservation biology (Weitzman 1998): which species—or populations and ecosystems—should deserve priority for conservation in order to minimize loss in biodiversity "under a limited budget constraint"?

1.3.2 The need of conserving Animal Genetic Resources

Animal Genetic Resources are commodities of primary conservation concern, since they represent specific adaptations to current environmental and market conditions (Anderson 2003), and constitute a potential reservoir of adaptive genes for future socio-environmental scenarios (Notter 1999). Therefore, characterization of AnGR is formally recognized as a *Strategic Priority Area* within the *Global Plan of Action for Animal Genetic Resources* (FAO 2011), as it constitutes the preliminary step to assess breeds' value for conservation and the basis for sustainable breeding programmes. However, although representing around two-thirds of the total livestock biodiversity, AnGR of locally adapted and indigenous breeds living in developing countries are scarcely characterized (Ajmone-Marsan & The GLOBALDIV Consortium 2010; Hoffmann 2010a). Such a lack of information might prove detrimental, as these AnGR are expected to become crucial in the near future to respond to changes in climatic conditions, disease/parasite distribution or market demands (Hoffmann 2010b).

Therefore, an adequate characterization of livestock biodiversity and subsequent setting of conservation priorities are required to avoid losing such a unique reservoir of genetic variants and evolutionary potential.

1.4 Animal Genetic Resources and local adaptation

The characterization of genes conferring adaptation to specific environmental conditions is a core topic in evolutionary biology (Tenaillon & Tiffin 2008), with key implications for AnGR conservation under the light of current climate change and upcoming demands in food safety and production (Savolainen *et al.* 2013).

To allow spatially divergent selection to take place, populations from different geographical sites must experience heterogeneous selective pressures on ecologically relevant traits. Divergent selection is considered the main driver prompting 'local adaptation' (Kawecki & Ebert 2004), which is the process leading a population to present a "higher fitness at its native site than any other population introduced to that site" (Savolainen *et al.* 2013). Local adaptation is a genetic adaptive process requiring the existence of alternative alleles and genotypes for the same locus within the considered demes[13]. The genetic nature of local adaptation distinguishes it from adaptive phenotypic differentiation, in which a single genotype can result in multiple phenotypes due to phenotypic plasticity (Chevin *et al.* 2010).

Theoretically, if (i) spatially divergent selection is sufficiently constant over time, and sufficiently strong to counteract the homogenizing effect of gene flow, (ii) locally adapted optimal genotypes are favoured in the native site but strongly disadvantaged in the others, (iii) evolution of adaptive phenotypic plasticity is hindered by some evolutionary costs or

[13] *Deme*: local population displaying a distinct gene pool.

constraints, and (iv) populations are large enough to render the confounding effects of genetic drift negligible, then conditions are expected to be favourable for local adaptation to evolve and be detected (Kawecki & Ebert 2004; Yeaman & Otto 2011). Conversely, the lack of sufficient standing genetic variation within populations is expected to hinder a rapid process of local adaptation (Kawecki & Ebert 2004; Savolainen *et al.* 2013).

1.4.1 The genetics of local adaptation

The study of the genetics underlying local adaptation can be tackled by either 'top-down' or 'bottom-up' approaches.

In the first case, candidate demes for local adaptation have to be first identified and adaptive traits of interest measured. Reciprocal transplant experiments represent the traditional framework for identifying locally adapted demes. In this kind of tests, individual phenotypic characteristics (e.g. reproductive output) are recorded to measure the average fitness of at least two demes in their native and non-native habitats, respectively (Savolainen *et al.* 2013) (Figure 1.1a and 1.1b). When evidence of local adaptation exists for the studied demes, recorded traits are then related with underlying genotypes through quantitative trait loci mapping (QTL) (Rellstab *et al.* 2015). Two basic genetic mechanisms are argued to sustain local adaptation at an individual locus or QTL (Anderson *et al.* 2013): (i) 'antagonistic pleiotropy', which occurs when alternative alleles confer higher fitness in different habitats (Figure 1c); and (ii) 'conditional neutrality', which occurs when an allele confers a fitness advantage in one habitat, while being neutral in the non-native site (Figure 1d).

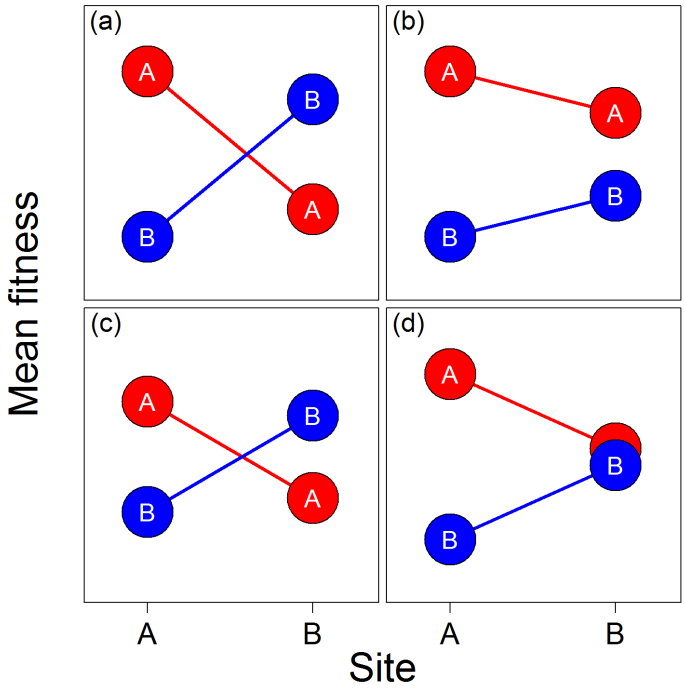

Figure 1.1 Fitness comparisons among demes (figures **a** and **b**) and alternative alleles at a single locus involved into local adaptation (figures **c** and **d**). Red circles represent mean fitness for demes and alleles native of site A; blue circles represent average fitness for demes and alleles originating in site B. (**a**) Both demes display higher fitness at their native sites when compared with 'foreign' demes, by satisfying the so-called 'local vs. foreign' criterion. (**b**) 'Home vs. away' pattern, in which both demes A and B show higher fitness in their own home-site and decrease fitness in the non-native sites. In this case, 'local vs. foreign' criterion is not met, as deme A performs better in both its native and non-native sites. As a result, local adaptation pattern is supported only in Figure 1.1a, where both 'home vs. away' and 'local vs. foreign' criteria are satisfied. (**c**) Native allele of site A confers higher fitness in its own home-site, as do the native allele from site B: antagonistic pleiotropy is suggested for the concerned locus. (**d**) Native allele from site A confers higher fitness in its own home-site, while showing no effect on fitness in the non-native site; in this case, conditional neutrality is suggested for the concerned allele.

Alternatively, 'bottom-up' approaches allow to bypass the transplant experiment design, by relating the highlighted loci with either specific evolutionary processes (e.g. positive selection) or the environmental driver promoting local adaptation (Rellstab *et al.* 2015). In turn, two types of 'bottom-up' approaches have been described:

1) Population genetic methods are used to measure differentiation between populations

at the DNA level (Savolainen *et al.* 2013). In particular, genome-scan methods can be used to obtain individual loci estimates of Wright fixation index for population differentiation (F_{ST}), and highlight F_{ST} outliers on the basis of empirical or expected distributions under neutral models of evolution (Akey *et al.* 2002; Bonin *et al.* 2007; Foll & Gaggiotti 2008). Theoretically, local adaptation is expected to produce high differentiation (i.e. $F_{ST}\approx1$) for those loci under selection, while not affecting neutral loci which are expected to show F_{ST} values within the ranges of the null expectations (de Villemereuil & Gaggiotti 2015). However, local adaptation is often driven by polygenic quantitative traits (Savolainen *et al.* 2013), whose underlying genotypes may show little differences in allele frequencies between populations (Rellstab *et al.* 2015) which might not be detected by F_{ST}-based methods (Pritchard & Di Rienzo 2010). Furthermore, population genetic methods are potentially unable to discern true local adaptation from anthropogenic signatures of selection in the case of domestics, by imposing caution in the interpretation of the obtained outliers in this context.

2) Environmental (or genetic-environment) association analysis allows to directly associate variations in habitat features with the genetic variability of populations, thus potentially revealing adaptive loci (Mitton *et al.* 1977). The rationale behind a genetic-environment association analysis is that genetic variants (alleles or genotypes) showing a significant association with a particular habitat feature are likely to be involved into adaptation mechanisms with the concerned environmental feature (e.g. precipitation, soil type or a disease).

1.4.2 Landscape genomics

One of the last developments within the domain of genetic-environment association analysis is represented by landscape genomics, which took advantage of the concurrent development of next-generation sequencing (NGS) and high-throughput genotyping techniques, as well as recent improvements in the environmental datasets describing habitat characteristics (e.g. temperature, precipitation, vegetation, etc.) (Rellstab *et al.* 2015). Landscape genomics aims at uncovering the environmental drivers of local adaptation and the underlying candidate genes/gene networks (Manel *et al.* 2010). To this end, it searches for significant associations between the habitat characteristics and the genetic makeup of sampled individuals or populations. Therefore, the approach requires the collection of both genetic and environmental information at the same locations (Joost *et al.* 2007), and a careful planning of the sampling design in terms of both environmental variability coverage and replication (Joost *et al.* 2007; Rellstab *et al.* 2015).

1.4.2.1 The need to account for neutral population structure

Associative tests used in landscape genomics introduce the possibility of detecting a number of spurious signals due to the possible confounding effect of the underlying genetic structure of the studied demes (Excoffier *et al.* 2009). Population structure evolves as a result of historical demographic processes like gene flow and genetic drift shaping allele frequencies at neutral loci. Individuals from the same deme are likely to share a common demographic history, and may be genetically more similar to each other at neutral loci than individuals coming from different sites. Therefore, if demes are genetically structured while inhabiting areas with different habitat features, environmental and neutral variability may result collinear, and population structure can mimic the effect of divergent selection inducing false positive

detections among the neutral markers (Rellstab *et al.* 2015).

Therefore, accounting for neutral genetic population structure is considered of primary importance in landscape genomics models to reduce the number of spurious detections (De Mita *et al.* 2013). Several approaches have been suggested to correct for genetic structure, which rely on: pairwise Euclidean distances between sampling locations (Guillot *et al.* 2014), spatial autocorrelation of individuals within populations (Poncet *et al.* 2010), individual Q-scores derived from global ancestry analyses (Pritchard *et al.* 2000; Alexander *et al.* 2009), and principal component scores derived from principal component analysis (PCA) performed on individual genotypes (Eckert *et al.* 2010). Ideally, analyses based on molecular information should be run on the neutral loci exclusively, in order to avoid losing putative adaptive signals.

1.4.2.2 Statistical associative models in landscape genomics

Landscape genomics techniques can be population- or individual-based (Rellstab *et al.* 2015): if both genetic and environmental information are expressed at the population level (i.e. a locus is represented by the frequency of one of its alleles in the populations under study), then population-based methods can be used to investigate significant genome-environment associations (see e.g. Turner *et al.* 2010); conversely, if genome-environment associations are modelled at the level of single individuals (i.e. each individual represents a separate sampling unit, with both genetic and environmental information available), then an individual-based approach can be applied (see Box 2 in Rellstab *et al.* 2015).

Since its implementation within the Spatial Analysis Method (SAM; Joost *et al.* 2007), logistic regression (LR) has represented a valuable individual-based approach to detect signatures of local adaptation in several animal and plant species (see e.g. Nielsen *et al.* 2009; Colli *et al.*

2014; Quintela *et al.* 2014). In the context of environmental association analysis, LR allows to model the probability of each individual to carry a particular allele or single-locus genotype as a function of the habitat features at the sampling site. Since each genotype is by definition georeferenced, the goal of the analysis is to detect environmental factors significantly associated with (and thus putatively affecting) the spatial distribution of the genetic variants under study (Rellstab *et al.* 2015). Recently, SAM approach has been improved to allow multivariate logistic regression analysis through the software SAMβADA (Stucki *et al.* 2016). Multivariate logistic regression allows to correct genome-environment associations for neutral population structure, an implementation which is expected to reduce the relatively high rate of false positives characterizing univariate logistic regression tests (De Mita *et al.* 2013).

Mixed-effects regression modelling has been recently proposed to provide the possibility of concurrently testing genome-environment associations while accounting for the neutral structure of the studied populations. Within this framework, spatial distribution of allelic or single-locus genotypic frequencies is predicted as a function of the tested environmental factors and the neutral population structure, the former being modelled as fixed effects and the latter as a random effect. Mixed-effects population-based models can be run with the software BAYENV (Coop *et al.* 2010; Gunther & Coop 2013), which can detect low rates of false positives (De Mita *et al.* 2013); conversely, an individual-based sampling design can be accommodated by LFMM (Frichot *et al.* 2013; Frichot & François 2015), an approach able to concurrently control for random effects due to population structure and spatial autocorrelation, and to provide rates of false positives comparable to BAYENV (Rellstab *et al.* 2015).

1.4.2.3 Merits of landscape genomics and future research

Although biased by higher rates of false positives when not adequately correcting for

population structure, landscape genomics was shown to be more powerful than F_{ST}-based methods in detecting signatures of local adaptation (De Mita *et al.* 2013; Savolainen *et al.* 2013). In fact, statistical models applied in genetic-environment association analysis are generally able to detect even subtle differences in allele frequencies between demes, a pattern often associated with local adaptation processes either occurring in the presence of high gene flow between demes (Rellstab *et al.* 2015), or due to ecologically relevant polygenic traits (Rockman 2012; Sork *et al.* 2013).

Therefore, the principal merits of landscape genomics are (i) the increased statistical power while accounting for neutral population structure, and (ii) the possibility of directly uncovering the environmental drivers of local adaptation. These characteristics make landscape genomics a valid option to investigate the genetic bases underlying local adaptation processes in both natural and livestock populations, especially those reared under management systems with limited human intervention (Pariset *et al.* 2012).

Nevertheless, further research is needed to develop approaches explicitly accounting for the polygenic nature of quantitative adaptive traits (but see Legendre & Legendre 2012), and to post-hoc validate the discovered putative variants in the field and/or in the laboratory (Rellstab *et al.* 2015).

1.5 Aim of the thesis

The main objective of this thesis is to contribute to the process of characterization and conservation of biological resources prompted by the Convention on Biological Diversity (*Secretariat of the Convention on Biological Diversity. Handbook of the Convention on*

Biological Diversity Including its Cartagena Protocol on Biosafety 2005) and the Food and Agriculture Organization of the United Nations (FAO 2011).

Within such a context, this thesis aims at achieving three specific goals:

1) To review methods proposed to prioritize biodiversity for conservation, suggest a classification framework, and propose a decision-aiding scheme for the selection of the most appropriate methodologies given a conservation goal (**Chapter 2**). Such a scheme aims at (i) unifying prioritization methods for conserving natural and agricultural biodiversities, and (ii) identifying methodological gaps in the current literature. As a result, possible new research avenues are envisaged and discussed.
2) To characterize the genetic diversity and provide hints on the evolutionary history of *Bubalus bubalis* (water buffalo) (**Chapter 3**). In this case study, the new 90K Affymetrix Axiom® Buffalo Genotyping Array was used for the first time after its development by the International Buffalo Consortium[14]. Water buffalo is one of the most economically important domestic species (Scherf 2000), providing both dairy products and animal traction especially in India and South-East Asia. While the scientific community seems now to converge on two independent domestication events for the river-type *B. bubalus bubalis* and the swamp-type *B. bubalis carabanensis* (Kumar *et al.* 2007a; Yindee *et al.* 2010), debate is still open around the geographical locations of the putative domestication centres and the post-domestication migration routes. The present work addresses both questions while providing a worldwide view of the genetic diversity patterns within the species.

[14] The International Buffalo Consortium collected research institutions from several countries of the world to sequence *B. bubalis* genome and provide a new species-specific SNP chip. The Institute of Zootechnics of the Università Cattolica del S. Cuore participated as a partner and was in charge of describing worldwide patterns of buffalo genetic diversity.

3) To uncover putative adaptive loci and genes underlying local adaptation towards East Coast Fever (ECF) while providing hints about their ancestral origin (**Chapter 4**). ECF is an endemic vector-borne disease caused by the protozoan *Theileria parva parva* and affecting susceptible cattle populations of Sub-Saharan Africa. A landscape genomic approach was used to relate SNP data from indigenous cattle populations of Uganda with two environmental proxies of the disease selective pressure, i.e. the spatial distribution of the *T. parva parva* vector (the brown ear tick *Rhipicephalus appendiculatus*), and the infection risk by *T. parva parva*. Further, the evolutionary origin of the highlighted genomic regions was investigated by means of local ancestry analyses, i.e. methods allowing to infer the ancestry of specific chromosome segments on the basis of a chosen set of reference populations (Brisbin *et al.* 2012).

2. Prioritizing ecosystems, taxa and genes: a unified framework for conserving wild and agricultural biodiversity

Elia Vajana, Licia Colli, Pablo Orozco-terWengel, Mario Barbato, Stefano Capomaccio, Paolo Ajmone-Marsan* & Michael W. Bruford*

*Co-senior authorship

2.1 Abstract

The biodiversity crisis is jeopardizing both natural and agricultural systems: an increasing number of species is becoming extinct, and the evolutionary potential of both wild and domestic populations is at risk. Typically, economic resources invested in conservation are limited, and priorities must be devised to stem losses in ecosystems, species and at the genetic level. The term 'prioritization' has been traditionally referred to the process of defining conservation rankings on the basis of criteria reflecting precise biological attributes of the systems concerned. More recently, it has also been associated to methods optimizing allocation of a defined amount of resources between competing strategies, projects or actions to maximize biodiversity protection. Here we review prioritization methods from the wildlife and livestock conservation literature and propose a general classification framework suitable for both sectors. First, methodologies are classified into 'biological prioritization methods' or 'resource allocation methods', then referred to a targeted level in biodiversity hierarchy (i.e. landscape, ecosystem or species), and are lastly identified by unambiguous prioritization criteria. As a result, we propose a decision tree to support selection of the most pertinent

approaches, given predefined prioritization goals and targets. We also discuss potential generalizations of methods normally applied in the sector of origin, by revealing great potential for profitable scientific exchange between wild and domestic communities. Finally, we envisage unexplored methodological integrations, and discuss the role that emerging genomic technologies will potentially play in the context of biodiversity prioritization.

Keywords: Natural and agricultural biodiversity, conservation, biodiversity prioritization, biological prioritization problem, conservation resource allocation problem, decision tree.

2.2 The biodiversity crisis and prioritization

Biodiversity is defined as the "variety of life" existing at all levels of biological organization, i.e. ecosystems, species and genes (Primack & Ralls 1995; Gaston 2000). More specifically, 'agricultural biodiversity' refers to the ecosystems, species and genetic variation which support human nutrition and agriculture (Frison *et al.* 2011).

Wild and agricultural biodiversity is experiencing a profound, generalized crisis (Thomas *et al.* 2006): ecosystems are degrading, undermining fundamental services at the basis of natural and agricultural balances; species are disappearing at an unprecedented rate (Ceballos *et al.* 2015); genetic diversity is being eroded with consequent reduction in species adaptive potential to future environmental or market conditions.

Anthropogenic change is the primary cause of decline for both components of biodiversity (Galaz *et al.* 2015). Climate change and biosphere pollution are global phenomena with profound implications at the landscape and ecosystem levels, while habitat loss and the spread of alien invasive species mainly threaten wild species' survival. Artificial fragmentation of

populations is a common threat to the genetic health of wild and agricultural species, whereas modern breeding schemes represent a particular risk for the gene pool diversity of cosmopolitan breeds in the livestock industry (Taberlet *et al.* 2008).

Safeguarding biological diversity is among the most pressing and fundamental challenges facing humanity, since it represents a basic requirement to guarantee a sustainable future for coming human generations. Despite efforts in the last decades, ongoing conservation programs have proved to be insufficient in slowing down the rate of biodiversity loss (Eizaguirre & Baltazar-Soares 2014). This partial failure can be mainly attributed to a constantly increasing anthropogenic pressure on the biosphere (Butchart *et al.* 2010), and, importantly, the scarcity of economic resources that have been invested in conservation (Master 1991; Boettcher *et al.* 2010). Because of these budget constraints, protection cannot be granted equally to all threatened ecosystems, species or populations, and priorities must be set in order to optimize conservation of what remains (Vane-Wright *et al.* 1991). To this aim, a number of methods have been proposed, and prioritization has become a core approach for NGOs, government agencies and institutions devoted to biodiversity conservation (Game *et al.* 2013).

Despite the topic's importance, a general scheme disentangling the network of prioritization techniques coming from the wild and the domestic literatures is still missing. The present review therefore aims to (i) propose an ontology of prioritization methods currently available for preserving wild and agricultural biodiversities, (ii) provide a decision tool for selecting the most appropriate methodology given specific conservation targets, (iii) suggest, whenever possible, more generic application of the reviewed prioritization methods (i.e. the possibility to utilize methods in both conservation sectors, natural and agricultural), and (iv) discuss methodological improvements or gaps in the current literature to address future research goals.

2.3 An ontology for prioritization methods

2.3.1 Biological prioritization and resource allocation problems

The problem of how identifying priorities in conservation can be described as following two approaches.

The first addresses the question: Which are the ecosystems or taxa deserving the highest priority for conservation, when provided with a set of possibilities and defined conservation criteria? This issue will be referred to as the 'biological prioritization problem', in that priorities are ascribed on the basis of precise biological attributes of the system studied (e.g. regional species richness or genetic diversity). In this case, neither competing conservation actions nor related costs are considered. Biological prioritization methods (BPMs) can be further distinguished between 'direct' and 'indirect': the former being explicitly conceived for prioritizing biological resources, the latter being developed for different purposes but can be adapted to be applied to biological prioritization.

The second approach addresses the question: What are the best actions for optimizing biodiversity conservation, given a defined prioritization criterion, a set of options, and an explicit conservation budget to be invested? We borrow the expression 'conservation resource allocation problem' from (Wilson *et al.* 2006) for referring to this approach. Being devised within the framework of decision support science, resource allocation methods (RAMs) generally prioritize actions guaranteeing the best investment returns (e.g. the effective number of species protected) given a fixed quantity of conservation funds. In some circumstances, RAMs can provide optimal resource allocation among the priorities first highlighted by BPMs.

2.3.2 A decision tree approach for classifying prioritization methods

Here, a decision tree approach is proposed for classifying prioritization methods through four decision steps (Figure 2.1):

1. Selection of the general prioritization approach (biological prioritization or conservation resource allocation).
2. Selection of a level in the biodiversity hierarchy targeted (landscape, ecosystem or species). Typically, landscape level-methods focus on ecological communities; ecosystem level-methods rank and allocate resources among species (not necessarily coming from the same ecosystem); species level-methods prioritize and distribute resources among populations within the same species (including based on genetic data).
3. Selection of a prioritization criterion. At the landscape level, choices are made based upon ecosystem uniqueness, species richness, endemism content, community composition, taxonomic diversity as well as evidence for ongoing evolution. At the ecosystem level, BPMs allocate priorities using among-species genetic diversity, taxonomic and genetic distinctness, environmental threats or extinction risk; RAMs rely on effective numbers of species protected, demographic indicators of conservation status, and among-species genetic diversity. At the species level, priorities mirror contributions to total genetic diversity (either in terms of among- and within-population diversity or adaptive and neutral diversity), adaptive variability, demographic dependence, extinction risk, or genetic uniqueness.
4. Selection of a prioritization method.

In the following sections, a review is provided featuring representative methods addressing

both types of prioritization problem. In the case of BPMs, discussion is separated between direct and indirect methods.

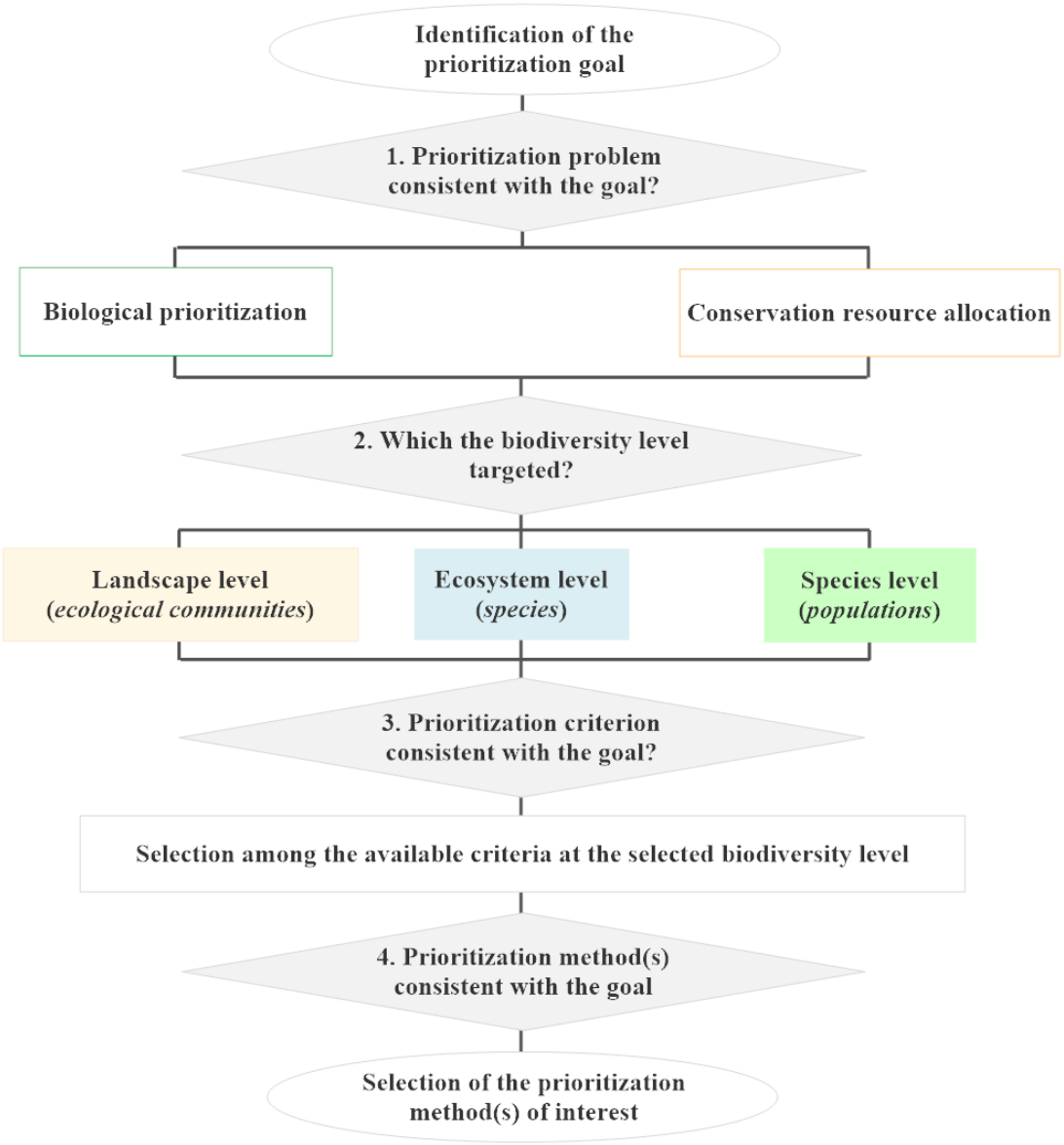

Figure 2.1 Decision tree-like approach supporting selection of the available prioritization methods. Having identified a precise prioritization goal, decision steps (grey boxes) include: (1) the addressed prioritization problem (a choice which reduces to the possibility/willingness of accounting for the economic aspect related to the prioritization goal); (2) the targeted level in biodiversity hierarchy (in brackets are the targeted biological units, i.e. ecological communities, species or populations); (3) the prioritization criteria given the selected problem and biodiversity level; (4) the available methods for addressing the specific prioritization goal.

2.4 The biological prioritization problem

2.4.1 Direct biological prioritization

A large number of methods were proposed to directly prioritize biodiversity for conservation (Figure 2.2, Table 2.1). The fundamental principles of 'complementarity' and 'rarity' were firstly introduced in the context of spatial prioritization. The former states that the addition of a new site to a set of protected areas only makes sense if this place adds new biodiversity value (Justus & Sarkar 2002), implying that sites with higher endemism (i.e. "rare sites") should deserve priority for conservation (Sarkar 2014). A number of approaches rely on these principles for defining *conservation area networks* (CANs), groups of geographical regions optimizing biodiversity content or composition. *Critical faunal analysis* (Ackery & Vane-Wright 1984), for instance, applies both complementarity and rarity to identify the minimal set of areas containing at least one population of all the considered species. The *biodiversity hotspots approach* (Myers 1988) designates priority areas on the basis of endemism and considering the level of threat to ecosystems. *Theoretical priority area analysis* (Vane-Wright *et al.* 1991) incorporates critical faunal analysis and the cladistic method (May 1990) to provide a set of areas maximizing the percentage of phylogenetic diversity conserved. The *ecoregion approach* is similar to the biodiversity hotspots approach but focused on ecosystem uniqueness rather than a region's endemism (Olson & Dinerstein 2002). Different ecosystem typologies harbour unique communities, whose protection can only be guaranteed if at least a part of the ecosystem—i.e. an *ecoregion*—is prioritized for conservation. Marine and terrestrial ecoregions were then tested for irreplaceability and distinctiveness, and a representative list of Earth's ecosystems (the 'Global 200') suggested as priorities for

conservation. In contrast, Erwin 1991 introduced the concept of the '*evolutionary front*' to stress the importance of prioritizing lineages that are actively evolving, to optimize the largest amount of evolutionary potential regardless of its rarity value.

At the ecosystem level, *Weitzman's diversity theory* (Weitzman 1992, 1993) represents a cornerstone for biological prioritization. Relationships between species are evaluated by a genetic distance matrix, and total diversity is defined as the length of the derived phylogenetic tree. This approach requires the definition of species-specific extinction probabilities, so that 'marginal diversities' can be computed to quantify the expected decrease in the total diversity occurring if the extinction probability of a species in the set would increase by one unit, due to an absence of conservation actions. The product of the extinction probability and marginal diversity defines the 'conservation potential' for each component of the set, by providing an objective way for defining biological priorities as a function of genetic distinctiveness and extinction risk (Boettcher *et al.* 2010). Although the Weitzman method was first demonstrated for prioritizing wild species (Weitzman 1993), it has instead found wide application in domestic populations. As a result, many more livestock breeds have been prioritized on the basis of their relative contribution to total and marginal diversities (Cañón *et al.* 2001; Reist-Marti *et al.* 2003) than have wild populations. However, several authors have criticized application of the Weitzmann approach at the species level, as total diversity coincides with the between-population diversity component, thus disregarding within-population variability which also represents a significant component of diversity and which is known to correlate itself with extinction risk (Caballero & Toro 2002; Toro & Caballero 2005). Unfortunately, Weitzman priorities often coincide with the most distant and inbred populations (European Cattle Genetic Diversity Consortium 2006), a case not always desirable in domestic species

where a significant goal for conservation is maximizing the amount of both within- and between-breed variability.

In order to address such criticisms, García *et al.* (2005) applied a diffusion process approach to compute genetic instead of physical extinction probabilities, and proposed their use to represent within-population diversity. Genetic extinction probabilities were defined to reflect homozygosity in populations, and computed as a population-specific probability of fixation averaged across the considered loci.

Alternatively, total genetic diversity can be explicitly partitioned into a between- and a within-population component. In this context, Ollivier & Foulley (2005) proposed to derive '*aggregate diversities*' to represent partial contributions to global variability, and set conservation priorities accordingly. Total within-population diversity was expressed as the mean expected heterozygosity over the studied units, and Weitzman methodology subsequently applied to compute partial merits to both between and within-population components. Therefore, aggregate diversities were derived to represent relative contributions to global diversity, by linearly combining population-specific partial merits. Marginal diversities and conservation potentials were also calculated either referring to the between- or within-population components, to provide a further basis for priority setting. Both the García and the aggregate diversity methods were proposed and applied for livestock breed conservation, but would remain conceptually valid also in the case of natural populations.

Conversely, Petit *et al.* (1998) did not rely on Weitzman methodology to evaluate between- and within-population components of total genetic diversity. Instead, they used Nei's diversity measures (Nei 1973) to define population-specific contributions to total gene diversity. Two components, i.e. 'diversity' and 'differentiation', were estimated for each population to

account for its contribution to the overall gene variability. In this way, populations mostly contributing to diversity can be evidenced, together with the reason of their contribution (i.e. high diversity, differentiation, or both).

Following on from the latter methods, Caballero & Toro (2002) proposed an approach relating coancestry within populations and genetic distance among populations to total metapopulation coancestry, and this to total genetic diversity. In this case, relative contributions to total coancestry were derived to represent the amount of redundant diversity each population shared with the others, and, in turn, the amount they contributed to global metapopulation diversity. Priorities were then assigned to the populations with minor quotas of shared diversity. Interestingly, such an approach allowed also to derive the theoretical genetic dividend that populations could provide for optimizing diversity in a hypothetical germplasm bank. The method was first proposed to evaluate priorities among domestic breeds, but would be valid in the case of wild metapopulations.

Weitzman's limitation could also be addressed using Eding *et al.* (2002) '*core set*' approach, where total genetic diversity is defined as the maximal genetic variance obtainable in a hypothetical random mating population derived from the studied populations. The core set represents the smallest subset of populations optimizing total diversity, and it is identifiable by selecting the populations with the lowest mean kinship coefficient among the individuals. Once established, relative contributions can be assessed analogous to the previous methods, and priorities set accordingly. The approach was introduced in the context of domestic prioritization, but could also work for conserving genetic variability in natural metapopulations, where—at least for certain species—the assumption about the random mating among populations might appear more realistic. Weitzman and the 'core set'

approaches have been compared in the case of cattle breed prioritization, and have generally been found to produce different ranking in the populations to be prioritized (Tapio *et al.* 2006).

Until this point in the development of the field, neutral diversity—the component of genetic diversity shaped by recombination, genetic drift and gene flow—has constituted the implicit target for genetic conservation, being regarded as a reservoir for species evolutionary potential and reflecting important demographic events in their evolutionary history. However, the additional component of diversity, that which is directly subjected to selection and underlies patterns of local adaptation, life history and productive traits—i.e. adaptive diversity—remained substantially unaddressed. To fill this gap, some authors have devised methods to support prioritization using both typologies of genetic diversity, neutral and adaptive.

Marker-based genomic techniques represent a first option to investigate adaptive variability. By projecting conservation into the era of 'Omics' sciences (Allendorf *et al.* 2010), such approaches permit the recognition of genomic sites with atypical patterns of diversity, differentiation, or association with given selective pressures (Vitti *et al.* 2013). A 'population adaptive index' (PAI) (Bonin *et al.* 2007) has been developed, being a metric based on individual genome scans which uses the frequencies of loci under directional selection to quantify adaptive uniqueness of candidate populations for conservation measuring how distant a given population is from a hypothetical, pooled population with averaged frequencies at the adaptive loci. The PAI calculation was incorporated into an approach maximizing protection of total genetic diversity, given a constraint in the number of populations granted for conservation. Selected loci were highlighted on the basis of single-locus F_{ST} exceeding a theoretical neutral threshold in pairwise comparisons between populations. Therefore, neutral

and adaptive diversities were estimated for each population, the former relying on true neutral loci, the latter on the subset of selected loci, and conservation outputs were compared between competing prioritization strategies. PAI was first developed for evaluating adaptive diversity in wild populations of amphibians and plants, even if it might be generalized to populations of agricultural interest. Surprisingly, to date it has rarely been applied to either wild or domestic species.

Recently, next-generation sequencing (NGS) techniques and high density single nucleotide polymorphisms (SNP) chips allowed the characterization of an increasing number of livestock and natural species, by greatly enhancing possibilities in detecting adaptive loci. Funk *et al.* (2012) devised a pioneering pipeline exploiting this vast amount of information to define groups of populations to be considered discrete for management (i.e. conservation units, CUs), delineate adaptive groups, and support prioritization. The authors suggested to: (i) compute locus-specific global F_{ST} to individuate adaptive outlier loci; (ii) delimit evolutionarily significant units (ESUs) and management units (MUs) by relying on the entire set and the subset of neutral loci, respectively; they justified this choice by arguing that ESUs are the broadest kind of CUs, defined by both neutral and adaptive processes, whereas MUs are groups of demographically independent populations whose definition is likely to be reflected by diversity patterns at neural loci (Lowe & Allendorf 2010); (iii) use the subset of adaptive loci to delimit adaptive groups among MUs, and accordingly set priorities encompassing the adaptive differentiation within the species.

Adaptive diversity has been traditionally approached using quantitative genetic methods. Provided a set of populations have been recorded for a trait, Wellmann *et al.* (2014) devised a novel approach for estimating total and neutral trait diversities, and derive trait adaptive

diversity—i.e. the portion of total diversity not explained by neutral diversity alone—as the difference between these estimates. The approach is extendable to multiple traits to obtain an overall estimate of adaptive diversity. Thus, these authors introduced the concept of '*adaptivity coverage*' to express the capacity of a set of populations to adapt to a series of diversified environments in a short time span, and suggested the computation of population-specific conservation values to quantify the proportion of diversity (or adaptive coverage) that would go lost in case of extinction of the concerned group.

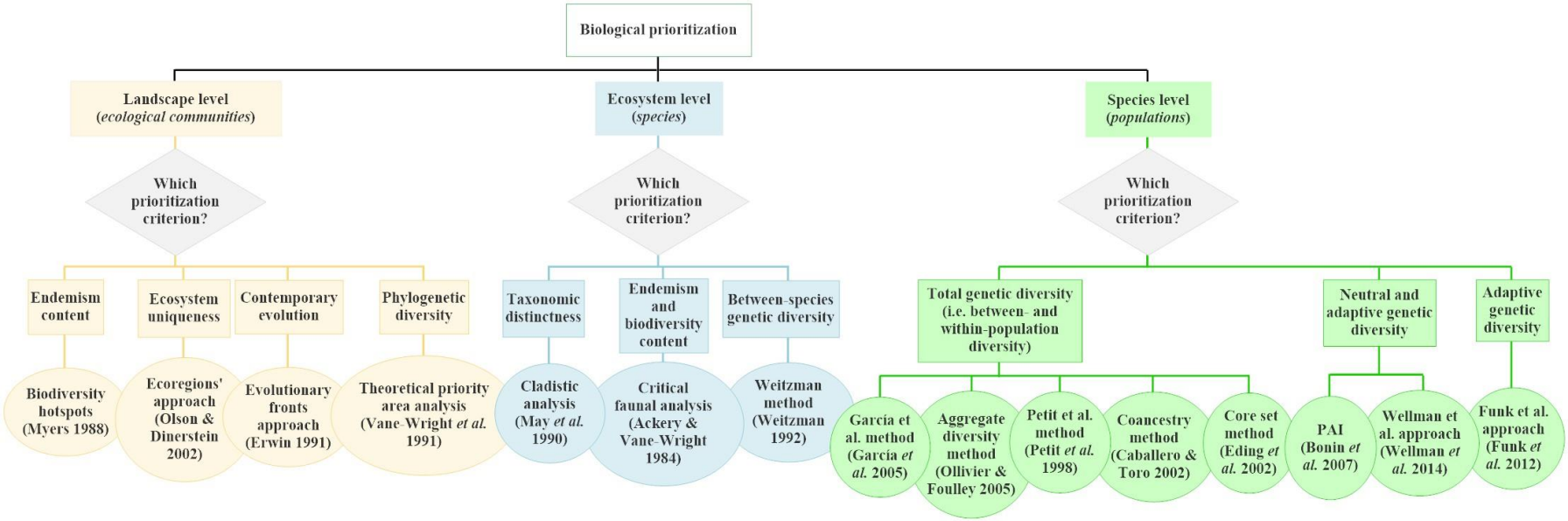

Figure 2.2 Decision tree for the reviewed direct biological prioritization methods. Colour key follows figure 2.1: orange designates criteria and methods addressing landscape level; blue refers to ecosystem level, and green to species level. Tree tips (circular boxes) correspond to the reviewed methodologies, each of which is identified on the basis of the addressed prioritization problem, the targeted level in biodiversity hierarchy and the precise prioritization criterion according to which biological priorities are assigned.

Table 2.1 Direct biological prioritization methods discussed in this review.

Method	Level[a]	Criterion[b]	Aim	Origin[c]	General[d]	Applied[e]	Notes[f]	References
Biodiversity hotspots	Landscape	Endemism content	Protection of communities reach in endemic species	W	Yes	No	Prioritization of areas rich in indigenous breeds	Myers (1988); Commission on Genetic Resources for Food and Agriculture (2012);
Ecoregions' approach	Landscape	Ecosystem uniqueness	Protection of different ecosystem types	W	No	-	-	Olson & Dinerstein (2002)
Evolutionary fronts approach	Landscape	Contemporary evolution	Protection of evolving lineages	W	No	-	-	Erwin (1991)
Theoretical priority area analysis	Landscape	Phylogenetic diversity	Protection of areas optimizing phylogenetic diversity	W	Yes	No	Prioritization of areas optimizing taxonomic diversity of the analysed set of breeds	Vane-Wright *et al.* (1991)
Cladistic analysis	Ecosystem	Taxonomic distinctness	Protection of taxonomic distinctness	W	Yes	No	Prioritization of breeds contributing more to total taxonomic diversity	May *et al.* (1990); Vane-Wright *et al.* (1991)
Critical faunal analysis	Ecosystem	Endemism and biodiversity content	Protection of target species	W	Yes	No	Prioritization of areas guaranteeing the protection of the whole set of considered breeds	Ackery & Vane-Wright (1984)
Weitzman method	Ecosystem	Between-species genetic diversity[g]	Protection of species maximizing total between-species genetic diversity	W	Yes	Yes	Application almost restricted to the sole domestic community	Weitzman (1992, 1993)
García *et al.* method	Species	Between- and within-population diversity	Protection of populations maximizing total genetic diversity	L	Yes	No	Application of the same methodology in the case of natural populations	García *et al.* (2005)
Aggregate diversity method	Species	Between- and within-population diversity	Protection of populations maximizing total genetic diversity, or total between- or within-population	L	Yes	No	Application of the same methodology in the case of natural populations	Ollivier & Foulley (2005)

			components					
Petit *et al.* method	Species	Between- and within-population diversity	Protection of populations maximizing total genetic diversity, by representing their 'diversity' and 'differentiation' contributions	L	Yes	No	Application of the same methodology in the case of domestic populations	Petit *et al.* (1998)
Coancestry method	Species	Between- and within-population diversity	Protection of populations maximizing total genetic diversity	L	Yes	No	Application of the same methodology in the case of natural populations	Caballero & Toro (2002)
Core set method	Species	Between- and within-population diversity	Protection of populations maximizing total genetic diversity	L	Yes	No	Application of the same methodology in the case of natural populations	Eding *et al.* (2002)
Population adaptive index	Species	Neutral and adaptive genetic diversity	Protection of populations maximizing neutral diversity and adaptive uniqueness	L	Yes	No	Application of the same methodology in the case of domestic populations	Bonin *et al.* (2007)
Funk *et al.* approach	Species	Adaptive genetic diversity	Protection of MUs optimizing the amount of within-species adaptive variability	L	Yes	No	Application of the same methodology in the case of domestic populations	Funk *et al.* (2012)
Wellman *et al.* approach	Species	Neutral and adaptive genetic diversity	Protection of populations maximizing adaptive potential to various environmental conditions	L	Yes	No	Application of the same methodology in the case of natural populations	Wellman *et al.* (2014)

a: targeted level in the biodiversity hierarchy: landscape (when prioritization is among different ecosystems, and thus ecological communities); ecosystem (when it is among different species, not necessarily belonging to the same ecosystem); or species (when it is among populations of the same species, often involving genetic data). **b**: criterion used for prioritization. **c**: whether the method was firstly proposed in the wild (W) or livestock (L) conservation community. The classification derives either from the case study in which the method was originally applied or from the scientific sector of the journal where it was presented. **d**: is the method theoretically general? **e**: are there any examples of its application in the other (i.e. different from the sector of origin) conservation sector? **f**: general notes. When no examples of generalization exist, notes can regard possible hints about how to expand applicability into the corresponding conservation sector. **g**: Weitzman method is suitable for quantifying any kind of between-species (or taxa) diversity. For sake of simplicity, however, we refer here to between-species genetic diversity as the method has been applied almost uniquely with genetic distances.

2.4.2 Indirect biological prioritization methods

Several methodologies developed in the fields of ecology, statistics and genetics can be adapted to identify biological priorities for conservation (Figure 2.3, Table 2.2). α, β and γ similarity measures were introduced to quantify and compare biodiversity within and between different geographical regions (Jaccard 1912; Simpson 1943; Sørensen 1948; Baselga 2010), and may serve to reveal areas of conservation concern. Considering a series of sampled sites, α-diversity estimates the average richness in species composition over all sites, γ-diversity the total regional diversity, and β-diversity, being the ratio between γ and α (Whittaker 1960, 1972), the number of effective ecological communities among the sampled assemblages (Grieves 2015): the higher this value, the higher the number of distinct ecological communities within the region. Estimation of species richness in local assemblages and similarity measures might represent an indirect way to set conservation priorities within single and multiple geographical regions. To this aim, β-diversity has been used for delimiting '*biogeographic crossroads*' (Spector 2002), ecotonal zones where transient environmental conditions support the coexistence of diversified communities, high species richness, and active evolutionary processes. When comparing different regions, further arguments for priority setting might derive from the estimation of nestedness and spatial turnover components of β-diversity, namely the degree of redundancy and species replacement between sites of the same region (Baselga 2010; Baselga & Orme 2012). No parallelism seems to exist between biogeographic crossroads and some analogous method for prioritizing agricultural landscapes. Given an opportune definition of the geographical scale for comparisons, however, β-diversity might appear appropriate to compare regional breed richness, and

identify critical areas for conservation.

Macroecological modelling (Mokany *et al.* 2014) might represent an alternative to diversity measures for defining priority areas at the landscape level. By relying on environmental predictors, correlative models are built to foresee regional species richness, compositional dissimilarity and community composition, so that to individuate unsampled areas of potential high conservation concern.

At the ecosystem and species levels, the biological prioritization problem might be addressed using *ecological niche modelling*. Ecological niche models (ENMs) (sometimes referred to as species distribution models, SDMs) are correlative techniques exploring associations between species spatial occurrences and environmental features at the sampled sites (Elith & Leathwick 2009; Thuiller *et al.* 2009), and returning probabilistic estimates of species potential distributions (Guisan & Thuiller 2005). ENMs have been employed to propose CANs for safeguarding threatened species (Urbina-Cardona & Flores-Villela 2010), to investigate the impact of climate change on communities composition (Peterson *et al.* 2002; Midgley *et al.* 2003) and to extrapolate species potential distributions in the future, by driving attention towards critical predicted shifts (Elith *et al.* 2010). In that regard, Razgour *et al.* (submitted) recently combined ENMs extrapolations with data concerning current adaptive patterns to climate and environmental heterogeneity to produce a priority rank for a set of bat populations and suggest strategies for their adaptive management. ENMs are commonly used to infer potential distributions of wild flora and fauna, being rather ignored by livestock conservation community (but see Robinson et al. 2014). However, the introduction of breed distribution models might represent a useful tool for prioritizing agricultural biodiversity at the species level, especially if evaluation of environmental risk were complemented with genetic,

demographic, economic and conservation status information.

Multivariate analysis can provide several indirect BPMs. Given conservation-relevant variables, *principal component analysis* (PCA) may be used to summarize information and rank species or populations on the basis of their principal components scores (Boettcher *et al.* 2010). When performed on genetic data, PCA can represent genetic relationships between species, genetic structure among putative populations, and highlight uniqueness to be investigated afterwards (Jombart *et al.* 2009). If samples are both genotyped and georeferenced, *spatial analysis of principal components* (sPCA) may figure out genetic relationships between populations by accounting for the effect of hidden spatial structures (Jombart *et al.* 2008). sPCA defines linear combinations of allele frequencies (or genotypes) optimizing the product between the overall genetic variance and spatial genetic autocorrelation, so that fine spatial genetic patterns can be uncovered, and hypotheses can be tested about global and local structures—i.e. the existence of clines and clusters, or marked differences between neighbours. In fact, sPCA has been shown to reveal genetic signatures and spatial structuring which would have remained otherwise unnoticed (Laloë *et al.* 2010). Just like PCA, it can be exploited to target attention towards natural or livestock populations of major conservation concern.

The vast array of mathematical techniques performing *population viability analysis* (PVA) constitutes a notable tools for alerting about the conservation status of species or populations. PVA relies on demographic, life history and sometimes genetic information to estimate the minimum viable population (MVP) size of the concerned taxa, assess their likelihood to decline below such a demographic threshold at some time point in the future, and suggest if they are threaten by extinction or not (estimated census below or above MVP size,

respectively) (Morris & Doak 2002; Traill *et al.* 2007). After the pioneering study by Shaffer (1978), these techniques were extended to evaluate the extinction risk of both natural (Bakker *et al.* 2009; Tian *et al.* 2011) and livestock populations (Bennewitz & Meuwissen 2005), identify drivers of census decline, and test the effectiveness of competing management actions (Sebastián-González *et al.* 2011). PVA implicitly offers the possibility of targeting conservation efforts towards sensitive taxa, including those with realistic recovery possibilities and those most threatened by extinction. However, such criteria should to taken into account with extreme caution: although PVA predictive accuracy was proved to be good in the presence of extensive and informative data (Brook *et al.* 2000), some serious concerns remain about its reliability with insufficient information, as well as its ability in modelling unpredictable catastrophic events and future vital rates (Coulson *et al.* 2001). Unfortunately, real-life conservation studies often clash with these limitations, by making PVA an elegant, useful but often uncertain method for prioritizing species or populations for conservation.

With the aim of defining MUs among harbor seal populations, Olsen *et al.* (2014) proposed an integrated approach coupling genetic information with life history and demographic data. Genetic units were (i) delineated using molecular markers, (ii) tested for demographic independence comparing their census and MVP sizes, and (iii) considered actual MUs whenever census exceeded MVP size threshold. Following this rationale, priorities may then be accorded to natural or domestic genetic units which are threatened by extinction because of demographic dependence on other populations.

Q_{ST}–F_{ST} analysis (Leinonen *et al.* 2013) may be used to investigate adaptive divergence and indirectly suggest priorities at the species level. Q_{ST} is a measure of genetic differentiation between populations similar to F_{ST} but estimating the degree of divergence in quantitative

traits instead of physical loci (Spitze 1993). Provided a measured quantitative trait of interest and a set of true neutral loci, Q_{ST} and F_{ST} can be computed. F_{ST} provides a reference value to test if observed divergence in the quantitative trait evolved by genetic drift ($Q_{ST}=F_{ST}$), because of directional selection ($Q_{ST}>F_{ST}$), or because of stabilizing selection ($Q_{ST}<F_{ST}$). In practice, the analysis enables a user to detect genetic differentiation between natural populations attributable to directional selection (Sæther *et al.* 2007; Leinonen *et al.* 2013), but to our knowledge has never been proposed to directly set priorities for conservation. To this end, pairwise comparisons between populations would probably be useful, by permitting to identify populations where directional selection is taking place and different adaptive solutions have evolved. Similar to the core set approach, this would ideally define a group of populations encompassing the largest amount of adaptive variability related to the traits under study, and thus deserving conservation priority. Such a framework based on Q_{ST}–F_{ST} analysis might be considered for both wild and agricultural species.

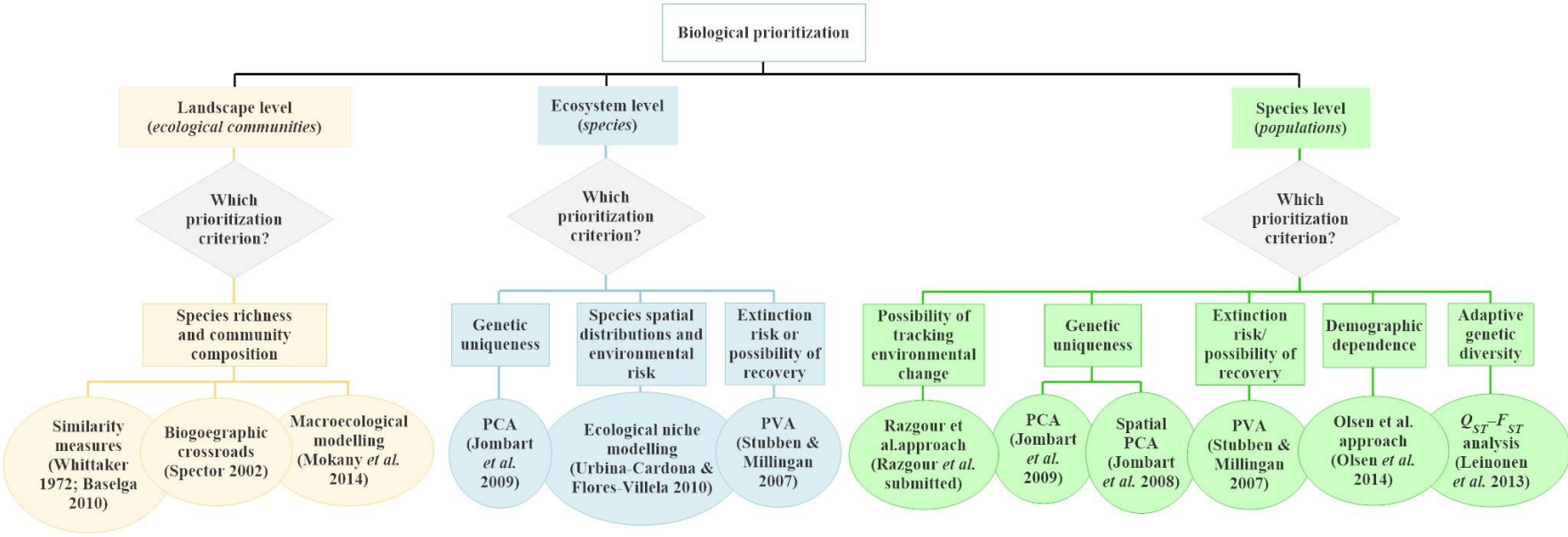

Figure 2.3 Decision tree for the reviewed indirect biological prioritization methods. Colour key follows figure 2.1: orange designates criteria and methods addressing landscape level; blue refers to ecosystem level, and green to species level. Tree tips (circular boxes) correspond to the reviewed methodologies, each of which is identified following the decision path described in section 2.3.2.

Table 2.2 Examples of indirect biological prioritization methods discussed in this review[a].

Method	Level	Criterion	Aim	Origin	General	Applied	Notes	Free software[b]	References
Similarity measures	Landscape	Species richness and community composition	Protection of regions with the highest number of ecological communities	W	Yes	No	Comparisons among regional breed richness, and prioritization of the most diversified agricultural areas	`betapart` R package (Baselga & Orme 2012)	Whittaker (1972); Baselga (2010)
Biogeographic crossroads	Landscape	Species richness and community composition	Protection of regions with diversified communities, high species richness, and active evolutionary processes	W	No	-	-	`betapart` R package (Baselga & Orme 2012)	Spector (2002)
Macroecological modelling	Landscape	Species richness and community composition	Protection of the most diversified regions (in terms of species richness, and community composition)	W	No	-	-	-	Mokany *et al.* (2014)
Principal component analysis	Ecosystem Species	Genetic uniqueness[c]	Representation of genetic structure and individuation of genetic singularities	-	Yes	Yes	-	`adegenet` R package (Jombart 2008; Jombart & Ahmed 2011)	Jombart *et al.* (2009)
Ecological niche modelling	Ecosystem	Species spatial distributions and environmental risk	Proposal of CANs[d] and estimation of the expected shifts in optimal habitats because of environmental change	W	Yes	No	Description of breed potential distributions, and prioritization of breeds whose current niche is expected to shift because of environmental and socio-economic change	`biomod2` (Thuiller *et al.* 2016) and `KISSMig` (Nobis & Normand 2014) R packages; QGIS (QGIS Development Team 2016); ZONATION (Moilanen *et al.* 2005)	Urbina-Cardona & Flores-Villela (2010)
Population viability	Ecosystem	Extinction risk or	Protection of taxa	W	Yes	Yes	-	`popbio` R package	Bennewitz &

analysis (PVA)	Species	possibility of recovery	threatened by extinction (or with realistic recovery chances), as well as identification of effective management strategies					(Stubben *et al.* 2007)	Meuwissen (2005)
Razgour *et al.* approach	Species	Possibility of tackling environmental change	Protection of locally adapted populations which are unable to track optimal habitat shift	W	Yes	No	Prioritization of locally adapted breeds whose optimal habitat is expected to shift because of environmental, and socio-economic change	`biomod2` R package (Thuiller *et al.* 2016); Spatial analysis method (SAM) and SAMβADA (Joost *et al.* 2007; Stucki *et al.* 2016); `LEA` R package (Frichot & François 2015)	Razgour *et al.* (submitted)
Spatial principal component analysis	Species	Genetic uniqueness	Representation of genetic and spatial structuring and individuation of genetic singularities	W	Yes	Yes	-	`adegenet` R package (Jombart 2008; Jombart & Ahmed 2011)	Jombart *et al.* (2008)
Olsen *et al.* approach	Species	Demographic dependence	Protection of demographically dependent genetic units	W	Yes	No	Application of the same methodology in the case of domestic populations	-	Olsen *et al.* (2014)
Q_{ST}–F_{ST} analysis	Species	Adaptive genetic diversity	Protection of populations maximizing the amount of adaptive variability under study	W	Yes	No	Application of the same methodology in the case of domestic populations	-	Leinonen *et al.* (2013)

a: refer to Table 2.1 footnotes for an explanation of column headings. **b**: free software implementing the concerned method. **c**: see text for alternative uses of principal component analysis in setting conservation priorities. For a general use of the technique, refer to the R functions `prcomp` or `princomp` of `stats` package (R Core Team 2015). **d**: Conservation Area Networks.

2.5 The conservation resources allocation problem

Wilson *et al.* (2006) framed the conservation resource allocation problem into a decision support science context (Figure 2.4, Table 2.3). Given a predefined set of priority areas and a fixed budget, the goal was to maximize biodiversity protection through the definition of an optimal CAN. Heuristic algorithms were proposed to identify optimal solutions about where, how much and when conservation funding should be allocated. Strategies were formulated by accounting for conservation costs, regional threats to biodiversity and regional value in biodiversity (e.g. numbers of endemic bird species), and evaluated on the basis of investment return (the amount of biodiversity protected). Management guidelines were then formulated for different situations: when candidate regions presented similar levels of endemism but different levels of threat, the best resource allocation strategy was to minimize short-term biodiversity loss; and if uncertainty existed about funding and the candidate regions experienced similar threat levels, maximization of short-term gains in biodiversity protection turned out to be the best decision.

More recently, Joseph *et al.* (2009) devised a cost-benefit analysis to efficiently allocate resources among species conservation projects. Project prioritization protocols based on different criteria were evaluated for their ability in optimizing the number of funded projects. They found that protocols explicitly stating conservation costs and probability of success proved to protect more species than protocols based only on species value or threat status.

Similarly, a cost-efficiency analysis was developed to prioritize habitat-management actions optimizing protection of target species, given budget constraints (Sebastián-González *et al.* 2011). First, actions were prioritized on the basis of the expected increase in target species

abundance, and second, expected achievements were validated by means of PVAs performed on a subset of well-characterized target species. Formal approaches based on decision science and allocating resources among conservation strategies, projects or actions, have proved to outperform traditional biological prioritization in optimizing biodiversity protection (Marris 2007).

If prioritization criterion is to maximize among-taxa diversity, the Weitzman framework can again provide a basis upon which to formulate optimal funding strategies. By considering extinction probabilities to be mainly governed by effective population sizes (N_e), Simianer *et al.* (2003) introduced explicit relationships describing the direct effects of funding allocation on N_e. Given a fixed budget, several functions were developed to describe with more realism the management of domestic populations. Funding-driven changes in N_e and extinction probabilities were related to marginal diversities in order to describe the predicted effects on total between-breed diversity, and formulate optimal resource allocation strategies. The future development of specific functions describing plausible impacts of resource allocation on extinction probabilities in wildlife would also enable to generalize the method to the case of natural species or populations.

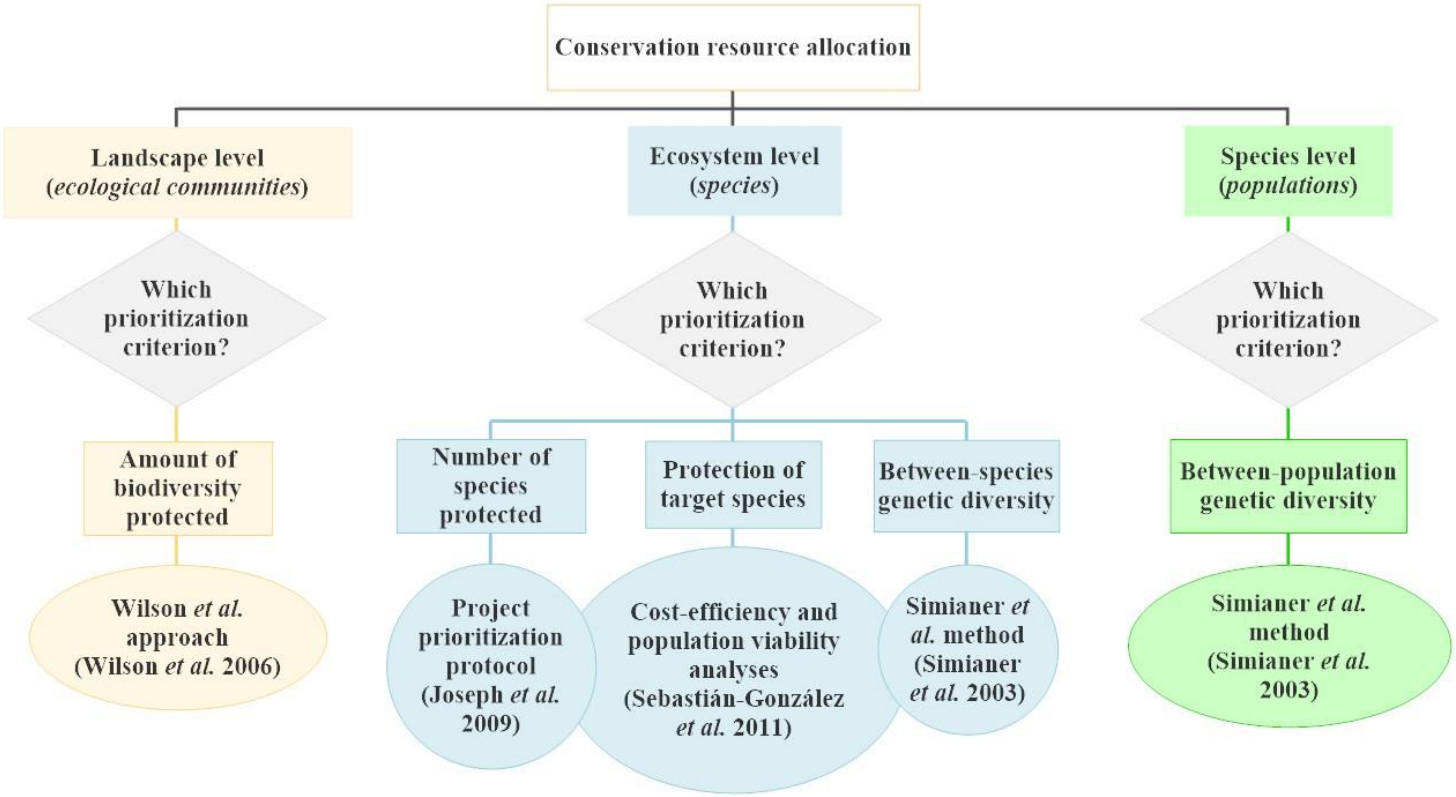

Figure 2.4 Decision tree for the reviewed resource allocation methods. Colour key follows figure 2.1: orange designates criteria and methods addressing landscape level; blue refers to ecosystem level, and green to species level. Tree tips (circular boxes) correspond to the reviewed methodologies, each of which is identified following the decision path described in section 2.3.2.

Table 2.3 Examples of resource allocation methods discussed in the present review[a].

Method	Level	Criterion	Aim	Origin	General	Applied	Notes	Free software	References
Wilson *et al.* approach	Landscape	Amount of biodiversity protected	Definition of optimal CANs[b] to protect biodivesity	W	Yes	No	Given the prior individuation of critical agricultural areas (see notes for similarity measures in Table 2.2), the approach might be applied to reveal optimal spatial strategies maximizing investment return (i.e. the amount of protected breeds or strains).	ZONATION (Moilanen *et al.* 2005)	Wilson *et al.* (2006)
Project prioritization protocol	Ecosystem	Number of species protected	Optimal resource allocation to maximize the number of funded projects (i.e. protected species)	W	Yes	No	Might be suitable for devising project prioritization protocols for breeds or strains	-	Joseph *et al.* (2009)
Cost-efficiency and population viability analysis	Ecosystem	Protection of target species	Optimal resource allocation among actions to maximize protection of some target species	W	Yes	No	Application of the same methodology in the case of domestic populations	-	Sebastián-González *et al.* (2011)
Simianer *et al.* method	Ecosystem Species	Between-species (or population) genetic diversity	Optimal resource allocation to maximize between-species (or population) genetic diversity	L	Yes	No	Development of ad hoc functional relationships describing the effects of resource allocation on extinction probabilities of wild species (or populations)	-	Simianer *et al.* (2003)

a: Refer to Table 2.1 and Table 2.2 footnotes for an explanation of column headings. **b**:Conservation Area Networks.

2.6 Discussion

A rough search in Google Scholar with the keywords ‘prioritization’ and ‘conservation biology’ returns around 9,000 results. This amount of literature makes the attempt of drawing a general picture rather difficult. In the present review, approximately thirty methods have been analysed, and some have certainly been disregarded. However, the analysed literature permitted a global appraisal of priority setting in conservation by highlighting conceptual and methodological trends A classification and decision-aid scheme was proposed (Figure 2.1), first subdividing methods into two broad categories (BPMs and RAMs), and subsequently referring them to a targeted biodiversity level (landscape, ecosystem or species). The scheme is expected to remain valid also for methods not discussed in this paper: for instance, Carwardine *et al.* (2008), Moilanen *et al.* (2008), and Volkmann *et al.* (2014) would fall into direct BPMs at the species level, while Reist-Marti *et al.* (2006) or Carwardine *et al.* (2008), Moilanen *et al.* (2008) and Volkmann *et al.* (2014) into RAMs at the landscape level.

The examination of techniques described in wild and livestock literatures suggested that generalizations could be possible in about 70% of the cases. Typically, approaches developed in the wildlife community may be adapted to focus on domestic animal populations, where diversity within species is the actual target for agricultural conservation (Table 2.1, 2.2 and 2.3). Spatial methods might prove useful in (i) highlighting areas with high strain richness (Whittaker 1972, Baselga 2010) or indigenous breeds concentration (Myers 1988), (ii) maximizing protection of breed diversity (Ackery & Vane-Wright 1984; Vane-Wright *et al.* 1991), (iii) revealing locally adapted breeds threatened by shifting niches (Razgour et al.

submitted), and (iv) defining optimal resource allocation (Wilson *et al.* 2006). Phylogenies may also be inferred and breeds prioritized on the basis of taxonomic distinctness (May *et al.* 1990). An even more straightforward transposition would be possible for genetic methods, since genetic fragmentation is threatening wild and livestock within-species diversity similarly (Taberlet *et al.* 2008). Again, no evidence of such a methodological exchange appears from the reviewed literature: integrations to Weitzman method (García *et al.* 2005; Ollivier & Foulley 2005) and alternative options addressing between- and within-populations neutral genetic diversity (Petit *et al.* 1998; Caballero & Toro 2002; Eding *et al.* 2002) seems confined to the sector of origin. The same applies for methods accounting for adaptive diversity (Bonin *et al.* 2007; Funk *et al.* 2012; Leinonen *et al.* 2013; Wellman *et al.* 2014).

Complementary approaches might be evidenced and integrated to enhance prioritization capacity in both wildlife and livestock conservation. Funk *et al.*'s approach appears directly applicable for delineating CUs in the wild species, but of more difficult application in domestics, where non-neutral genomic regions are shaped by both natural and anthropogenic selection and the global F_{ST} method might also identify not truly adaptive signals. However, particularly in the case of populations living under a "natural" regime (e.g. livestock kept under traditional extensive management systems), global F_{ST} method might remain valid to identify neutral loci to be used in the delineation of MUs, while an environmental association analysis (Rellstab *et al.* 2015) might be subsequently employed to identify putative adaptive loci underlying a selective pressure of interest (e.g. adaptation to climate or diseases). In this way, highlighted loci might then be used to identify adaptive groups within (or across) MUs and biological priorities as devised by Funk and colleagues.

The frameworks proposed by Funk and Olsen (Funk *et al.* 2012; Olsen *et al.* 2014) might also be combined to provide a genomic method integrating demographic information and addressing biological prioritization within wild and livestock species. In particular, ESUs and genetic units might be delimited using total and neutral loci, respectively. A population viability analysis may then be performed to test genetic units for demographic independence. In this way, MUs and demographically endangered units would be evidenced, and adaptive loci used to outline adaptive clusters. Prioritization would finally target endangered units, which might be supplemented by the most adaptively similar MUs to decrease chances of outbreeding depression (Funk *et al.* 2012). Such a combination, therefore, would increase our capacity of outlining CUs and targeting adaptive management towards effectively declining genetic units.

Applications of genome-editing techniques have been recently suggested as a tool to address conservation-relevant issues (Taylor & Gemmell 2016). In fact, the ability of deleting, inserting and replacing specific sites in individual genomes is opening new prospects for the genetic biocontrol of invasive species, the management of bottlenecked populations (e.g. by directly removing genetic disorders or supplementing diversity in target genomic regions) and the reshaping of endangered species habitat requirements (Johnson *et al.* 2016). In such a context, biological prioritization represents the preliminary step for delimiting CUs to subsequently target by genome editing. For instance, Creole cattle breeds from Latin America are receiving considerable attention for conservation because of their high degree of genetic diversity and peculiar natural adaptations to tropical environments like the SLICK mutation affecting hair phenotype and conferring tolerance to high temperatures (Ginja *et al.* 2013).

Recently, the SLICK variant has been identified (Huson *et al.* 2014; Littlejohn *et al.* 2014) and introduced by genome-editing methodologies into the Holstein genome, thus leading to positive results in terms of decreased heat-stress and improved production performances during the hot season (Dikmen *et al.* 2014). Although promising, however, extensive usage of genome-editing should be carefully evaluated in conservation due to serious ethical concerns and gaps in knowledge, especially regarding potential side effects like horizontal gene transfer or unwanted alterations of genomic processes in the natural context (Webber *et al*. 2015).

To conclude, the present review focused on similarities—rather than differences—among approaches proposed for wild and agricultural biodiversities. Formal proof of the suggested generalizations and integrations was beyond our scope, and future research will be required to test their effectiveness. Given the potential for generalization that emerged from our investigation, however, we believe that a more extensive communication and reciprocal scientific exchange between the wildlife and livestock sector would be desirable to achieve the common goal of optimizing biodiversity conservation.

3. Water buffalo genomic diversity and post-domestication migration routes

Licia Colli*, Marco Milanesi*, Elia Vajana*, Daniela Iamartino, Lorenzo Bomba, Francesco Puglisi, Marcello Del Corvo, Paolo Ajmone Marsan, and the International Buffalo Consortium

*Equally contributing authors

3.1 Abstract

The 90K Affymetrix Axiom® Buffalo Genotyping Array has been used to genotype river buffalo samples from Pakistan, Iran, Turkey, Egypt, Romania, Bulgaria, Italy, Mozambique, Brazil and Colombia, and swamp buffaloes from China, Thailand, Philippines, Indonesia and Brazil. Model-based clustering algorithms and phylogenetic tools have been applied to estimate the levels of molecular diversity and population structure, and infer migration events. In agreement with documented importations of animals for breed improvement purposes, three distinct gene pools in pure river as well as in pure swamp buffalo populations were highlighted, together with some genomic admixture occurring in the Philippines and in Brazil. The Mediterranean from Italy and the Carabao from Brazil represent the most differentiated gene pools within the river and swamp group, respectively, which is most likely due to genetic bottlenecks, isolation and selection. Inferred gene flow events highlighted a possible contribution from the river buffalo gene pool to the admixed swamp populations and, within river-type buffaloes, from the Mediterranean to the Colombian and Brazilian breeds. Furthermore, our results support archeozoological evidence for the domestication of the river

buffalo in the Indian subcontinent, and of the swamp type buffalo in Southeast Asia, while suggesting some unexpected migration routes out of the proposed domestication centres.

Keywords: Water buffalo, river buffalo, swamp buffalo, *Bubalus bubalis*, SNP, genomic diversity

3.2 Introduction

The domestic water buffalo *Bubalus bubalis* (Linnaeus, 1758) is native to the Asian continent but through historical migration events and recent importations, it reached a worldwide distribution during the last century (Cockrill 1974). It represents the most important farm animal resource in several highly populated developing countries of the tropical and subtropical region, and contributes largely to the local economy of rural areas and tribal communities (Mishra *et al.* 2015). As a source of milk, meat, dung, hide, horns and traction power, the water buffalo is estimated to provide livelihood to a larger number of people than any other livestock species (Scherf 2000). Two types of water buffalo are traditionally recognised, the river and the swamp buffalo (Macgregor 1941), respectively assigned to different subspecies, *Bubalus bubalis bubalis* and *Bubalus bubalis carabanensis*. Besides displaying distinct morphological, cytogenetic (chromosome number: river 2*n*=50, swamp 2*n*=48) and behavioural traits, they also have different purposes and geographical distributions: the river buffalo is mainly a dairy animal with several recognized breeds, spread from the Indian subcontinent to the eastern Mediterranean countries (the Balkans, Italy and Egypt) and imported to Indonesia, southern America and central Africa during the XXth

century. The swamp buffalo has no recognized breeds and is primarily used for draught power in a wide area ranging from eastern India (Assam region), through south-eastern Asia, Indonesia to eastern China (Yangtze river valley) (Zhang *et al.* 2016), and was recently introduced (XX° cen.) into Australia and southern America.

Being interfertile, the two types naturally interbreed in the area of geographical overlap located between north-east India and south-east Asia (Mishra *et al.* 2015), but in several countries they have been intentionally crossed to increase the productivity of swamp buffaloes (Borghese 2011).

Even if the wild buffalo *Bubalus arnee* is generally accepted as the probable ancestor of the water buffalo, the details of the domestication dynamics have been debated for a long time, with the two major hypotheses envisaging either a single (Kierstein *et al.* 2004) or two independent events for river and swamp types (Lau *et al.* 1998; Ritz *et al.* 2000; Kumar *et al.* 2007a; 2007b; Lei *et al.* 2007; Yindee *et al.* 2010; Zhang *et al.* 2016). With the lack of conclusive archeozoological data, a growing body of molecular evidence, based on the analysis of mitochondrial (Lau *et al.* 1998; Kumar *et al.* 2007a; 2007b; Lei *et al.* 2007), Y chromosome (Yindee *et al.* 2010; Zhang *et al.* 2016) and autosomal DNA (Ritz *et al.* 2000), seem to support the scenario of two independent domestication events that have involved wild ancestor populations that had long since diverged.

The same evidence also suggests north-western India as most likely domestication centre for river buffaloes (Nagarajan *et al.* 2015) and the region close to the border between China and Indochina for swamp buffaloes (Zhang *et al.* 2011, 2016). From their respective domestication centres, river buffaloes migrated west across south-western Asia, to Egypt, Anatolia and

reached the Balkans and the Italian peninsula in the early Middle ages (VII° cen. AD; (Clutton-Brock 1999), while the swamp buffaloes likely dispersed Southwestwardly to Thailand and Indonesia, and northward to central and eastern China (Zhang *et al.* 2016), wherefrom they further spread to the Philippines (Zhang *et al.* 2011).

Several studies have relied on nuclear microsatellite markers to describe the levels and the distribution of molecular diversity in water buffalo populations from different countries (Moioli *et al.* 2001; El-Kholy *et al.* 2007; Zhang *et al.* 2011; Saif *et al.* 2012; Ünal *et al.* 2014). However, so far it has not been possible to obtain a comprehensive view of the molecular variation of the species across its distribution area due to the adoption of different or only partially overlapping marker panels.

In the last decades, the demographic trends of a number of water buffalo populations have shown a steady contraction in population sizes (Borghese 2011), which usually brings along an increased risk of loss of biodiversity. An effective evaluation of the genomic "health status" of livestock breeds and populations is a basic prerequisite for the definition of adequate plans to safeguard and/or restore diversity, and also to identify demographic discontinuities with detrimental effects, such as a lack of gene flow, excessive inbreeding or indiscriminate crossbreeding. In recent years, standardized marker panels as medium or high density SNP chips have become available for the major livestock species and have proven particularly useful to analyse farm animals genomic variability both at the global (Kijas *et al.* 2012; Decker *et al.* 2014) and at the local level (Nicoloso *et al.* 2015), and to shed light on their post-domestication evolutionary history.

The attempts made to characterize water buffaloes via cattle-specific high- (Borquis *et al.*

2014) and medium-density SNP panels (Michelizzi *et al.* 2011) returned either very low percentages of polymorphic markers (2.2%; Michelizzi *et al.* 2011), or high numbers of markers with very low level of polymorphism (about 650K markers out of 800K had Minor Allele Frequency <0.05; Borquis *et al.* 2014), or very low values of the individual genotype call rates (0.54-0.90, mean value 0.85, compared to the >0.98 usually scored in cattle; Borquis *et al.* 2014).

Recently the Axiom® Buffalo Genotyping Array has been developed in collaboration with the International Buffalo Genome Consortium, and includes about 90K polymorphic SNP markers with a high genome-wide coverage (Iamartino *et al.* in preparation). The SNP discovery panel was represented mostly by river buffalo breeds (Mediterranean, Murrah, Jaffarabadi, and Nili-Ravi) but about 25% of the markers resulted to be polymorphic also when tested over a number of swamp buffalo populations.

Here we present the result of the characterization of the genomic diversity in 31 buffalo populations of river, swamp and crossbred river *x* swamp origin, covering most of the worldwide distribution of the species.

3.3 Materials and methods

3.3.1 Sampling and genotyping

The DNA samples were provided by the members of the International Water Buffalo Consortium. A total of 346 individuals were sampled from 31 populations covering a large

part of the worldwide geographical distribution of water buffalo (Figure 3.1 and Table 3.1).

Figure 3.1 Geographical origin of the sampled populations. The correspondence between numbers and populations is given in Table 3.1.

In particular, 15 river and 16 swamp buffalo breeds were targeted, together with one lowland anoa (*Bubalus depressicornis*) population. River and swamp buffalo samples were collected from India, Pakistan, Iran, Turkey, Egypt, Italy, Bulgaria, Romania, Mozambique, Colombia, Brazil and from China, Philippines, Thailand, Indonesia, Brazil, respectively.

After testing DNA quality and concentration on 1.5% agarose gel, all samples have been genotyped with the Axiom® Buffalo Genotyping Array 90K from Affymetrix (http://www.affymetrix.com). This panel includes about 90K markers evenly distributed along the genome and provides a genome-wide coverage of polymorphic SNPs in the water buffalo species. Genotype data are available from the authors upon request.

3.3.2 Dataset construction

Since the Axiom® Buffalo SNP panel has been developed starting from a set of river-type

buffalo breeds (Iamartino *et al.* in preparation), a lower level of polymorphism was expected in swamp-type populations due to an Ascertainment Bias (AB) effect already reported by previous preliminary investigations (Iamartino *et al.* in preparation).

Thus, to reduce the impact of AB, the main dataset was built by including individuals from both river and swamp-type populations (named *poly-SW* hereunder) and only those SNP markers that were polymorphic in swamp buffalo. In order to check the effects of this strategy, we first compared the average values of observed heterozygosity obtained within this dataset to those obtained from a second version of the dataset which included all SNP markers that resulted polymorphic overall, named *poly-ALL* hereunder.

3.3.3 Quality control procedures and statistical analysis

Raw genotypic data were subjected to quality control (QC) procedures performed with the function `check.marker` of the R package `GenABEL` (Aulchenko *et al.* 2007) and the following threshold values: individual call rate <0.95, SNP call rate <0.95, threshold value for acceptable Identity By State (IBS) <0.99 (evaluated on 5000 randomly selected markers), Minor Allele Frequency (MAF) <0.01.

To evaluate the relationships between individual multilocus genotypes, Multi-dimensional Scaling (MDS) plots based on the IBS distances were obtained with the `cmdscale` function of the `stats` R package. The number of most informative dimensions was evaluated from the bar plot of the components' eigenvalues.

The software ARLEQUIN v.3.5.2.2 (Excoffier & Lischer 2010) was used to: (i) calculate

observed (H_{obs}) and expected heterozygosity (H_{exp}), subsequently corrected over the number of usable loci; (ii) compute Wright's F_{ST} fixation index (Wright 1965) and the inbreeding coefficient F_{IS} (Weir & Cockerham 1984); (iii) perform an Analysis of MOlecular VAriance (AMOVA; Excoffier *et al.* 1992); and (iv) compute a matrix of Reynolds unweighted distances (DR) between breeds (Reynolds *et al.* 1983). Starting from DR distance matrix, a neighbour-net was subsequently built with the software SPLITSTREE v.4.14.2 (Huson & Bryant 2005).

Gene flow, estimated as the number of migrants per generation exchanged between populations, was calculated with the composite-likelihood method implemented in JAATHA v.2.7.0 (Naduvilezhath *et al.* 2011; Lisha *et al.* 2013). The following parameter values were set: split time (τ) comprised within the interval [0.01-5], scaled migration rate (M) within [0.01-75], mutation parameter (θ) within [1-20], and recombination parameter equal to 20.

A model-based estimation of population structure was obtained through maximum-likelihood criterion with the software Admixture v.1.22 (Alexander *et al.* 2009) for K values from 2 to 40, under the assumptions of Hardy-Weinberg equilibrium (HWE) and complete linkage equilibrium, and with the 'unsupervised' method. To identify the best cluster solution, both 5-fold Cross-Validation errors and the number of iterations needed to reach convergence were considered for each K value.

The occurrence of migration events was evaluated with the software TREEMIX v.1.12 (Pickrell & Pritchard 2012), by including 14 lowland Anoa (*B. depressicornis*) individuals to serve as an outgroup. By relying on a drift-based evolutionary model, TREEMIX estimates the relationships occurring among the studied populations, and then models a user-defined number

of migrations (m_i) within the tree, while estimating the proportion of admixture displayed by the receiving groups. In order to avoid issues related to missing values, all marker positions displaying missing data were removed after adding the outgroup. Furthermore, to assess the robustness of the modelled migrations, the following bootstrap-based procedure was adopted: (i) a varying number of migrations was modelled up to a maximum of m=15 (m_{15}) and with a number of SNPs per block equal to 50; (ii) the most meaningful number of migrations (m_{best}) was identified based on the variance "in relatedness between populations" explained by the model (Pickrell & Pritchard 2012), the log likelihood of the model, the *p-values* associated with each migration(s), and the biological meaning of the migrations themselves; (iii) 100 bootstrap replicates of the analysis with m_{best} migrations were performed, and a consensus tree was built with the "CONSENSE" executable implemented in PHYPIP v.3.696 (Felsenstein 1989, 2016), following the majority rule; (iv) finally, the consensus tree was loaded into TREEMIX and a number of migrations equal to m_{best} was re-estimated together with the *f3*-statistics, as computed for each populations' triplet through the software THREEPOP (Reich *et al.* 2009).

3.4 Results

Nineteen individuals with low quality genotypes were dropped during QC procedures, leading to the complete removal of one Chinese population (SWACN_WEN, 3 individuals). Thus, the working version of the dataset included 20,463 SNPs, 327 individuals and 31 populations after QC. Population size ranged from 3 to 15, with an average of 10.55. Table 3.1 provides a summary of pre- and post-QC dataset statistics.

Table 3.1 Analysed anoa, river and swamp buffalo populations. String (pop. label) and number code (n.) are reported for each population with the number of samples pre (n. samples pre QC) and post QC (n. samples post QC).

Species	n.	Breed	pop. Label	Country	Region	n. samples pre QC	n. samples post QC
Lowland anoa *Bubalus depressicornis*	1	–	ANOA	Indonesia		14	14
River buffalo *Bubalus bubalis bubalis*	2	Mediterranean	RIVIT_MED	Italy		15	15
	3	Mediterranean	RIVMZ	Mozambique		7	7
	4	Mediterranean	RIVRO	Romania		13	9
	5	Murrah	RIVPH_IN_MUR	India*		6	4
	6	Murrah	RIVPH_BU_MUR	Bulgaria*		10	8
	7	Murrah	RIVBR_MUR	Brazil		15	15
	8	Anatolian	RIVTR_ANA	Turkey	Istanbul, Afyonkarahisar (western Anatolia) and Tokat (central Anatolia) Provinces	15	15
	9	Egyptian	RIVEG	Egypt		16	15
	10	Azari	RIVIR_AZA	Iran	Urmia, West Azerbaijan Province	9	9
	11	Khuzestani	RIVIR_KHU	Iran	Ahvaz, Khuzestan Province	10	10
	12	Mazandarani	RIVIR_MAZ	Iran	Miankaleh peninsula, Mazandaran Province	8	8
	13	Aza Kheli	RIVPK_AZK	Pakistan		3	3
	14	Kundhi	RIVPK_KUN	Pakistan		10	10
	15	Nili-Ravi	RIVPK_NIL	Pakistan		15	15
	16	–	RIVCO	Colombia		12	12

			total				**164**	**155**
Swamp buffalo *Bubalus carabanensis*	*bubalis*	17	–	SWAPH	Philippines		15	15
		18	–	SWAPH_ADM	Philippines		10	9
		19	Carabao	SWABR_CAR	Brazil		10	10
		20	–	SWATH_THS	Thailand		6	6
		21	–	SWATH_THT	Thailand		8	8
		22	–	SWACN_ENS	China	Enshi	15	15
		23	–	SWACN_FUL	China	Fuling	15	15
		24	–	SWACN_GUI	China	Guizhou	11	11
		25	–	SWACN_HUN	China	Hunan	15	15
		26	–	SWACN_WEN	China	Wenzhou[a]	3	-
		27	–	SWACN_YAN	China	Yangzhou	14	12
		28	–	SWACN_YIB	China	Yibin	15	15
		29	–	SWAID_JAV	Indonesia	Java	13	12
		30	–	SWAID_NUT	Indonesia	Nusa Tenggara	7	7
		31	–	SWAID_SUM	Indonesia	Sumatra	13	12
		32	–	SWAID_SUW	Indonesia	South Sulawesi	11	10
			total				**181**	**172**
Grand total							**346**	**327**

§: these numbers identify the different populations on the map in Figure 3.1; *Animals of Indian/Bulgarian origin but reared in the Philippines; [a]South-East China (Chinese coasts north of Taiwan).

The dataset version based on markers polymorphic overall contained 67,206 SNPs, 155 individuals and 31 populations.

The comparison of the observed heterozygosity values obtained with the *poly-SW* and the *poly-ALL* versions of the dataset showed that the reduction in the number of markers did not change the trend of H_{obs} values for river-type buffaloes (Supplementary 3.8.1 and 3.8.2, left panels), while swamp-type populations increased their heterozygosity of 0.15 on average (Supplementary 3.8.1 and 3.8.2, right panels). For river-type buffaloes, the values of H_{obs} and H_{exp} corrected over the number of usable loci (Table 3.2) ranged from 0.334 (RIVMZ population) to 0.417 (RIVPK_NIL population), and from 0.362 (RIVMZ) to 0.406 (RIVCO) respectively. For pure swamp-type buffaloes, the values varied between 0.334 (RIVMZ population) and 0.417 (RIVPK_NIL population), and between 0.220 (SWAID_NUT) and 0.294 (SWATH_THS) respectively. Corrected H_{obs} and H_{exp} estimates for SWAPH_ADM, a population of known river *x* swamp admixed origin, were 0.413 and 0.391, respectively.

Among water buffalo populations, F_{IS} ranged between -0.064 (SWABR_CAR) and 0.067 (SWATH_THT), and was never statistically significant ($P<0.05$) (Table 3.2). On the contrary, a statistically significant F_{IS} of 0.338 was obtained for lowland anoa.

Table 3.2 Expected and observed heterozygosity for each population together with the estimated inbreeding coefficient (F_{IS}).

Population	H_{obs}	S.D $H_{obs.}$	H_{exp}	SD $H_{exp.}$	N. usable loci	N. polymorphic loci	H_{obs} (corrected)^	H_{exp} (corrected)^	F_{IS}
ANOA	0.160	0.132	0.238	0.164	12601	2235	0.028	0.229	0.338*
RIVIT_MED	0.381	0.164	0.385	0.130	19983	18842	0.359	0.372	0.009
RIVMZ	0.411	0.204	0.390	0.136	20057	16337	0.334	0.362	-0.062
RIVRO	0.401	0.185	0.400	0.128	19793	18250	0.370	0.377	-0.009
RIVPH_IN_MUR	0.455	0.244	0.459	0.114	20100	18176	0.412	0.401	0.004
RIVPH_BU_MUR	0.422	0.192	0.419	0.118	20157	19246	0.403	0.393	-0.010
RIVBR_MUR	0.413	0.153	0.417	0.111	19984	19614	0.406	0.403	0.007
RIVTR_ANA	0.393	0.160	0.409	0.117	19498	19068	0.384	0.395	0.038
RIVEG	0.395	0.160	0.400	0.123	19218	18620	0.383	0.386	0.008
RIVIR_AZA	0.407	0.184	0.411	0.122	19815	18865	0.388	0.388	0.006
RIVIR_KHU	0.387	0.177	0.403	0.125	19882	18865	0.367	0.383	0.039
RIVIR_MAZ	0.402	0.193	0.404	0.128	19837	18119	0.367	0.378	0.000
RIVPK_AZK	0.481	0.262	0.485	0.108	20327	17384	0.411	0.404	0.009
RIVPK_KUN	0.423	0.178	0.420	0.115	20091	19552	0.412	0.399	-0.009
RIVPK_NIL	0.422	0.154	0.418	0.109	19994	19755	0.417	0.404	-0.013
RIVCO	0.415	0.171	0.424	0.108	19936	19596	0.408	0.406	0.019
SWAPH	0.302	0.176	0.315	0.157	18905	16078	0.257	0.331	0.037
SWAPH_ADM	0.426	0.187	0.414	0.118	20029	19451	0.413	0.391	-0.032
SWABR_CAR	0.369	0.198	0.348	0.148	20221	16010	0.292	0.331	-0.064
SWATH_THS	0.364	0.200	0.373	0.139	20341	16433	0.294	0.342	0.026
SWATH_THT	0.332	0.184	0.355	0.145	20332	16653	0.272	0.332	0.067
SWACN_ENS	0.324	0.178	0.332	0.152	19858	16141	0.264	0.321	0.021
SWACN_FUL	0.328	0.180	0.333	0.152	19950	16104	0.264	0.322	0.014

SWACN_GUI	0.327	0.179	0.342	0.149	20131	16147	0.262	0.327	0.045
SWACN_HUN	0.328	0.179	0.327	0.153	19974	16876	0.277	0.316	-0.003
SWACN_YAN	0.337	0.184	0.336	0.150	19424	15864	0.275	0.322	-0.006
SWACN_YIB	0.324	0.177	0.332	0.152	19805	16081	0.263	0.321	0.021
SWAID_JAV	0.334	0.182	0.342	0.150	19376	13453	0.232	0.328	0.019
SWAID_NUT	0.357	0.197	0.377	0.139	20223	12453	0.220	0.350	0.055
SWAID_SUM	0.333	0.181	0.335	0.148	17467	14738	0.281	0.321	-0.005
SWAID_SUW	0.334	0.184	0.357	0.146	20046	13489	0.225	0.340	0.066

^ Corrected over the number of usable loci; * highlights statistically significant tests (P<0.05).

Wright's fixation index F_{ST} was always significant (P<0.05; Supplementary 3.8.3), with the exception of the following pairwise comparisons: RIVPK_NIL vs. RIVPH_IN_MUR, RIVPK_AZK vs. both RIVPK_KUN and RIVPK_NIL, and SWATH_THS vs. SWATH_THT. F_{ST} values ranged from 0.004 (SWACN_GUI vs. SWACN_YIB) to 0.448 (SWAID_JAV vs. RIVMZ) overall; from 0.006 (RIVPK_AZK vs. RIVPH_IN_MUR) to 0.199 (RIVIR_MAZ vs. RIVMZ) among the river buffalo group; from 0.004 (SWACN_GUI vs. SWACN_YIB) to 0.232 (SWAID_NUT vs. SWABR_CAR) among the swamp buffalo group; from 0.104 (RIVPK_AZK vs. SWAPH_ADM) to 0.448 (SWAID_JAV vs. RIVMZ) between river and swamp populations.

According to the results of JAATHA, the number of migrants varied between 0.010 and 75, with the most extensive gene flows occurring between river buffalo populations and between the swamp populations from China (Supplementary 3.8.3 and 3.8.4). In detail, the occurrence of extensive exchanges represents a general trend within the river group, with the few exceptions of RIVMZ from Mozambique and RIVPK_AZK from Pakistan, and to a lesser extent RIVRO from Romania, RIVIT_MED from Italy and RIVIR_MAZ from Iran.

Among the swamp buffaloes, very high levels of gene flow were estimated among the Chinese populations, between SWATH_THT and SWATH_THS populations from Thailand, and from SWATH_THT to the Chinese population SWACN_GUI. In addition, the admixed swamp population from the Philippines SWAPH_ADM shows signs of gene flow with several river-type populations (RIVCO, RIVPK_NIL, RIVPK_KUN, RIVEG, RIVTR_ANA, RIVPH_IN_MUR).

The Multi-Dimensional Scaling plot (Figure 3.2) allowed to evaluate the relationships among

the individual multi-locus genotypes in a multivariate framework. According to the estimated eigenvalues 3.8.4), around 59% of the total molecular variance is explained by the first three dimensions. In particular, dimension one (*X*-axis in both panels of Figure 3.2) explains 53.55% of the original molecular variance, separating river- from swamp-type individuals, with the admixed individuals from the Philippine being placed at an intermediate position. The second dimension (2.80% of variation; *Y*-axis of the left panel in Figure 3.2) separates the groups of river-type individuals based on their geographical provenance and genomic relationships, but also the Carabao population from Brazil (SWABR_CAR) from the other swamp buffaloes. In detail, from top to bottom of the second dimension axis we can identify: (i) a first group of points representing the populations from Italy and Mozambique (RIVIT_MED and RIVMZ), (ii) the group of river buffaloes from Romania (RIVRO), (iii) a group including the Murrah breed populations from Bulgaria, Brazil and India, together with the population from Colombia; (iv) the group of animals from Turkey, Egypt and Pakistan (RIVTR_ANA, RIVEG,RIVPK_AZK, RIVPK_KUN, RIVPK_NIL) in close continuity with the populations from Iran (RIVIR_AZA, RIVIR_KHU, RIVIR_MAZ). Notably, the position of the swamp Carabao breed on the second axis corresponds to that of the river population from Romania.

Similarly, the third dimension (2.56% of variation; Figure 3.2 right panel, *Y*-axis) separates the swamp populations as follows: three populations of Java, Nusa Tenggara and South Sulawesi from Indonesia (SWAID_JAV, SWAID_NUT, SWAID_SUW) are positioned on top of the axis, and are separated by a large gap from the Indonesian population of Sumatra (SWAID_SUM), which lies closer to the group formed by the individuals from Thailand

(SWATH_THT, SWATH_THS) and the Brazilian Carabao (SWABR_CAR), while the individuals from China and the Philippines are positioned at the bottom of the axis.

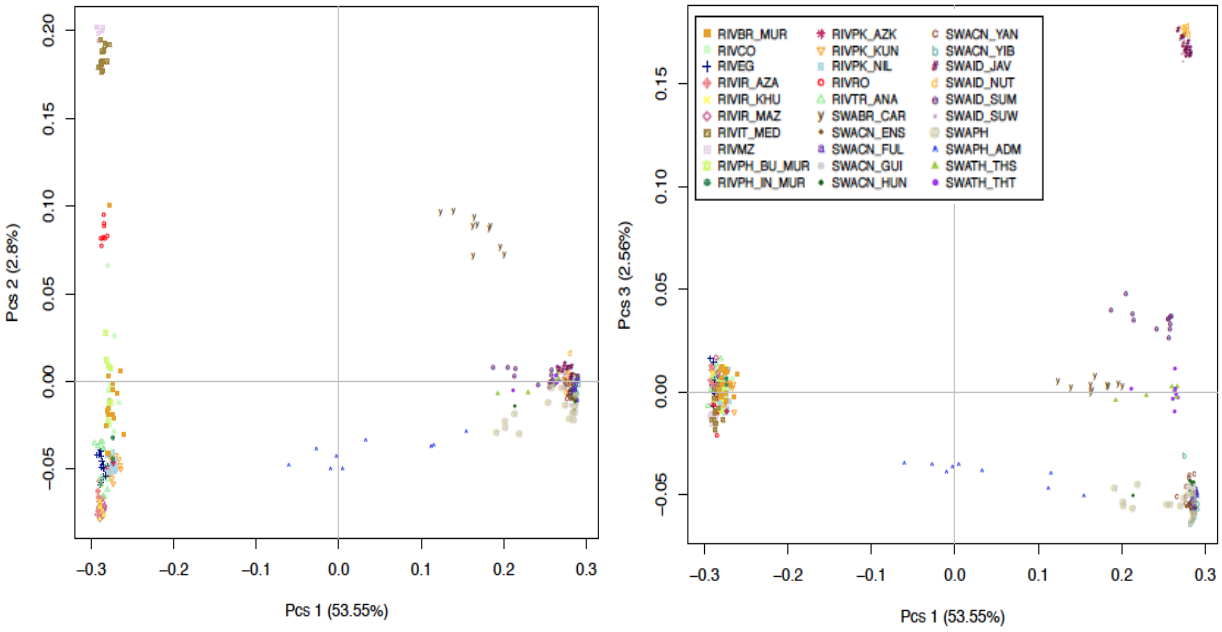

Figure 3.2 Multi-Dimensional Scaling plot of first vs. second dimension (left panel) and first vs. third (right panel). The percentages of variance explained by each dimension are reported into brackets.

Both AMOVA and the neighbour-net reconstructed from the DR matrix corroborate the results of the MDS. In fact, a large fraction of the variance (25.71%; Table 3.3a) explains the subdivision into river- vs. swamp-type groups, and the percentage further increases to 26.72% when the admixed population from the Philippines is removed from the analysis (Table 3.3b). About 5.75% of the variance is assigned to the "among populations within groups" component (Table 3.3b), while the variation among individuals within populations is very low (0.69%; Table 3.3b).

Table 3.3a Analysis of molecular variance performed on river-type and swamp-type populations.

Source of variation[a]	d.f.[b]	Sum of squares	Variance components	Percentage of variation
Among groups	1	422395.22	1263.31	25.71
Among populations within groups	28	271650.32	291.78	5.94
Among individuals within populations	297	1006390.28	29.62	0.60
Within individuals	327	1088674.00	3329.28	67.75
Total	653	2789109.82	4913.99	100.00

[a]All values have been calculated after removing the anoa population from the dataset; [b]d.f.: degrees of freedom

Table 3.3b Analysis of molecular variance performed on river-type and swamp-type populations after removing admixed individuals from the Philippines.

Source of variation[a]	d.f.[b]	Sum of squares	Variance components	Percentage of variation
Among groups	1	430136.13	1321.17	26.72
Among populations within groups	27	258177.63	284.45	5.75
Among individuals within populations	289	974756.17	34.35	0.69
Within individuals	318	1050726.00	3304.17	66.83
Total	635	2713795.93	4944.14	100.00

[a]As above; [b]d.f.: degrees of freedom

The neighbour-net confirms the subdivision into the two groups and the intermediate position of SWAPH_ADM (Supplementary 3.8.6). Among the river-type populations (right side of Supplementary 3.8.6), RIVBR_MUR and RIVPK_NIL are placed in a basal position, while the remaining populations are split into three sub-networks, the first one formed by RIVCO, RIVIT_MED, RIVMZ, RIVRO and RIVPH_BU_MUR, the second by RIVEG, RIVTR_ANA, RIVIR_AZA, RIVIR_KHU and RIVIR_MAZ; the third by RIVPH_IN_MUR, RIVPK_AZK and RIVPK_KUN. Moreover, the river buffaloes from Mozambique are characterized by the longest branch, which stems directly from that of the Italian

Mediterranean population.

Also among the swamp-type populations (left side of Supplementary 3.8.6) three main network subdivisions are recognizable: (i) the branch of the Indonesian population from Sumatra (SWAID_SUM) stemming close to (ii) the sub-network which includes the buffaloes from Java, Nusa Tenggara and South Sulawesi (SWAID_JAV, SWAID_NUT, SWAID_SUW) and which is also characterized by very long branches; (iii) a further sub-network encompassing the Chinese swamp buffaloes (SWACN_GUI, SWACN_ENS, SWACN_FUL, SWACN_YIB, SWACN_HUN, SWACN_YAN), and the branch of the population from the Philippines (SWAPH).

The two populations from Thailand (SWATH_THT and SWATH_THS) are placed in a basal position, while the Brazilian Carabao branch forks at a distance from the network formed by the remaining swamp populations.

According to ADMIXTURE analysis, the first subdivision (K=2) is between river- and swamp-type groups of populations (Figure 3.3). ADMIXTURE bar plots show an admixed ancestry for SWAPH_ADM and some degree of introgression of the river-type gene pool into the swamp populations of Brazil (SWABR_CAR), the Philippines (SWAPH), Sumatra (SWAID_SUM) and Thailand (SWATH_THT and SWATH_THS). The river populations from Bulgaria, India, Pakistan and South America show signs of a small but widespread contribution from the swamp-type gene pool. At K=3 (Supplementary 3.8.7), a further split occurs within the river cluster, separating the Italian Mediterranean breed and the population from Mozambique. The same genomic component is present at high percentage in the river populations from Romania, Bulgaria and South America (RIVBR_MUR, RIVCO), as well as in the swamp Carabao from

Brazil. At K=4 (Figure 3.3), the aforementioned behaviour is confirmed, but a further component comes into view within the swamp-type group, grouping the Indonesian populations from Java, Nusa Tenggara and South Sulawesi. This component is also found at a high percentage in the populations from Sumatra, those from Thailand and the Carabao. The subsequent component (K=5; Supplementary 3.8.7) appears in the Thai populations, while characterizing Carabao as a distinct cluster. The six-cluster model showed the lowest cross-validation error (together with a low number of iterations required to reach convergence), and was therefore considered the optimal solution (Supplementary 3.8.8). The corresponding bar plot (Figure 3.3) discloses an additional component within the river group, typical of the populations from Pakistan, India, Bulgaria, South America, and also present to a lesser extent in Egypt, Romania and Turkey. The same signal occurs in the swamp populations from Sumatra and the Philippines.

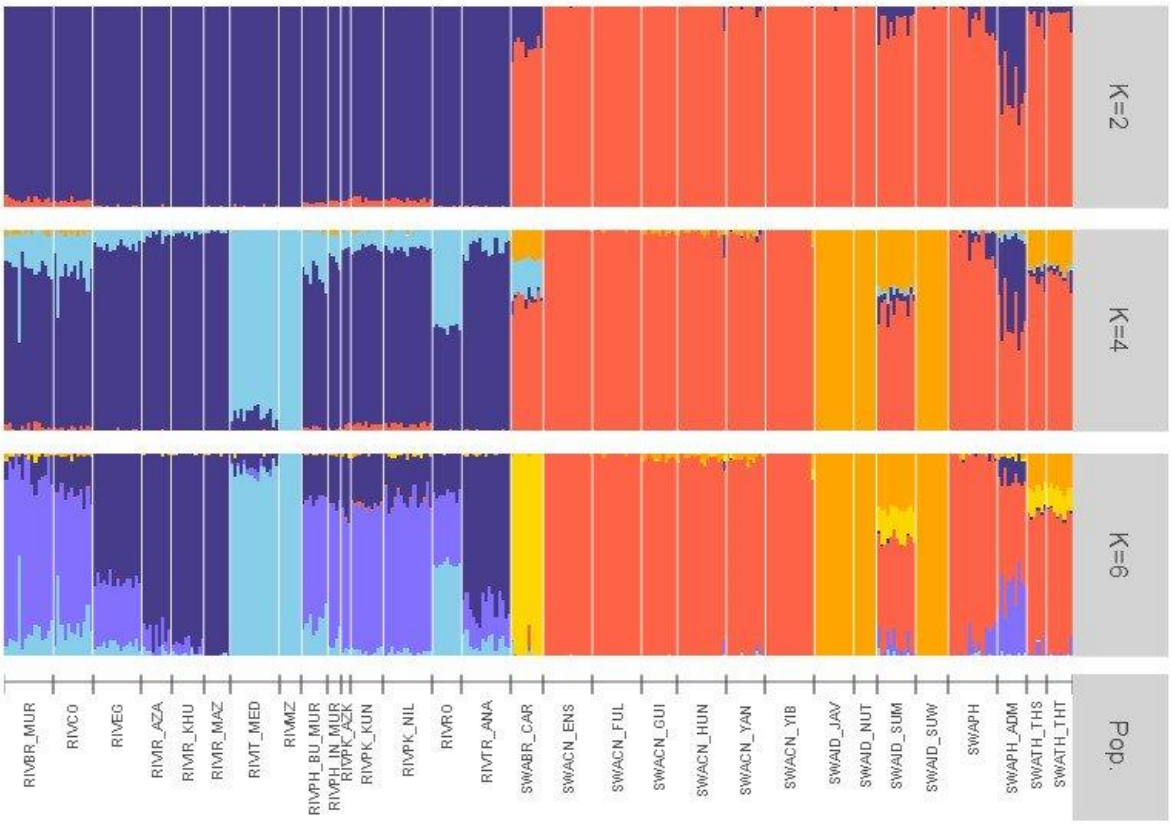

Figure 3.3 Bar plots of ADMIXTURE results at K=2, 4 and 6 (best clustering solution).

After the addition of 14 anoa individuals (outgroup) and the removal of the markers with missing data, the dataset for TREEMIX analysis involved 341 individuals and 12,601 SNPs. The starting tree (m_0) accounts for 99.16% of the variance and this percentage gradually grows up to 100% as the number of migrations increases to 15 (Supplementary 3.8.9 and 3.8.10). Based on the cumulated value of variance explained (99.96%), the fraction of statistically significant migrations modelled (100%) and literature support for the inferred migration edges, the graph with five migrations was selected to run the subsequent bootstrap-based analysis (Supplementary 3.8.10).The consensus tree obtained from the 100 replicates shows all nodes to be supported by bootstrap values above 50, except for the branch separating RIVPK_NIL

and RIVPK_KUN from RIVPK_AZK, and the branch corresponding to the split of SWABR_CAR from the Indonesian and Chinese populations. The graph obtained at m_5 (Figure 3.4) displayed—in order of decreasing weight—the following migration edges:

1) from the branch of RIVPK_NIL to SWAPH_ADM;

2) from the branch of RIVRO to RIVPH_BU_MUR;

3) from the branch basal to RIVIT_MED and RIVMZ to SWABR_CAR;

4) from RIVRO to the basis of the branch of RIVPH_IN_MUR and RIVPH_BU_MUR;

5) from RIVPK_KUN to SWAPH.

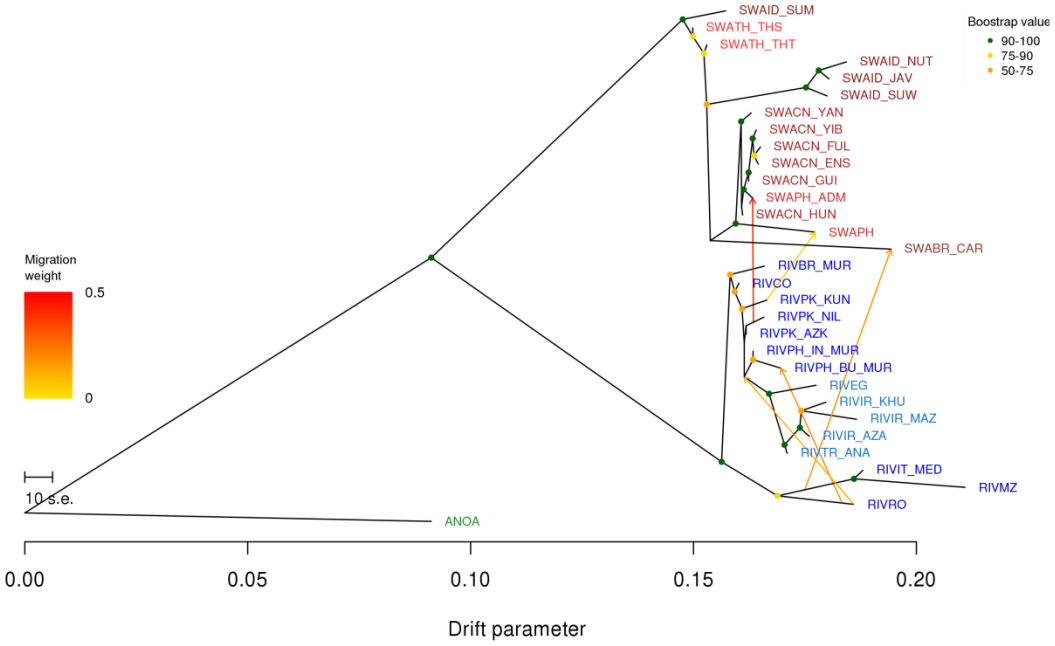

Figure 3.4 TREEMIX graph depicting five assumed migration events. The robustness of the branches was calculated over 100 bootstrap replicates, and is indicated by the following colour key: green dots=90-100, yellow dots=75-89, orange dots=50-74, red= <50.

The highly admixed nature of SWAPH_ADM population was further supported by the related *f3*-statistics (Reich *et al.* 2009) (data not shown), where SWAPH_ADM was significantly detected as receiver in 119 tests involving one swamp and one river source population as donor pairs. Moreover, *f3*-statistics pointed out the Chinese populations as the most certainly admixed (54 significant tests out of 119 performed).

3.5 Discussion

3.5.1 Performance of the Axiom® Buffalo Genotyping Array

According to our results, the Axiom® Buffalo Genotyping Array proved to be an efficient tool for the molecular characterization of water buffalo populations. In fact, compared to the results obtained when cattle-specific tools were used on water buffalo (Michelizzi *et al.* 2011; Borquis *et al.* 2014), the 90K array allows to increase 56.7 times the number of polymorphic markers (52,520 polymorphic markers of the present work vs. 926 in Michelizzi *et al.* 2011) and by 40.5 percentage points the level of polymorphism scored (51,765 out of 89,988 markers with MAF>0.05, i.e. 57.5%, vs. 131,991 out of 777,962, i.e. 17.0% in Borquis *et al.* 2014). Thus, this tool represents the best option available at present for the molecular characterization of *B. bubalis* in terms of both cost-effectiveness and information content, although with some caveats.

However, due to the over-representation of river buffalo breeds in the SNP discovery panel, the array proved to be affected by a moderate-to-high degree of Ascertainment Bias, as also

described by Iamartino *et al.* (in preparation) and confirmed by our results: only about 22.74% of the markers on the chip were polymorphic in swamp buffalo populations.

Anyway, the strategy adopted here (i.e. the use of the polymorphic markers in swamp buffaloes only), allowed to reduce the AB impact, as shown by the increase in the observed heterozygosity among swamp populations (Supplementary 3.8.1 and 3.8.2). Nevertheless, this approach was probably not sufficient to completely remove the bias, since both in the MDS (second dimension) and in Admixture analysis (K=3), the trends occurring among river populations were always revealed earlier than those among swamp populations.

Regarding the possible utilization of the array outside the water buffalo species, the chip turns out to be heavily affected by AB, since only 4,090 markers out of 89,988 (4.55%) were scored as polymorphic in the Lowland anoa (*B. depressicornis*). However, it is worth stressing that anoa experimented a strong reduction in population size in the recent decades (Burton *et al.* 2005), a fact which might affect the actual level of polymorphism in the species. Nevertheless, we consider advisable to evaluate the performance of the SNP array on a wider set of species before extensively using this tool to characterize wild buffaloes.

3.5.2 Molecular variability of river and swamp buffalo populations

Among the river buffalo breeds, the Pakistani Nili-Ravi (RIVPH_NIL, H_{obs}=0.417), Kundhi (RIVPK_KUN, H_{obs}=0.412) and Aza Kheli (RIVPK_AZK, H_{obs}=0.411) showed the highest values of observed heterozygosity together with the Murrah population of Indian origin reared in the Philippines (RIVPH_IN_MUR, H_{obs}=0.412). This evidence agrees with previous

research based on microsatellite (Kumar *et al.* 2006; Vijh *et al.* 2008) and mitochondrial markers, which suggested North-Western India as the most probable domestication centre for river-type buffaloes (Nagarajan *et al.* 2015). However, the higher values of heterozygosity observed in Murrah and Nili-Ravi may have also been influenced by AB, since these breeds were among those included in the SNP discovery panel (Iamartino et al. in preparation). Assuming a uniform impact of AB on the breeds used in the discovery panel, nevertheless, a similar inflation in H_{obs} should have also been expected for the Mediterranean breed, which ranks, on the contrary, among the most heterozygous ones (RIVIT_MED, H_{obs}=0.359).

A general agreement among SNP- and microsatellite-based heterozygosity estimates emerges from our comparisons with literature. However, a discrepancy regards the Egyptian population: contrarily to a previously reported microsatellite-based estimate of 0.872-1.000 (El-Kholy *et al.* 2007), we observe a considerably lower observed heterozygosity (H_{obs}=0.383) in line with those of the neighbouring populations (Table 3.2). Such an evident difference might be explained either by the "animals exchange policy between the different regions over Egypt", which could have produced a systematic outbreeding among the analysed breeds in (El-Kholy *et al.* 2007), or a biased selection of the used microsatellites towards the most polymorphic ones.

The observed trend in H_{obs} is mostly confirmed by the corrected H_{exp} values (RIVPK_NIL, H_{exp}=0.404; RIVPK_KUN, H_{exp}=0.399; RIVPK_AZK, H_{exp}=0.404; RIVPH_IN_MUR, H_{exp}=0.401), which also indicated the river populations from Colombia (H_{exp}=0.406) and the Murrah from Brazil (H_{exp}=0.403) as highly heterozygous. In particular, the high H_{exp} values observed in South America might mirror the Indian ancestry of the analysed populations,

combined with a limited—but detectable—crossbreeding with Mediterranean water buffaloes.

Concerning the swamp-type populations, the highest H_{obs} values were observed in Thailand (SWATH_THS, H_{obs}=0.294), in agreement with previous microsatellite-based findings (Barker *et al.* 1997; Zhang *et al.* 2011). Lower values of H_{obs} are observed in the insular populations from Java (H_{obs} =0.232) and South Sulawesi in Indonesia (H_{obs}=0.225), in agreement with Zhang *et al.* (2011) and Barker *et al.* (1997). Most of the Chinese populations had similar H_{obs} values (SWACN_ENS, H_{obs}=0.264; SWACN_FUL H_{obs}=0.264; SWACN_GUI, H_{obs}=0.262; SWACN_YIB, H_{obs}=0.263), with only those from South-eastern China showing slightly higher values (SWACN_HUN, H_{obs}=0.277; SWACN_YAN, H_{obs}=0.275). Such a finding is in agreement with the previously described uniformity among the Yangtze river valley populations (Zhang *et al.* 2011), and the higher differentiation reported in the populations inhabiting the South-eastern regions of China. Admixed individuals from the Philippines (SWAPH_ADM) stand out among swamp populations, by displaying an observed heterozygosity up to 0.413, deriving from crossbreeding with the river-type gene pool.

F_{IS} values ranged from slightly positive (SWATH_THT, F_{IS}=0.067) to slightly negative (SWABR_CAR, F_{IS}=−0.064), and they were never statistically significant (P<0.05) (Table 3.2).

Marques *et al.* (2011) reported statistically significant F_{IS} values calculated from microsatellite markers for Carabao (0.057) and Brazilian Murrah (0.135) breeds, by evidencing a trend opposite to our findings (-0.064 and 0.007, respectively). Such a difference may be explained by the possible occurrence of null alleles, genotyping errors or sampling bias. In particular, the

animals were selected in highly structured herds from different states of Brazil, possibly leading to a Wahlund effect with consequent deviations from HWE expectations.

Our results point to the existence of a number of distinct and well differentiated gene pools within the analysed buffalo populations. As expected, the most evident subdivision is between river- and swamp-type buffaloes. This subdivision was clearly highlighted by all the analyses we performed, accounting for 26.72% of the total molecular variance in AMOVA, and being depicted by the first MDS dimension (Figure 3.2). Therefore, even considering the effect of ascertainment bias, the considered set of markers shows a remarkable type-specific differentiation in the level of variability, by supporting the assignment of river and swamp buffaloes to different subspecies (Macgregor 1941).

Within-type subdivisions highlight the presence of genetic clusters that share a common ancestry either due to geographical origin (as in the case of the river breeds from Egypt, Turkey and Iran, or the swamp populations from Java, Nusa Tenggara and south Sulawesi), or to human-mediated translocations (as in the case of the Mozambique population imported from Central Italy (Cockrill 1974).

This scenario is made more complex by the occurrence of a number of admixture events both between- and within-type, and mostly dating back to the last century. Between-type admixture seems to be mainly unidirectional from the river towards the swamp gene pool: South-eastern Asian populations (from the Philippines, Sumatra and Thailand) show clear signals of a river-type genomic contribution that, according to the results of JAATHA (Supplementary 3.8.3), ADMIXTURE (Figure 3.3, K=6) and TREEMIX (Figure 3.4), likely originated from the breeds of the Indo-Pakistani region. Conversely, the river-type input received by the Brazilian Carabao

seems to derive from the Mediterranean gene pool (Figure 3.3 and 3.4), an evidence further supported by the MDS (Figure 3.2).

All these findings agree with bibliographic records that account for the establishment of crossbreeding programs in several countries to increase milk production in swamp populations (Iannuzzi & Di Meo 2009). More in detail, the literature accounts for: (i) the common practice of crossing river and swamp buffaloes in the Philippines (Reyes 1948 cited in Cockrill 1974); (ii) an importation of Bulgarian Murrah animals to the Philippines in the 1990s (Borghese 2011); (iii) a limited introduction of Murrah buffaloes to Sumatra (Cockrill 1974); (iv) several importations of Mediterranean buffalo from Italy into Brazil (starting from the late XIXth century until the mid XXth, (Cockrill 1974), and (v) the extensive crossbreeding between the river and swamp types carried out in several southern American countries (Iannuzzi & Di Meo 2009).

Within-type admixture occurs both in river and in swamp buffaloes, even if to a larger extent in the river-type. According to JAATHA results, in fact, riverine populations exchange a high number of migrants with each other (Supplementary 3.8.3 and 3.8.4), with a few exceptions represented by the Mediterranean breeds (particularly individuals from Mozambique), Aza Kheli breed from Pakistan (RIVPK_AZK) and Mazandarani breed (RIVIR_MAZ) from Iran.

The highlighted gene flow events occurred between the Romanian population (RIVRO) and the Murrah from Bulgaria and India (RIVPH_BU_MUR and RIVPH_IN_MUR) are confirmed by historical information describing the importation of Murrah animals from India to Bulgaria in 1962, their subsequent crossing with the indigenous Mediterranean to establish the Bulgarian Murrah, which was later crossed with the Romanian populations (Borghese

2011).

Molecular analyses and bibliographic record both suggest southern American river buffaloes to derive from the Indo-Pakistani breeds with a further, although minor, contribution from the Mediterranean gene pool. This hypothesis is supported by both ADMIXTURE (K=6, Figure 3.3), which reveals a strong similarity between the genetic makeup of the aforementioned groups, and the neighbour-network (Supplementary 3.8.6), in which RIVBR_MUR and RIVCO are placed at an intermediate position between the edges corresponding to Pakistani and Mediterranean populations. Furthermore, model residuals from TREEMIX analysis (Supplementary 3.8.11) show that the pairs formed by RIVCO with the three populations of clear Mediterranean ancestry (RIVIT_MED, RIVMZ and RIVRO) all have highly positive values, thus indicating that the overall fitting of the model could be increased if migration edges between these populations were postulated.

According to previous research and historical records, the first buffaloes reaching Sao Paulo (in 1904 and 1920) and Minas Gerais (in 1919) were native to India. A large part of the present-day population derives from this initial nucleus, with the Indian Murrah and Jaffarabadi representing the principal river breeds in Brazil (Cockrill 1974). Contextually, Mediterranean buffaloes have been imported to Brazil several times starting from the end of the XIXth century throughout the whole XXth century (e.g. see the case of the recorded arrival of Italian buffaloes to Sao Paulo in 1948, Cockrill 1974).

Gene flow within swamp-type buffaloes seems to be generally less pronounced and to involve mostly the Chinese populations (Supplementary 3.8.3 and 3.8.4). An extensive exchange is also detectable between SWACN_GUI, the southernmost Chinese population, and

SWATH_THT from Thailand, a finding which appears consistent with SWACN_GUI geographical position (Figure 3.1)

Overall, a lack of differentiation and low level of variability are suggested for the Chinese swamp buffalo populations by the majority of our analysis: in ADMIXTURE plots, they remain tightly assigned to the same cluster until K=10 (data not shown); in the MDS plot (dimension one vs. three), they overlap completely in a very reduced area of the graph (Figure 3.2, right panel); in the Neighbour-network, they are placed on very short edges close to the basal network (Supplementary 3.8.6).

This evidence confirms previous analyses based on microsatellite data showing (i) the differentiation among Chinese populations to be generally much lower than that occurring across the South-East Asia, and (ii) the populations of South-East China to be more closely related to the Indochinese ones then those from South-West China, more similar to Indonesia and the Philippines (Zhang *et al.* 2007, 2011). Further support is provided by studies based on mitochondrial control region data, suggesting a weak phylogeographic structure and extensive gene flow between Chinese swamp buffalo populations (Yue *et al.* 2013).

According to our analyses, a moderate level of gene flow and an extensive genomic uniformity also characterize the Indonesian populations from Java, Nusa Tenggara and South Sulawesi (Supplementary 3.8.4, Figure 3.2 and 3.3). These populations appear separated from the remaining swamp buffalo nuclei, probably due to the effect of geographical isolation and genetic drift, as suggested by: (i) their positioning in the upper-left corner of the MDS plot (Figure 3.2, right panel), (ii) their placement on long branches in the Neighbour-network (Supplementary 3.8.6), and (iii) the assignment to a well-defined cluster in admixture analysis

starting from K=4 (Figure 3.3) to K=15 (data not shown). The population from Sumatra, on the contrary, seems to be closely related to the Thai swamp buffaloes, although no evidence of gene flow was obtained by our analyses between the groups.

According to Cockrill (1974), Dutch colonizers introduced swamp buffaloes to Southern America (i.e. Suriname) as draught animals for work in the sugarcane plantations, and Kierstein *et al.* (2004) stated that part of the present day Carabao population in Brazil was imported from the Philippines. However, our results suggest the considered Brazilian Carabao population to have more likely originated from Thailand or Sumatra, as supported by the dimension three of the MDS (Figure 3.2), and the admixture analysis (Figure 3.3).

Furthermore, we hypothesize the genomic relatedness between swamp buffaloes from Sumatra and Thailand to be more probably linked to the ancestral origin of these populations rather than to recent demographic events (see Supplementary 3.5.3).

3.5.3 Domestication and post-domestication migration routes

Two alternative hypotheses on water buffalo domestication have been long debated, contemplating either a single (Kierstein *et al.* 2004) or two separate domestication events for river and swamp buffaloes (Lau *et al.* 1998; Ritz *et al.* 2000; Kumar *et al.* 2007a; 2007b; Lei *et al.* 2007; Yindee *et al.* 2010; Zhang *et al.* 2016).

Based on the most recent and extensive molecular evidence, it is likely that the two types have been domesticated starting from different populations of the same wild ancestor *B. arnee* in different geographical areas of the Asian continent, in particular, North-western India

(Nagarajan *et al.* 2015) for river buffaloes and the region close to the border between China and Indochina (Zhang *et al.* 2011, 2016) for swamp buffaloes.

From the archaeological point of view, the analysis of bone measurements and demographic profiles performed on ancient buffalo remains from southern Asia and Neolithic China (Patel & Meadow 1998; Liu *et al.* 2004) also points to the former area as a probable centre of buffalo domestication. This hypothesis is further supported by the presence of domestic buffalo bones at Ban-Tamyae site in Central Thailand (2,600-2,200 years BP; Higham 1989), Ban-Chiang site in northern Thailand (4,300-2,500 BP; Higham 2002), and Phum Snay in northwestern Cambodia (2,200-1,760 BP; O'Reilly *et al.* 2006), while the findings at the sites of Kuahuqiao (8,000-7,500 BP) and Luo Jiajiao (7,000 BP) in the Zhejiang region of China (Liu *et al.* 2004) probably belonged to the extinct wild species *Bubalus mephistopheles*, thus disproving the hypothesis of a Chinese swamp buffalo domestication centre. Nor ancient DNA analyses carried out on samples from Neolithic-to-Bronze Age sites of the Shaanxi Province of China could confirm this area as a probable domestication centre, but rather highlighted a genetic discontinuity between the pre-historical and the present day Chinese water buffalo populations (Yang *et al.* 2008).

Concerning the post-domestication dispersal of the species, literature based on archaeological and historical evidence reports that the seal impressions from the Mohenjo-Daro civilization of the Indus Valley (5,000-4,500 BP; Clutton-Brock 1999, Zeuner 1963) and from the Ur royal cemetery in Mesopotamia (4,500 BP; Clutton-Brock 1999) are among the oldest findings testifying the presence of domesticated buffaloes outside their area of origin. According to the same literature, neither wild nor domestic water buffaloes were known west of Mesopotamia

in the ancient world (Manson 1974; Clutton-Brock 1999), and they did not reach the Mediterranean until the middle Ages, even though there is no general agreement on the century of arrival. The first documented record of the presence of domestic buffaloes in the eastern Mediterranean is from 723 AD in the Jordan valley, where they seem to have been brought from Mesopotamia by the Arabs (Manson 1974), who likely mediated also the introduction of domestic buffaloes to Egypt after its conquest in the IX century (Sidky 1951, cited by Manson 1974). Bökönyi (1974, cited in Clutton-Brock 1999) reports that, from about the VII century AD, domestic buffaloes had already become common draught and dairy animals in Italy and South-Eastern Europe. Similarly, Iannuzzi & Di Meo (2009) state that the Italian Mediterranean buffalo has never been crossed with other breeds since its introduction to Italy from Northern Africa (Egypt) or central Europe during the V to VII century AD, contrary to other European countries whose Mediterranean buffalo populations have frequently been crossed primarily with the Indian Murrah.

Other authors suggest a later arrival in Europe: according to Kaleff (1942) domestic buffaloes were brought back by the returning Crusaders, and could be found in sizable numbers in Thrace, Macedonia and other parts of Bulgaria at the beginning of XIII century. They subsequently spread to the rest of Eastern Europe and reached central Italy, where their presence in the Pontine marshes was recorded at the end of the XIII century (Ferrara 1964).

Regarding swamp buffalo post-domestication dispersal routes, the species was known in China by the fourth millennium BP at the time of the Shang dynasty (*c.* 1,766-1,123 BCE) and appeared to have been introduced from bordering areas of South-eastern Asia (Epstein 1969). According to records from ancient texts and art representations, Yue *et al.* (2013) report

domestic swamp buffalo to have probably appeared in southwestern China in the Yunnan region during the first century of the Common Era, subsequently spreading to the rest of the country. The authors also hypothesize that the southwestern Silk Road connecting Sichuan via Yunnan and Burma with southern Asia, may have played a role in the exchange of livestock, including water buffaloes.

Traditionally, from the molecular point of view, descriptors such as heterozygosity and allelic richness for microsatellites, or nucleotide and haplotype diversity for mtDNA have been used to identify the most probable domestication centres: when the populations bearing clear signs of recent introgression or outbreeding are excluded and the values of such statistics are placed in a geographical framework, it was shown that the areas with higher figures often correspond or lay close to centres of domestication previously suggested by archaeological findings. Moreover, it was shown that a gradual decrease in such values usually occurs along the migration routes out of the domestication centres (Troy *et al.* 2001; Beja-Pereira *et al.* 2004; Cañón *et al.* 2006; Groeneveld *et al.* 2010; Vahidi *et al.* 2014).

In the case of river buffalo, microsatellite-based estimates of diversity, although obtained with different marker panels, showed that the highest values of heterozygosity among river breeds were found in India (H_{exp}=0.71-0.78; Kumar *et al.* 2006) and moderately decrease to H_{exp}=0.58-0.68 in Italy (Moioli *et al.* 2001; Elbeltagy *et al.* 2008).

Similar evaluations applied to mtDNA and Y chromosome data from Asian water buffalo populations, confirmed that swamp buffalo domestication likely occurred in China-Northern Indochina, and also highlighted a complex scenario characterized by a weak phylo-geographic structure in river buffalo, a strong geographic differentiation of swamp buffaloes, and the post-

domestication introgression of wild buffalo lineages into the domestic stocks. Furthermore, the presence of a higher sequence diversity in swamp compared to river buffaloes suggested that a wider representation of wild ancestor lineages was sampled in the former case at the time of domestication (Zhang *et al.* 2016). According to these authors, for river buffalo the migration out of the domestication centre through Southwestern Asia to Europe occurred more gradually than for the majority of other livestock species (i.e. cattle, sheep, goat and horse) and without substantial bottlenecks. On the contrary, the diffusion of swamp buffalo was characterized by strong matrilocality and occasional incorporation of wild females into the herds, and probably occurred in association with the spread of rice cultivation: starting from the China/Indochina region, domesticated swamp buffalo simultaneously migrated northeast along the coasts of China, east and northeast along the Yangtze river valley both down- and upstream, and south on both sides of the Mekong river valley.

Considering our results, among the sampled river buffalo populations, the breeds from Pakistan, RIVPK_NIL, RIVPK_KUN and RIVPK_AZK, and the Indian Murrah reared in the Philippines, RIVPH_IN_MUR, are characterized by the highest figures for corrected H_{obs} (Table 3.2), and also lay on the branches closer to the midpoint in the neighbour-network (Supplementary 3.8.6) and to the root in the TREEMIX graph (Figure 3.4). Furthermore, the heat map of TREEMIX m_5 model residuals shows the pairs formed by ANOA with RIVPK_NIL, RIVPK_AZK and RIVPH_IN_MUR to have quite high and positive residual values, suggesting the addition of migration edges between these populations to potentially increase the model fitting to the data. Nevertheless, this evidence should be interpreted with caution due to the very low level of polymorphism scored in the ANOA population. Anyway,

it is interesting to note that the Indo-Pakistani river buffalo breeds from the region close to the putative domestication centre are also those that TREEMIX analysis highlights as related to the wild relative *B. depressicornis*.

Conversely, the Mediterranean breeds RIVIT_MED, RIVMZ and RIVRO display the lowest H_{obs} and H_{exp} values and also bear signs of a long-time isolation, as highlighted by their behaviour in the MDS (Figure 3.2, left panel) and by the separate subclades with very long branches that they form both in the neighbour-network (Supplementary 3.8.6) and in TREEMIX graph (Figure 3.4). The distinctiveness of the Mediterranean gene pool is also evident in both TREEMIX and ADMIXTURE analyses, since the first split occurring among river buffalo breeds is that parting the Mediterranean group from the rest, while a second split separates the group formed by the breeds from Egypt (RIVEG), Turkey (RIVTR_ANA) and Iran (RIVIR_AZA, RIVIR_KHU and RIVIR_MAZ).

Regarding the Iranian breeds, a previous study based on mitochondrial DNA (Nagarajan *et al.* 2015) highlighted a high degree of distinctiveness of Iranian buffaloes and lack of haplotype sharing with other populations (India, Egypt and Pakistan), a behaviour particularly striking in the case of Pakistani breeds, considering the geographical proximity of the two countries. This evidence was interpreted as the clue of an ancient migration of river buffaloes from India to Iran, occurred through maritime rather than terrestrial routes and followed by intense genetic drift. The authors also hypothesize a later arrival of buffaloes in Egypt due to a haplotypic composition more similar to the present day mitochondrial lineages of the Pakistani and Indian buffaloes.

Our results in part agree with the aforementioned mtDNA evidence by showing that, despite

the geographical continuity between Pakistan and Iran, the buffalo populations of these countries seem to belong to different gene pools, with the Iranian buffaloes being evolutionarily closer to those from Egypt and Turkey (Supplementary 3.8.6, Figure 3.3 and 3.4). However, according to the branching pattern of both TREEMIX and Neighbour-network graphs, the edges of the Anatolian and Egyptian populations split earlier than the Iranian ones, thus suggesting a relatively more recent origin of the latter. Such inconsistencies can be explained considering the different mode of inheritance of these markers, matrilinear for the mtDNA and biparental for the SNPs. Thus, starting from Nagarajan *et al.* (2015) hypothesis of an ancient origin of the mitochondrial variability of the Iranian populations, the similarity we found at the level of nuclear markers between the gene pools of Iranian, Anatolian and Egyptian populations can derive from a more recent and mainly male mediated gene flow. Alternatively, they may be due to a mere sampling effect: since Nagarajan *et al.* (2015) do not provide information on the sites of provenance of their Iranian samples, we cannot exclude that the observed differences mirror evolutionary events that have differentially affected the two sets of populations.

Similarly, according to TREEMIX graphs, the separation of the Mediterranean group seems to be a rather ancient event, but unfortunately, also in this case our results do not allow to precisely frame in a time perspective the evolutionary relationships between the population clades. Nevertheless, if we consider the overall geographical distribution of the different gene pools, it is evident that the present day pattern cannot be explained by a single migration wave originating from the Indian subcontinent and arriving to Europe and northern Africa, but rather seems to derive from a series of migration events occurred at different time and geographical

scales.

As pointed out by Zeuner (1963), the westward spread of river buffalo was probably slow, late and discontinuous. Therefore, we cannot exclude that the discontinuities in the gene pool distributions we observed may derive from at least two independent migration events: one more ancient wave that led the proto-Mediterranean gene pool through the Balkans to Italy, and a more recent wave bringing the proto-Middle eastern gene pool towards Mesopotamia and the Caspian sea and later followed by an expansion to Turkey and Egypt in conjunction with the spread of Islam.

Our evidence also show that the Italian Mediterranean and the population from Egypt belong to different gene pools, thus disproving the hypothesis reported in Cockrill (1974) that the Italian population may have derived from the introduction of Northern African buffaloes to southern Italy mediated by the Arabs.

Among the swamp buffalo populations considered here, our results clearly indicate the gene pool of those from Thailand and Indonesia as the most diverse and probably the most ancestral one: besides displaying the highest H_{obs} values (SWATH_THS H_{obs}=0.294 and SWAID_SUM H_{obs}=0.281; Table 3.2), in both the neighbour-network and the TREEMIX graph, SWATH and SWAID_SUM populations are placed on the edges closer to the midpoint/root. Furthermore, in ADMIXTURE bar plot (Figure 3.3) SWATH_THT, SWATH_THS and SWAID_SUM populations are shown to possess all the genomic components overall characterizing the swamp buffalo gene pool.

The other populations of the Indonesian islands (SWAID_NUT, SWAID_JAV and

SWAID_SUW) bear signs of geographical isolation, as indicated by the peripheral position and the small area occupied by their scatter of points in the MDS (Dimension one vs. three, Figure 3.2, right panel), by the long edges in the neighbour-network, and by the assignment to a well-defined cluster in ADMIXTURE analysis already at K=4 (Figure 3.3). Also the insular population from the Philippines SWAPH seems affected by geographical isolation; however, according to the general evidence (Figures from 3.2 to 3.4, Supplementary 3.8.4 and 3.8.6), its gene pool has closer similarities to that of the Chinese swamp buffaloes. Such relationship has already been revealed by microsatellite markers (Zhang *et al.* 2011) which highlighted that swamp buffaloes from South-eastern China—as are the populations included our sampling—have a closer similarity to those of the Philippines, compared to swamp buffaloes from southwestern China which were more similar to the rest of Indonesia. Furthermore, based on the clear separation of South-eastern Asian populations into two groups, the same authors suggested that, after domestication in southwestern China-northern Indochina, domesticated swamp buffaloes dispersal followed two different routes: one leading southward through peninsular Malaysia to the Indonesian islands of Sumatra, Java and Sulawesi, and a second leading towards north/northeast into Central China and then southwards through an insular route via Taiwan to the Philippines and Borneo.

Since our results generally agree with previously reported hypotheses on water buffalo domestication and post-domestication dispersal, to better highlight the patterns of molecular variation across the geographical area covered by our sampling, we calculated H_{obs} and H_{exp} after grouping the populations based on their geographical origin (Pakistan, Iran, Egypt, Anatolia, East Europe and Italy) and tested the significance of the differences between the

values following the approach of Skrbinšek *et al.* (2012), under the expectation of a decrease in genetic variability with increasing geographical distance from the centre of domestication (Groeneveld *et al.* 2010).

Even though the heterozygosity values could have been partially affected by ascertainment bias in the case of the Murrah, Nili-Ravi and Italian Mediterranean breeds due to their inclusion in the discovery panel, the evidence derived from our results fits well with the previously suggested origin and spread of domesticated water buffalo: after domestication in the Indian sub-continent, river buffalo populations migrated through South-western Asia and reached first Mesopotamia, and subsequently Egypt and Europe.

From their respective domestication centres, river buffaloes migrated west across south-western Asia, to Egypt, Anatolia and reached the Balkans and the Italian peninsula in the early Middle ages (VIIth cen. AD; Clutton-Brock 1999), while the swamp buffaloes likely dispersed South-westward to Thailand and Indonesia, and northward to central and eastern China (Zhang *et al.* 2016), wherefrom they further spread to the Philippines (Zhang *et al.* 2011).

3.6 Conclusions

Our results confirmed the utility of the Axiom® Buffalo Genotyping Array for the characterization of water buffalo breeds, even though its performance is likely reduced in the case of swamp-type or wild buffalo populations due to ascertainment bias. Nevertheless, when an adequate set of reference populations is available, this medium-density panel may allow to identify introgression and crossbreeding events between the two buffalo types, as shown in the

case of the admixed swamp *x* river buffalo population from the Philippines, or the Brazilian Carabao breed included in our dataset. Therefore, it may reveal useful to aid the implementation of marker-assisted breeding and inbreeding monitoring activities.

As for other livestock species, SNP data proved to be useful to assess the extent and geographical distribution of molecular diversity of domestic water buffalo, as well as to shed light on its domestication and post-domestication evolutionary history. In fact, our results largely confirmed previous archaeological, historical and molecular-based evidence on the existence of two different domestication sites for river- and swamp-type buffaloes, located in the Indo-Pakistani region and close to the border between China and Indochina, respectively. The subsequent diffusion out of the domestication centres seems to have followed two major divergent directions: river-type buffaloes apparently spread along a western route, while swamp buffaloes along an East-South-eastern route. To conclude, our and previous findings seem to suggest the present-day distribution of water buffalo diversity to derive from the combined effects of migration events occurred at different stages of the post-domestication evolution of the species.

3.7 Acknowledgements

We thank Elisa Eufemi for the help provided in screening the scientific literature and other written sources.

3.8 Supplementary information

3.8.1 Comparison of individual observed heterozygosity values

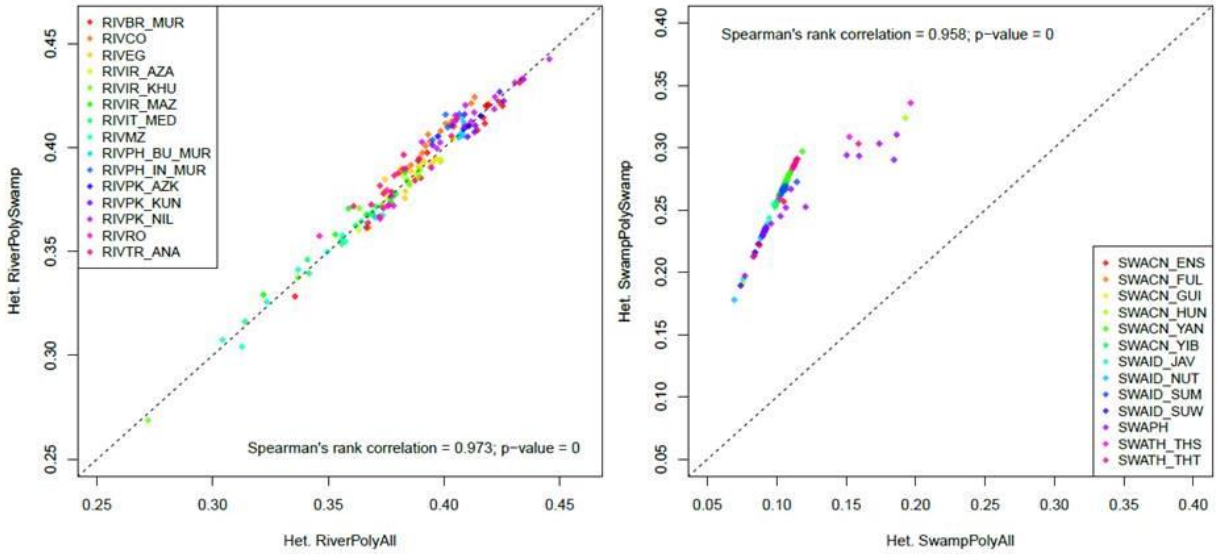

Figure 3.5 Comparison of individual observed heterozygosity values obtained when the whole set of markers (*X*-axis) and the set of markers polymorphic in swamp populations (*Y*-axis) were used. River populations are represented in the left panel, while swamp populations are in the right panel.

3.8.2 Comparison of average heterozygosity per population

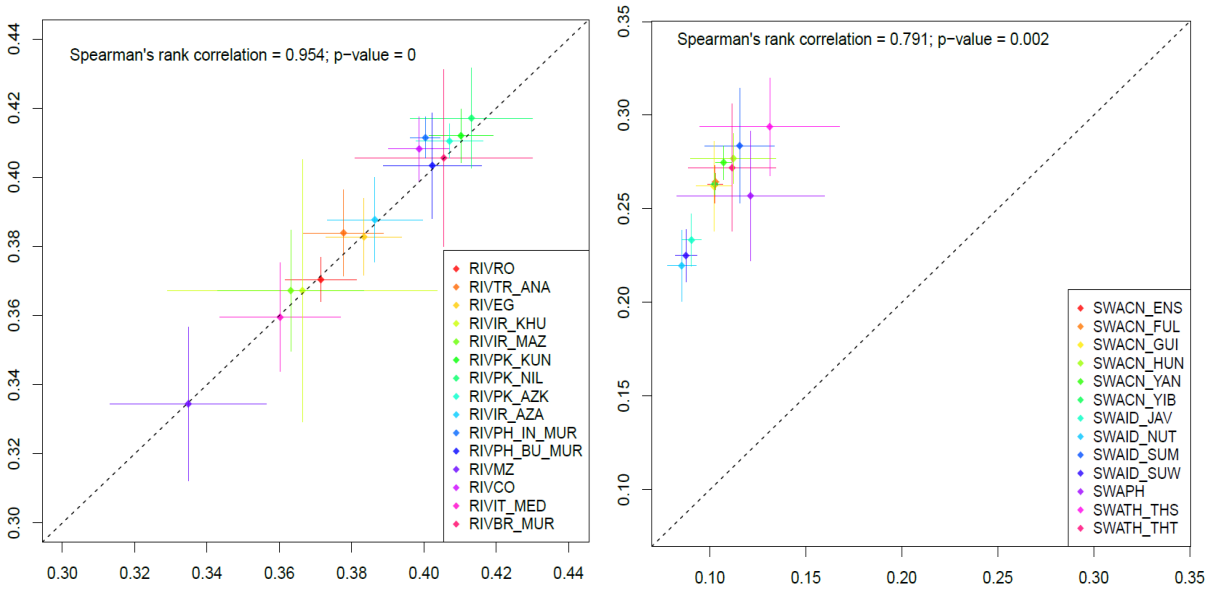

Figure 3.6 Comparison of population average observed heterozygosity values obtained when the whole set of markers (*X*- axis) and the set of markers polymorphic in swamp populations (*Y*-axis) were used. River populations are represented in the left panel, while swamp populations are in the right panel.

3.8.3 F_{ST} values and number of migrants

Table 3.4 F_{ST} values and number of migrants as estimated from ARLEQUIN and JAATHA. Rows' and columns' headers refer to the numerical code presented in Table 3.1. Estimated gene flow and F_{ST} vales are presented in the upper- and lower-diagonal matrix, respectively.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
2	–	0.093	43.490	0.591	0.398	65.338	39.216	74.142	24.609	48.007	0.031	0.087	0.152	75.000	57.307
3	0.070	–	0.769	0.010	0.777	0.236	0.380	0.010	0.010	0.123	0.125	0.496	0.016	0.312	0.012
4	0.097	0.153	–	75.000	46.534	75.000	30.729	39.319	38.657	19.351	0.156	0.326	47.127	55.706	59.268
5	0.106	0.160	0.079	–	75.000	75.000	75.000	60.005	75.000	75.000	75.000	0.010	75.000	75.000	75.000
6	0.099	0.148	0.064	0.018	–	75.000	75.000	60.005	65.875	40.370	0.221	75.000	75.000	75.000	60.005
7	0.100	0.144	0.088	0.031	0.046	–	75.000	75.000	75.000	74.456	75.000	0.016	75.000	75.000	75.000
8	0.104	0.149	0.081	0.029	0.043	0.050	–	75.000	75.000	75.000	48.007	0.024	75.000	75.000	75.000
9	0.120	0.169	0.097	0.049	0.057	0.062	0.040	–	75.000	60.005	33.560	0.012	75.000	75.000	75.000
10	0.120	0.170	0.098	0.041	0.054	0.058	0.013	0.047	–	75.000	35.613	0.505	75.000	75.000	75.000
11	0.129	0.179	0.108	0.052	0.063	0.066	0.023	0.055	0.021	–	43.070	1.130	70.198	75.000	75.000
12	0.146	0.199	0.126	0.072	0.083	0.083	0.044	0.075	0.038	0.045	–	0.010	26.833	48.007	75.000
13	0.114	0.174	0.089	0.006	0.027	0.028	0.027	0.045	0.038	0.049	0.069	–	0.016	0.024	0.010
14	0.116	0.159	0.092	0.021	0.040	0.039	0.041	0.053	0.048	0.057	0.076	0.014§	–	75.000	75.000
15	0.111	0.153	0.088	0.015§	0.035	0.037	0.039	0.051	0.046	0.055	0.071	0.012§	0.023	–	75.000
16	0.083	0.128	0.071	0.006	0.024	0.021	0.030	0.046	0.040	0.051	0.067	0.010	0.024	0.020	–
17	0.355	0.413	0.358	0.337	0.326	0.304	0.311	0.325	0.334	0.339	0.356	0.345	0.311	0.299	0.300
18	0.181	0.226	0.165	0.106	0.121	0.118	0.125	0.137	0.133	0.141	0.155	0.104	0.112	0.108	0.106
19	0.319	0.379	0.325	0.307	0.296	0.274	0.283	0.298	0.305	0.310	0.329	0.315	0.285	0.272	0.271
20	0.310	0.370	0.307	0.273	0.270	0.254	0.263	0.277	0.281	0.287	0.306	0.281	0.258	0.250	0.247

21	0.325	0.383	0.324	0.295	0.289	0.271	0.280	0.293	0.299	0.304	0.323	0.304	0.277	0.267	0.266
22	0.356	0.414	0.359	0.340	0.329	0.308	0.315	0.328	0.337	0.342	0.359	0.348	0.316	0.304	0.304
23	0.358	0.415	0.361	0.341	0.331	0.309	0.316	0.330	0.339	0.344	0.361	0.350	0.317	0.305	0.305
24	0.346	0.405	0.347	0.326	0.317	0.295	0.303	0.316	0.325	0.329	0.348	0.335	0.303	0.291	0.291
25	0.349	0.404	0.350	0.329	0.320	0.299	0.307	0.319	0.328	0.332	0.349	0.336	0.307	0.296	0.296
27	0.347	0.405	0.349	0.328	0.317	0.296	0.304	0.317	0.326	0.330	0.348	0.336	0.304	0.292	0.293
28	0.357	0.413	0.359	0.339	0.329	0.308	0.315	0.328	0.337	0.341	0.360	0.348	0.316	0.303	0.303
29	0.381	0.448	0.389	0.376	0.358	0.330	0.339	0.351	0.365	0.369	0.389	0.389	0.343	0.327	0.328
30	0.373	0.444	0.379	0.365	0.346	0.319	0.328	0.342	0.354	0.358	0.380	0.381	0.330	0.316	0.315
31	0.335	0.393	0.336	0.312	0.304	0.283	0.291	0.304	0.312	0.317	0.335	0.321	0.290	0.279	0.279
32	0.376	0.443	0.382	0.368	0.350	0.323	0.331	0.345	0.358	0.362	0.382	0.382	0.335	0.320	0.320

Second half of the table.

	17	18	19	20	21	22	23	24	25	27	28	29	30	31	32
2	0.010	0.288	0.016	0.163	0.020	0.169	0.261	0.274	0.074	0.064	0.117	0.016	0.073	0.152	0.144
3	0.016	0.100	0.100	0.296	0.024	0.085	0.057	0.057	0.245	0.012	0.382	0.073	0.012	0.031	0.024
4	0.010	0.019	0.031	0.012	0.243	0.024	0.249	0.024	0.038	0.012	0.012	0.157	0.010	0.100	0.016
5	0.171	75.000	0.626	0.010	6.611	0.085	0.010	0.101	0.020	0.138	0.340	0.196	0.016	0.010	0.306
6	0.194	27.766	0.010	0.010	0.169	0.029	0.020	0.119	0.147	0.350	0.138	0.020	0.010	0.010	0.016
7	0.068	41.436	0.225	0.012	0.119	0.010	0.034	0.295	0.086	0.074	0.016	0.031	0.010	0.108	0.125
8	0.086	75.000	0.092	0.010	0.010	0.197	0.010	0.574	0.086	0.537	0.180	0.012	0.033	0.010	0.010
9	0.016	75.000	0.119	0.016	0.224	0.129	0.174	2.719	0.064	0.138	0.129	0.010	0.016	0.103	0.079
10	0.024	19.213	0.101	0.016	0.083	0.518	0.167	0.010	0.142	0.178	0.211	0.306	1.256	0.078	0.306
11	0.242	23.595	0.016	0.153	0.209	0.025	0.020	0.031	0.241	0.210	0.172	0.020	0.012	0.124	0.341
12	0.188	0.371	0.057	0.010	1.479	0.104	0.016	0.010	0.229	0.039	0.016	0.024	0.026	0.022	0.306
13	0.076	0.328	0.371	0.010	0.010	0.187	0.016	0.430	0.115	0.024	0.102	0.010	0.024	0.010	0.024

14	0.309	75.000	0.145	0.200	0.012	0.010	0.031	0.010	0.125	0.020	0.114	0.072	0.053	0.248	0.010
15	0.137	75.000	0.184	7.071	0.010	0.170	0.010	0.178	0.087	0.016	0.168	0.043	0.033	0.010	0.024
16	0.080	75.000	0.256	4.451	0.010	0.754	0.338	0.275	0.016	0.016	0.172	0.010	0.024	0.104	0.020
17	–	0.804	0.010	0.131	2.179	0.033	0.234	0.025	0.131	0.341	0.021	0.064	0.020	0.065	0.095
18	0.117	–	0.373	0.271	0.289	0.371	0.024	0.103	0.012	0.247	0.074	0.016	0.048	0.010	0.141
19	0.194	0.142	–	0.525	0.010	0.332	0.069	0.010	0.010	0.349	0.031	0.067	0.016	0.079	0.062
20	0.077	0.065	0.134	–	75.000	0.039	0.262	31.774	28.418	28.234	0.210	0.226	0.258	0.150	0.160
21	0.081	0.080	0.141	0.008§	–	0.022	0.494	75.000	26.741	39.401	27.836	0.080	0.237	0.033	0.086
22	0.081	0.097	0.170	0.037	0.041	–	75.000	75.000	70.575	75.000	75.000	1.569	0.040	0.010	1.545
23	0.080	0.101	0.171	0.040	0.043	0.009	–	75.000	75.000	75.000	75.000	0.096	0.262	0.158	0.020
24	0.072	0.087	0.162	0.028	0.032	0.007	0.009	–	75.000	62.803	75.000	0.227	0.030	0.147	0.177
25	0.069	0.088	0.160	0.027	0.030	0.013	0.016	0.007	–	75.000	51.880	0.126	0.020	0.134	0.063
27	0.068	0.090	0.161	0.032	0.037	0.019	0.023	0.015	0.010	–	75.000	0.093	0.082	0.216	0.024
28	0.076	0.097	0.168	0.036	0.040	0.008	0.008	0.004	0.011	0.018	–	0.175	0.087	0.103	0.020
29	0.166	0.159	0.223	0.106	0.108	0.130	0.129	0.122	0.120	0.125	0.127	–	49.050	0.165	21.612
30	0.174	0.157	0.232	0.116	0.118	0.140	0.138	0.131	0.131	0.135	0.136	0.034	–	0.030	34.024
31	0.121	0.109	0.169	0.050	0.053	0.088	0.090	0.079	0.077	0.081	0.085	0.132	0.142	–	1.220
32	0.166	0.156	0.222	0.107	0.109	0.132	0.130	0.123	0.122	0.127	0.129	0.050	0.045	0.128	–

§ refers to the non-significant F_{ST} tests.

3.8.4 JAATHA heat map

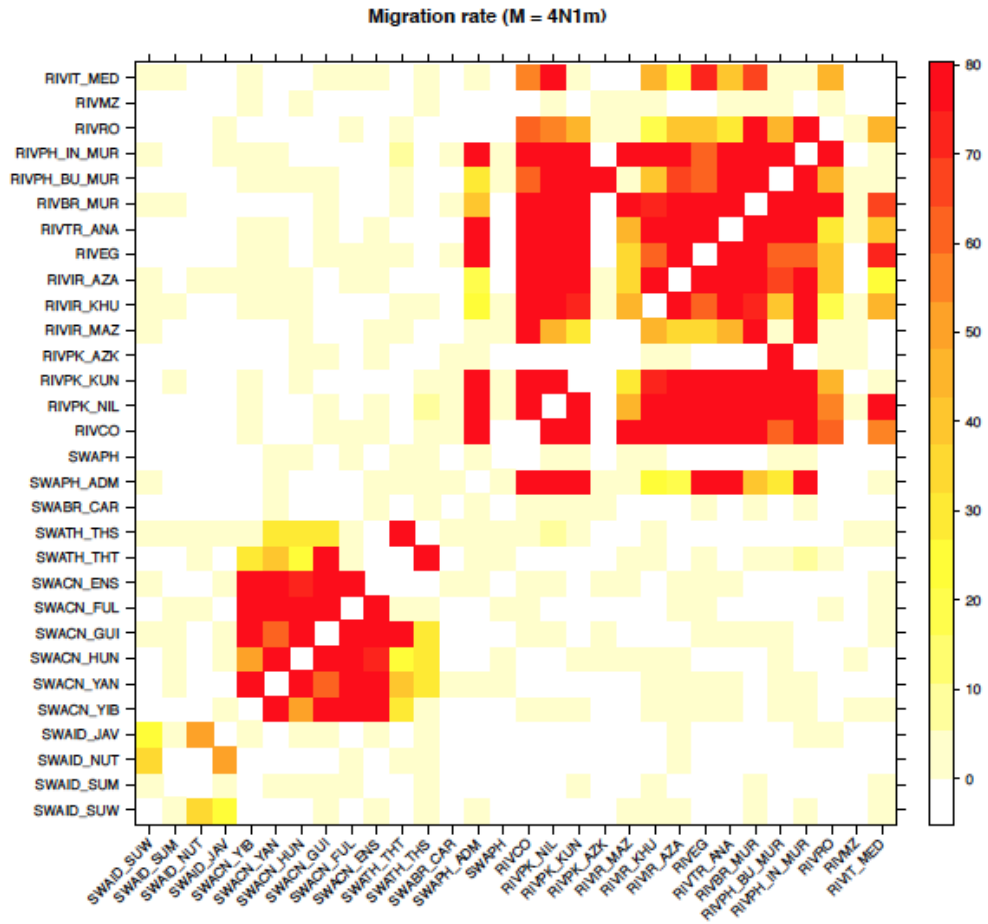

Figure 3.7 Heat map depicting the estimated number of migrants between each pair of populations.

3.8.5 Multi-dimensional Scaling analysis: eigenvalues

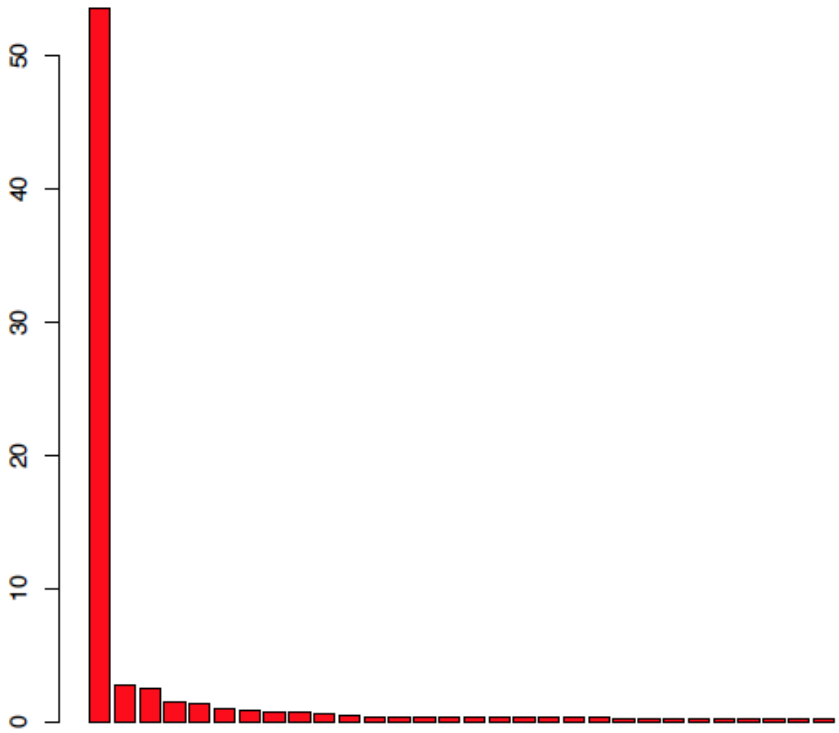

Figure 3.8 Bar plot of the eigenvalues corresponding to the first 30 dimensions of the Multi-Dimensional Scaling plot shown in Figure 3.2.

3.8.6 Neighbour-net analysis

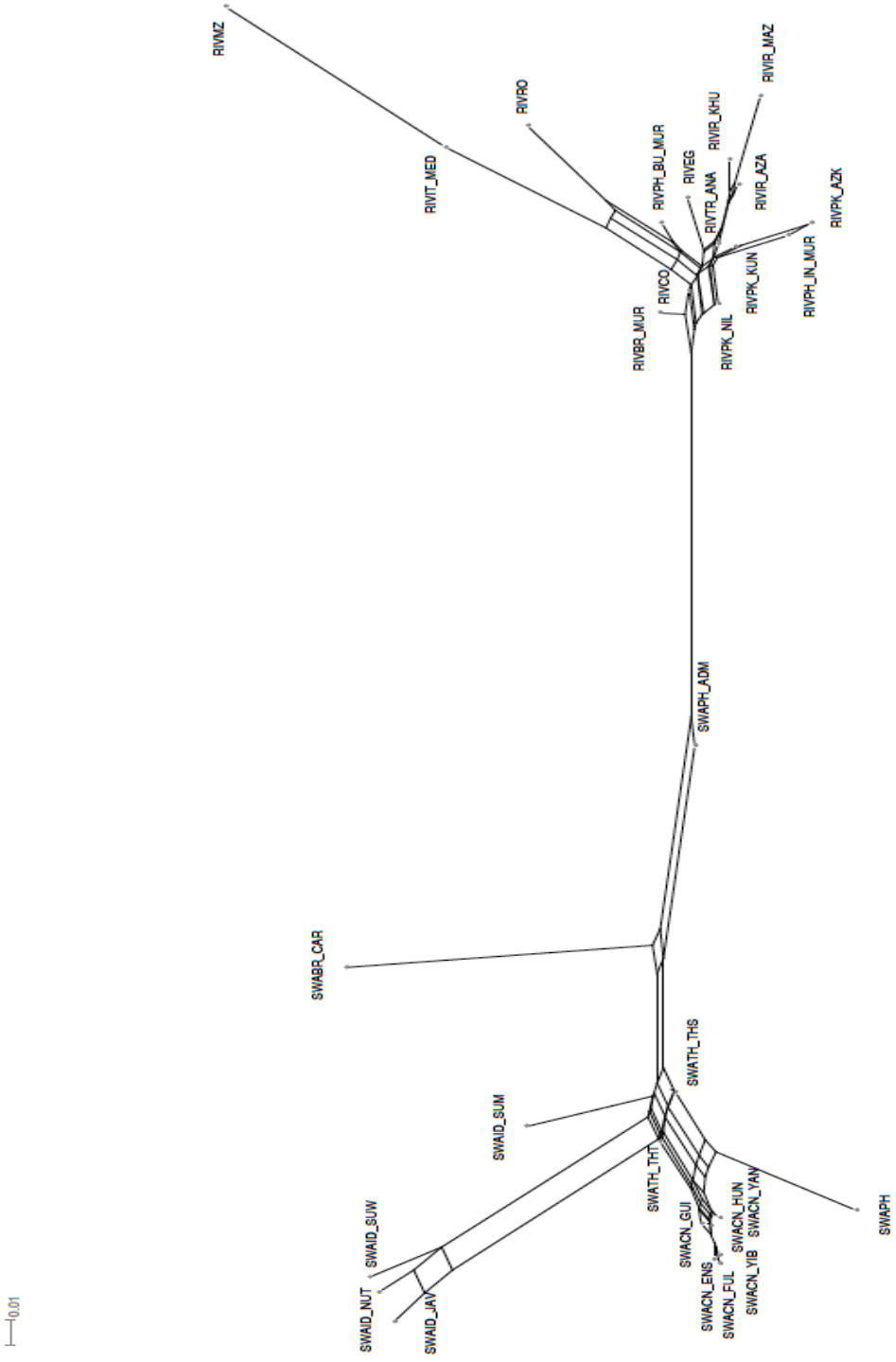

Figure 3.9 Neighbour-network based on a matrix of Reynolds genetic distances between breeds.

3.8.7 ADMIXTURE analysis: graphical representation

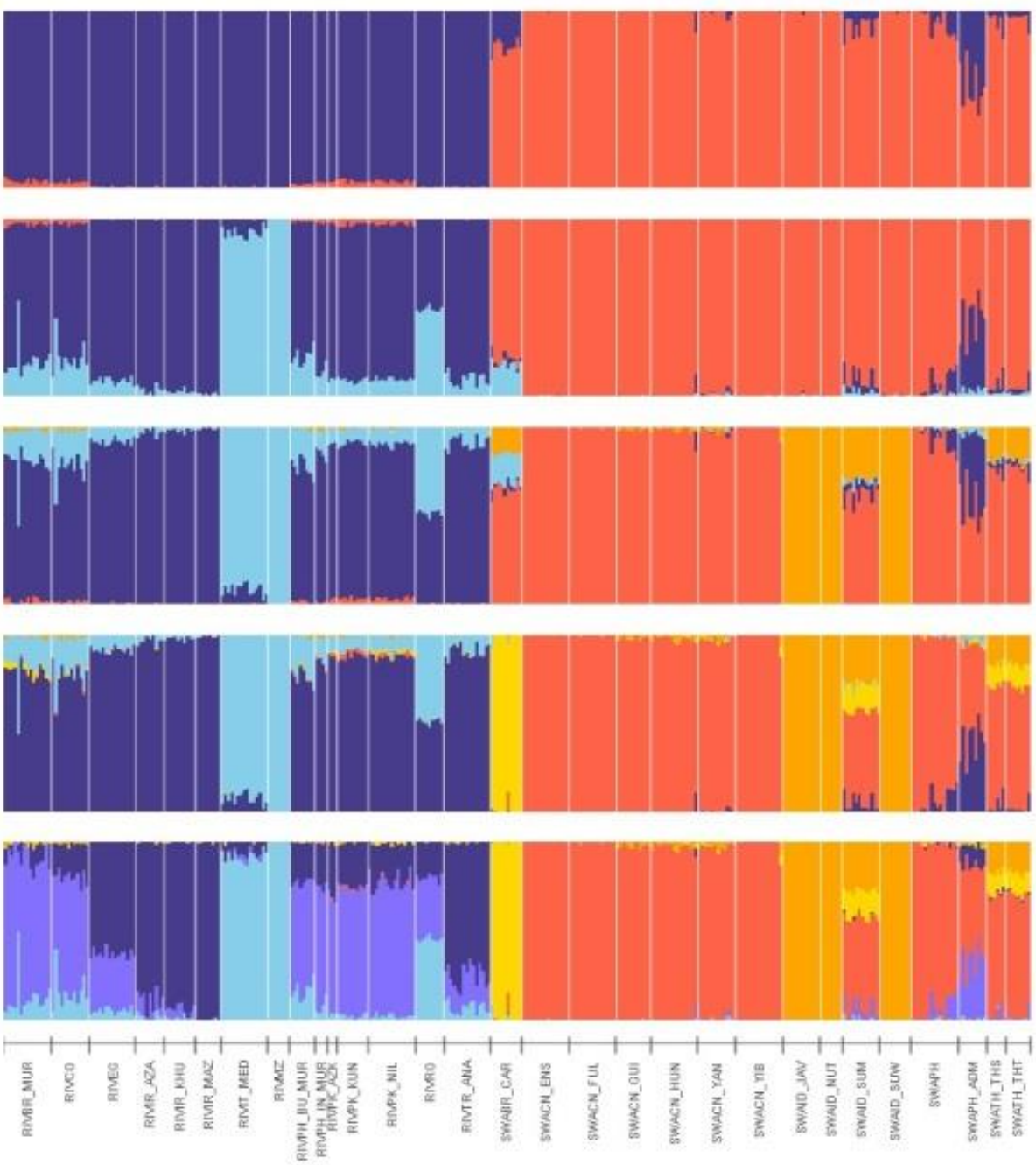

Figure 3.10 ADMIXTURE bar plots from K=2 (upper figure) to K=6 (lower figure).

3.8.8 ADMIXTURE analysis: selection of the clustering solution

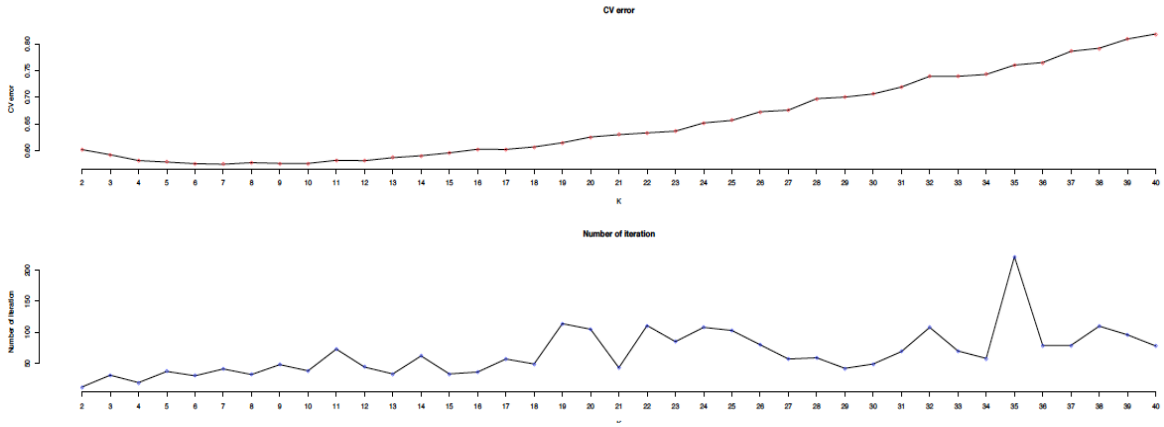

Figure 3.11 Upper panel: Cross-Validation error for any given cluster solution tested (from K=2 to K=40). Lower panel: number of iterations required to reach model convergence in any cluster solution tested.

3.8.9 TREEMIX: fraction of variance in relatedness between population explained

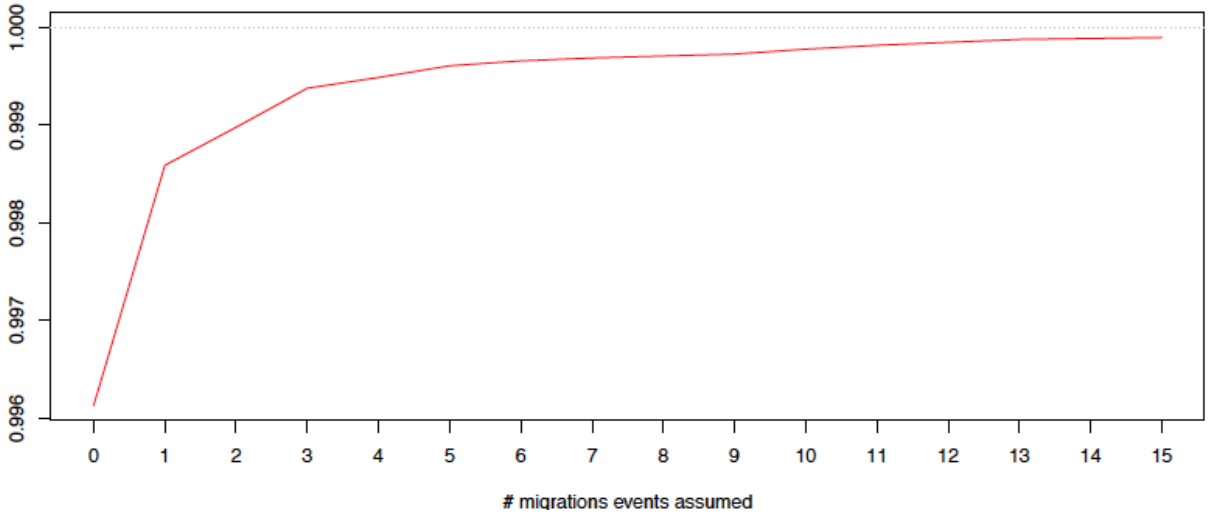

Figure 3.12 Fraction of variance in relatedness between populations explained for each tested model, from a tree with zero migration, to a graph with migration edges assumed. The fraction of variance was estimated following equation 30 in Pickrell & Pritchard (2012).

3.8.10 TREEMIX: results

Table 3.5 TREEMIX results for any tested model, from zero to 15 migration assumed. The variance accounted for each tested model (Var. expl.) is also plotted in Supplementary 3.8.9. The significance of each migration was computed, and the percentage of significant migrations in every model is reported (Perc.). For each tested model, the log-likelihood of the starting tree (log-lik m_0) and of the graph with the migration edges added (log-lik m_i) are reported.

m	Var. expl.	Perc.	log-lik m_0	log-lik m_i
0	0.99613	0	-4501.40	-4501.4
1	0.99859	100	-4501.40	2505.34
2	0.99898	100	-4501.40	2774.92
3	0.99938	100	-4501.40	2868.87
4	0.99949	100	-4495.76	2930.10
5	0.99961	100	-4495.23	2993.64
6	0.99966	100	-4495.76	3027.53
7	0.99969	100	-4495.23	3044.18
8	0.99971	100	-4495.76	3054.32
9	0.99973	100	-4495.23	3065.24
10	0.99978	100	-4495.76	3075.99
11	0.99982	100	-4495.23	3102.32
12	0.99985	100	-4495.76	3117.95
13	0.99988	100	-4501.40	3151.13
14	0.99989	92.86	-4495.23	3158.32
15	0.99990	93.33	-4501.40	3164.71

3.8.11 TREEMIX: residuals of m_5 model

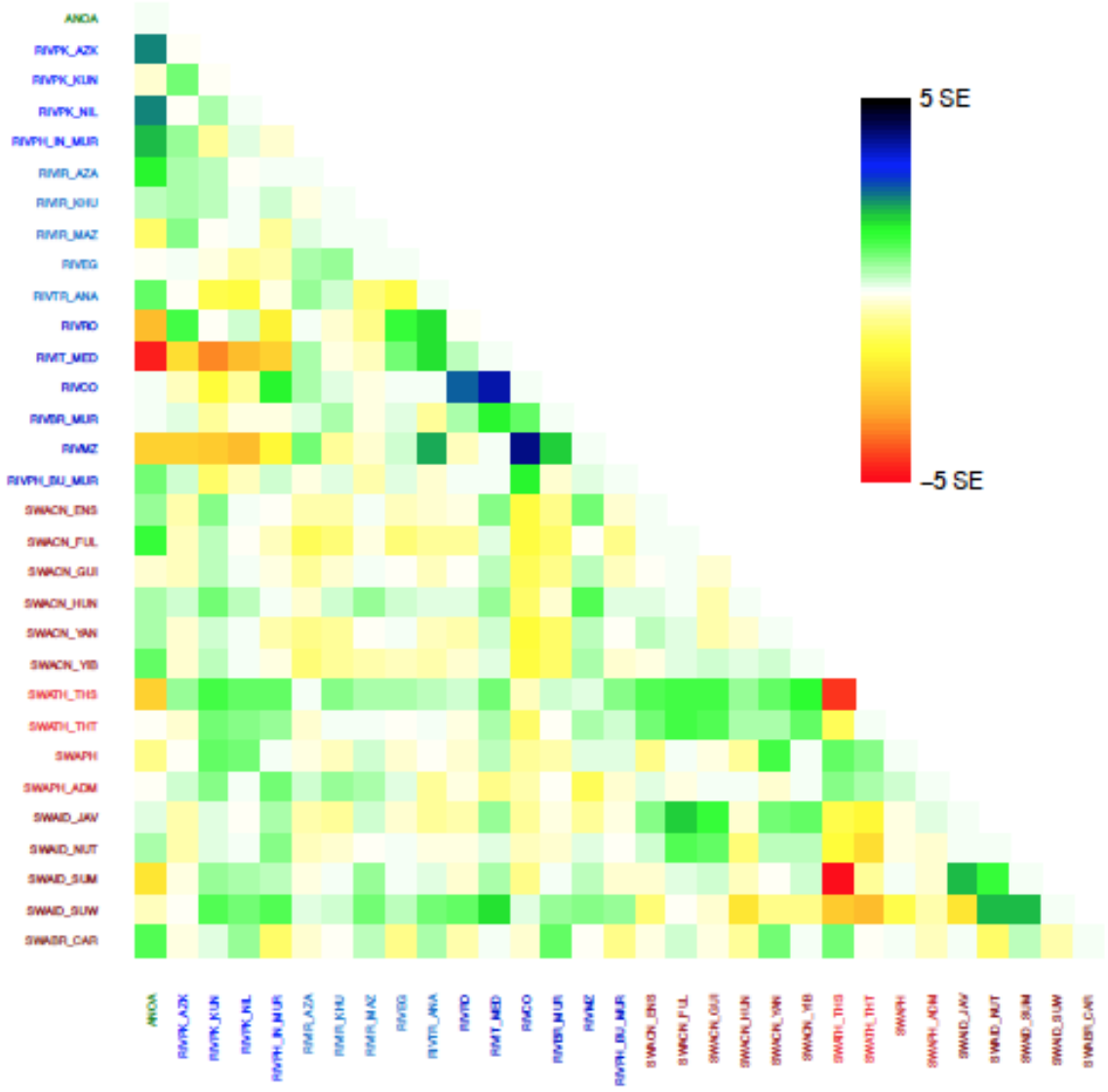

Figure 3.13 Heat map of the residuals of the m_5 model. Positive values (green to blue colours) indicate pairs of populations candidate to be linked by a migration edge (i.e. where the addiction of a migration edge could improve model fitting).

4. Combining landscape genomics and ecological modelling to investigate local adaptation of indigenous Ugandan cattle to East Coast Fever

Elia Vajana, Mario Barbato, Licia Colli, Marco Milanesi, Estelle Rochat, Enrico Fabrizi, Christopher Mukasa, Marcello Del Corvo, Charles Masembe, Vincent Muwanika, Fredrick Kabi, Riccardo Negrini, Stéphane Joost* & Paolo Ajmone-Marsan*, and the NEXTGEN Consortium.

*Co-senior authorship

4.1 Abstract

East Coast Fever (ECF) is a fatal sickness affecting cattle populations of Central and Eastern Africa. The disease is caused by the protozoan *Theileria parva parva*, transmitted by the hard-bodied tick *Rhipicephalus appendiculatus*. Indigenous herds, however, show tolerance to infection in ECF-endemically stable areas. Here, we investigated the postulated genetic bases underlying local adaption to *T. parva parva* by relying on molecular data and epidemiological information from 823 indigenous cattle from Uganda. *R. appendiculatus* potential distribution and *T. parva parva* infection risk were first estimated over the study area and subsequently tested into a genotype-environment association (GEA) analysis. The study found forty-one and seven candidate adaptive loci for tick burden and *T. parva parva* infection, respectively. Two genes were identified as putatively involved into local adaptation for ECF: PRKG1 and SLA2. The first was already described as associated with tick resistance in indigenous South African

cattle, possibly due to its role into inflammatory response. The latter is part of the regulatory pathways involved into lymphocytes' proliferation, which are known to be modified by *T. parva parva* infection. Finally, a preliminary investigation of the ancestral origin of the genomic regions candidate for ECF adaptation revealed a mixed African sanga and zebuine ancestry for PRKG1 region, and a prevalent sanga origin for SLA2 region.

Keywords: Indigenous cattle, *Theileria parva parva*, *Rhipicephalus appendiculatus*, East Coast Fever, Uganda, species distribution modelling, local adaptation, landscape genomics.

4.2 Introduction

East Coast Fever (ECF) is an endemic vector-borne disease affecting cattle populations of eastern and central Africa. The etiological agent of the disease is the emo-parasite protozoan *Theileria parva* Theiler, 1904, transmitted by the hard-bodied tick vector *Rhipicephalus appendiculatus* Neumann, 1901. ECF causes high mortality rates among exotic breeds and crossbreds, and reduces indigenous cattle productivity (Norval *et al.* 1992; Olwoch *et al.* 2008; Muhanguzi *et al.* 2014), consequently undermining the development of the livestock sector in affected countries.

African Cape buffalo (*Syncerus caffer* Sparrman, 1779) is believed to be *T. parva* native host, as well as its wild and asymptomatic reservoir (Oura *et al.* 2011). A primordial contact between buffalo-derived *T. parva* and domestic bovines is likely to have taken place around 4,500 years before present (YBP) (Epstein 1971). However, no consensus has been reached so far in establishing the migration date of *Bos taurus* and *Bos indicus* into ECF endemic regions

(Freeman 2006; Magee *et al.* 2014; Mwai *et al.* 2015), and therefore in determining if such "host jump" affected taurine or indicine cattle first. African taurine cattle represent the most ancient gene pool of the continent, and may have reached eastern Sub-Saharan regions between the large time span comprised between 8,000 and 1,500 YBP (Magee *et al.* 2014; Mwai *et al.* 2015). Conversely, the first zebuine colonization wave from the Far East is estimated of having occurred around 4,000-2,000 YBP, as suggested by the first certain archaeological record dated 1,750 YBP (Freeman 2006).

Plausibly, the first transmission of buffalo-derived *T. parva* to domestic bovines was mediated by infected ticks. Cattle-specific adaptations subsequently led to the differentiation at the genetic level between buffalo- and cattle-derived parasite strains: *T. parva lawracei* and *T. parva parva*, respectively (Hayashida *et al.* 2013; Sivakumar *et al.* 2014).

For centuries tropical diseases represented a barrier to livestock migration towards African southern regions (Hanotte *et al.* 2002). The coexistence of parasite and domestic host might have resulted in local adaptation, leading the indigenous livestock populations to coevolve with the parasite and develop a natural tolerance to the disease (Kabi *et al.* 2014; Bahbahani & Hanotte 2015). Innate tolerance, environmental conditions causing a constant tick challenge, and higher chances of being infected during the first months of life—when calves are protected by colostrum-derived immunity—are all factors believed to contribute to establish "endemic stability", an epidemiological condition in which a clinical disease manifests at negligible levels, and the host becomes an asymptomatic parasite carrier (Kivaria *et al.* 2004).

ECF endemically stable areas are currently populated by zebu, sanga and zenga breeds, Sanga are the result of crossbreeding between African *B. taurus* and *B. indicus* occurred *c.* 2,000

YBP (Hanotte *et al.* 2002), whereas zenga are a sanga *x* zebu backcross (Mwai *et al.* 2015). Therefore, at least three evolutionary hypotheses can explain the adaptive component of ECF endemic stability: (i) genetic adaptation appeared in the native African *B. taurus* at first and was subsequently introgressed into zebu and derived sanga and zenga crossbreds; (ii) adaptation first appeared in *B. indicus*, and either evolved independently on the African continent, or was imported from the Indian continent, where similar selective pressures are recorded (Singh *et al.* 1993; Boulter & Hall 1999), thus supporting the hypothesis of an indicine-derived tolerance in the local crossbreds; (iii) adaptive responses evolved in more recent times, after the appearance of sanga and zenga breeds.

Among the components that may have contributed to shape adaptive variation against infection is a specific immune response targeting the parasite, along with genetic adaptation conferring resistance towards the ECF vector, *R. appendiculatus*.

Specific areas within current Ugandan borders, e.g. some regions in the South-West and in the East of the country, are proved to be endemically stable for ECF (Kivaria *et al.* 2004; Rubaire-Akiiki *et al.* 2006; Kabi *et al.* 2014), thus making Uganda a candidate area for studying the putative genetic basis underlying ECF local adaptation. Local adaption is expected to evolve in a context of "ongoing or very strong recent spatially varying selection" (Savolainen *et al.* 2013), especially when demes are connected by gene flow (Kawecki & Ebert 2004). Further, it has been observed to evolve over short time scales—from thousands of years to few decades—in several animal and plant species (Stockwell *et al.* 2003; Crispo *et al.* 2010; Fraser *et al.* 2011). Such requirements are all likely to be met in Uganda, where the abrupt climate changes occurred during the middle and recent Holocene (Kiage & Liu 2006) plausibly

affected ECF vector distribution (Cumming. 2002), and where the genetic makeup of indigenous cattle populations has been proved to be highly admixed for the same ancestral components (Stucki *et al.* 2016).

Therefore, the present study aims to: (i) define genomic regions associated with ECF selective pressure in indigenous cattle populations from Uganda, and (ii) to provide a first reconstruction of the ancestral origin of such genomic regions.

Here, a comprehensive approach encompassing both ECF determinants (i.e. occurrence of the tick vector and the parasite) was developed in order to identify ECF related selection signals in Ugandan cattle. Firstly, tick and parasite selective pressures on indigenous cattle genomes were estimated by means of spatial modelling techniques; secondly, selective pressures were tested by a genotype-environment association approach relying on landscape genomics models to highlight target genomic regions and genes candidate for selection. Lastly, the ancestral origin of the adaptive genomic regions was investigated by local ancestry inference, to shed light about the evolutionary origins of ECF tolerance.

4.3 Materials and Methods

4.3.1 Species and infection distribution models

Spatial patterns of *R. appendiculatus* occurrence probability (ψ_R) and *T. parva parva* infection risk in cattle (γ) were modelled throughout Uganda and subsequently employed as predictors into landscape genomics models. Here, spatial variation in both γ and ψ_R was

assumed to describe the spatially heterogeneous selective pressure on indigenous host genomes.

As spatial proximity between Cape buffaloes and cattle herds may constitute a valuable factor for explaining *T. parva parva* infection incidence in livestock (Kabuusu *et al.* 2013), *S. caffer* occurrence probability (ψ_S) was estimated prior to environmental association analyses and used in combination with ψ_R to model γ over the study area.

Typically, ψ is estimated by means of species distribution modelling techniques. Species distribution models (SDMs) attempt to estimate the realized niche of species (*sensu* Hutchinson), by quantifying the relationship between a set of observed presences—and, whenever available, absences—and the habitat features taken into consideration at the sampling sites (Guisan & Thuiller 2005).

The following four sections will describe data and methods used to estimate ψ_R, ψ_S and γ over the study area.

4.3.2 Raster data

Bioclimatic variables (BIO) used to estimate *T. parva parva* infection risk as well as *R. appendiculatus* and *S. caffer* potential distributions were retrieved from the WorldClim database (Hijmans *et al.* 2005; v.1.4. release3), in the un-projected latitude/longitude coordinate reference system (WGS84 datum), at a resolution of 30 arc-seconds (around 1 km^2 at the equator), and for current conditions (i.e. corresponding to 1960-1990 decades).

Altitude information was retrieved from the SRTM 90m Digital Elevation Database (Jarvis *et*

al. 2008; v.4.1), which provides 5×5 degree tiles covering Earth's land surface in the latitude/longitude coordinate reference system (WGS84 horizontal datum, EGM96 vertical datum), at 90 m^2 resolution at the equator. The four tiles covering Uganda were downloaded and merged, and the obtained raster file served to compute the terrain slope through the function `terrain` implemented in the R package `raster` (Hijmans 2016).

A ten-years (2001-2010) averaged Normalized Difference Vegetation Index (NDVI) was derived from the "eMODIS products" of the U.S. Geological Survey (Swets *et al.* 1999; http://earlywarning.usgs.gov/fews/product/116), in un-projected latitude/longitude coordinates (WGS84 datum), and at a resolution of 250 m^2 at the equator.

Cattle density was acquired from the Livestock Geo-Wiki (Robinson *et al.* 2014; http://www.livestock.geo-wiki.org), in un-projected latitude/longitude coordinates with datum WGS84, and at 1 km^2 resolution at the equator.

Furthermore, a raster file describing each pixel distance from the nearest water source was obtained with the function `distance` of the `raster` package. The 'Land and Water Area' dataset (CIESIN 2016) from the Gridded Population of the World collection (GPV v.4) was used to define water bodies in Uganda at a resolution of 30 arc-seconds, in un-projected latitude/longitude coordinates with WGS84 datum.

Raster files were transposed into Africa Albers Equal Area Conic projection (EPSG:102022) to meet the main assumption of the SDM technique used to model ψ_R and ψ_S (see sections 4.3.3 and 4.3.4), i.e. that each pixel of the landscape presents the same probability to be randomly sampled to detect the species occurrence (Merow & Silander 2014). Subsequently,

raster files were standardised to have the same resolution (0.85 km^2), origin and extent. Water surfaces (e.g. lakes) were masked prior to the analyses to avoid the inclusion of background locations potentially misleading for characterizing the occurrence probability of terrestrial species (Barve *et al.* 2011). Quantum GIS (v.2.16.2) (QGIS Development Team 2016) and `raster` were used for carrying out operations on raster files.

4.3.3 *Rhipicephalus appendiculatus* distribution model: ψ_R estimation

Extensive studies on the environmental drivers affecting African tick species distribution have identified the interaction between temperature and precipitation as the main explanatory factor at broad geographical scales (Cumming 1999a, 2002). Therefore, BIO variables representing temperature, precipitation and temperature/precipitation interaction in the most extreme periods of the year, and thus likely acting as limiting factors for tick distribution, were chosen to be tested as environmental covariates of *R. appendiculatus* occurrence (Table 4.1 and Supplementary 4.7.1).

Table 4.1 Chosen environmental variables for *R. appendiculatus* distribution model.

Bioclim name	Description
BIO_8	Mean temperature* of the wettest three months (quarter) of the year
BIO_9	Mean temperature of the driest quarter
BIO_{10}	Mean temperature of the warmest quarter
BIO_{11}	Mean temperature of the coldest quarter
BIO_{16}	Precipitation* of the wettest quarter
BIO_{17}	Precipitation of the driest quarter
BIO_{18}	Precipitation of the warmest quarter
BIO_{19}	Precipitation of the coldest quarter

*Temperature was transformed from dC° to C° prior analyses; precipitation is in millimetres.

Due to the strong collinearity among the selected covariates, a principal components analysis (PCA) was performed using the R function `prcomp` (R Core Team 2016) to obtain independent synthetic environmental covariates. Univariate SDMs were devised to test the effect of each synthetic covariate, while multivariate SDMs were used to test different combinations of the synthetic covariates accounting for the greatest amount of variance in the original environmental dataset (see section 4.4.1). The performances of the devised SDMs were compared on the basis of the Bayesian information criterion (BIC) (Aho *et al.* 2014).

To ensure that the relative importance of each tested variable was described by the size of its standardized regression coefficient (Cade 2015), Bring's standardization (Bring 1994) was applied to the environmental variables prior to the analysis.

Fifty-one *R. appendiculatus* presence-points in Uganda were retrieved from the African ticks occurrence database collected by Cumming (1999b) (Figure 4.2). The R package `Maxlike` (Royle *et al.* 2012) was used to model ψ_R spatial distribution over Uganda. `Maxlike` is able to estimate ψ from presence-only (PO) data, by maximizing the likelihood of occurrences under the following logit-linear model (Royle *et al.* 2012):

$$ln\left(\frac{\psi_{Rx}}{1-\psi_{Rx}}\right) = \beta_0 + \boldsymbol{\beta} z(x)$$

where ψ_{Rx} denotes the tick occurrence probability in the x pixel of the landscape, β_0 the model intercept (i.e. the expected species prevalence across the study area), $\boldsymbol{\beta}$ the vector of slope parameters related to the considered environmental covariates, and $z(x)$ the vector containing the environmental variables for x. Tick occurrence probability in x can be derived from the inverse logit:

$$\psi_{Rx} = \frac{e^{\beta_0 + \boldsymbol{\beta} z(x)}}{1 + e^{\beta_0 + \boldsymbol{\beta} z(x)}}$$

In order to quantify model uncertainty, the delta method was used to compute the standard error (SE) and the 95% confidence intervals around each fitted ψ_{Rx}. Custom R functions were written to perform both Bring's standardization and SE computation.

4.3.4 *Syncerus caffer* distribution model: ψ_S estimation

The environmental variables affecting Cape buffalo potential distribution were identified according to the literature. Specifically, terrain slope (Matawa *et al.* 2012), NDVI (Pettorelli *et al.* 2011; Matawa *et al.* 2012), distance to water sources (Naidoo *et al.* 2012; Matawa *et al.* 2012), and annual precipitation (Naidoo *et al.* 2012) were identified as potential physical drivers of buffalo distribution, and were therefore acquired to predict ψ_S over Uganda. Since NDVI variable was distributed into 72 annual periods, a regression analysis was conducted to identify the period of the year mostly associated with the available *S. caffer* occurrences. Akaike information criterion (AIC) was used to select the best regression model, and the time span between April 6-15 was finally retained for subsequent analyses (Supplementary 4.7.2). In addition, altitude was also considered to account for the potential effect of elevation, by providing a total of five environmental predictors of Cape buffalo occurrence (Table 4.2).

Table 4.2 Considered environmental variables for ψ_S estimation.

Environmental variable
Slope
Altitude

BIO_{12}*
NDVI**
Distance from water (Wd)

*Annual precipitation. **Ten-years (2001-2010) averaged NDVI within the time span April 6-15.

No variables depicting the top-down regulatory effect of predators on buffalo populations were considered, as bottom-up ecological mechanisms (like quantity and quality of food resources) are argued to play the main role in determining large herbivores spatial occurrence (Winnie *et al.* 2008). At the same time, a potential limit of the presented model may be the lack of variables accounting for the anthropic effect on wild buffalo distribution (Matawa *et al.* 2012).

After checking for collinearity, all the possible models involving the selected environmental variables (i.e. 31 combinations from univariate up to penta-variate models) were tested to predict ψ_S over Uganda. BIC metrics was used for model selection, and the same pipeline developed for *R. appendiculatus* distribution model was applied to standardize variables and to calculate model uncertainty.

Sixty-one *S. caffer* presence-data were derived from the Global Biodiversity Information Facility (GBIF 2012) (Figure 4.3). `Maxlike` was used to estimate ψ_S, by relying on the same logit-linear structure used for *R. appendiculatus*:

$$ln\left(\frac{\psi_{Sx}}{1-\psi_{Sx}}\right) = \beta_0 + \boldsymbol{\beta} z(x)$$

from which

$$\psi_{Sx} = \frac{e^{\beta_0+\boldsymbol{\beta} z(x)}}{1+e^{\beta_0+\boldsymbol{\beta} z(x)}}$$

ψ_{Sx} being *S. caffer* occurrence probability in pixel x.

4.3.5 *Theileria parva parva* infection risk model: γ estimation

In the context of the European Project Nextgen (http://nextgen.epfl.ch), georeferenced blood samples from 587 Ugandan indigenous cattle were tested for the presence/absence of *T. parva parva* p104 antigen DNA sequence as described in Kabi et al. (2014). Samples were collected from 203 farms, distributed over a 51 cells grid covering the whole Uganda with an average of 12 (±4 s.d.) animals per grid-cell and three (±1 s.d.) animals per farm sampled (Figure 4.1).

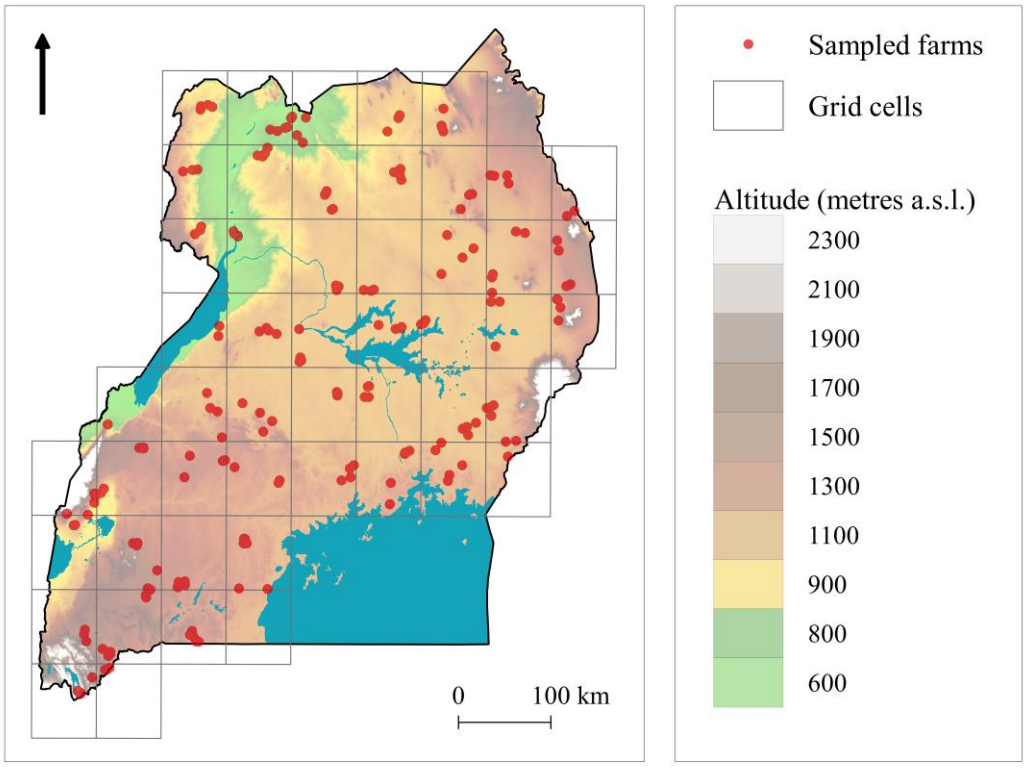

Figure 4.1 Sampling scheme used to collect blood samples from indigenous cattle populations of Uganda.

ECF epidemiology is complex and determined by both biotic and abiotic variables (Norval *et al.* 1992). *R. appendiculatus* presence is considered the primary risk factor of *T. parva parva* infection (Magona *et al.* 2008, 2011; Gachohi *et al.* 2011; Muhanguzi *et al.* 2014). Cattle density represents a further condition for *T. parva parva* occurrence (Olwoch *et al.* 2008), having been demonstrated to influence ECF outbreaks (Billiouw *et al.* 2002). Proximity with *S. caffer* carrier populations is also demonstrated to boost infection probability in cattle (Oura *et al.* 2011; Kabuusu *et al.* 2013). Laboratory investigations report environmental temperatures >28°C to inhibit *T. parva parva* life cycle (Young & Leitch 1981). Furthermore, farming system (extensive or intensive) was reported to be associated with different levels of ECF prevalence (Gachohi *et al.* 2012).

R. appendiculatus and *S. caffer* probabilities of occurrence, cattle density and environmental temperature were considered to predict *T. parva parva* infection risk (γ) over the study area. The maximal temperature in the warmest month of the year (BIO_5) was selected to account for the possible limiting effect of high temperatures on the parasite development (Table 4.3). Furthermore, the influence of farm-specific environmental conditions, management practices (e.g. differential use of acaricides), and unmeasured biological features (e.g. breed- or individual-specific response to tick burden), was considered by treating the sampled farms as random effects in the model (see hereafter).

Table 4.3 Considered biotic and abiotic variables for γ estimation.

Covariate
ψ_R
BIO_5^*

Cattle density[**]
ψ_S

[*]Maximal temperature of the warmest month of the year; [**]number of animals/km^2

Predictors' values were extracted in correspondence of the farm locations, checked for the presence of collinearity and outliers, and subsequently standardized following Bring's procedure prior to parameter estimation.

The probability of infection for each sampled animal ($\hat{\gamma}$) was modelled using a linear mixed-effects logistic regression, where ψ_R, BIO_5, cattle density (Cd) and ψ_S were specified as fixed effects, and random intercepts were computed for each specific farm. Although individual-based, the model can be specified in a farm-based fashion, since all the animals belonging to the *j-th* farm (located in the pixel *j*) present identical values for the considered predictors and, as a consequence, the same predicted $\hat{\gamma}$. Therefore, the model can be written in the form:

$$ln\left(\frac{\hat{\gamma}_j}{1-\hat{\gamma}_j}\right) = \left(\beta_0 + b_{0j}\right) + \beta_{\psi_R}\psi_{Rj} + \beta_{BIO_5}BIO_{5j} + \beta_{Cd}Cd_j + \beta_{\psi_S}\psi_{Sj} \quad (1)$$

$$b_{0j} \sim N\left(0, \sigma^2_{b_0}\right)$$

where $\hat{\gamma}_j$ represents the infection probability for all the animals belonging to the *j-th* farm, $\beta_0 + b_{0j}$ is the *j-th* farm random intercept, $\psi_{Rj}, BIO_{5j}, Cd_j$ and ψ_{Sj} define tick occurrence probability, temperature value, cattle density and buffalo occurrence probability in the pixel *j*, respectively. In this way, environmental conditions characterizing farms with both infected and uninfected individuals are expected to be associated with a lower infection risk if compared to environmental conditions identifying a set of infected individuals only.

Finally, infection risk in any given pixel *x* composing the study area (γ_x) was computed using the population model as estimated from equation (1):

$$\gamma_x = \frac{e^{\beta_0 + \boldsymbol{\beta} z(x)}}{1 + e^{\beta_0 + \boldsymbol{\beta} z(x)}}$$

Where β_0 represents the population intercept, $\boldsymbol{\beta}$ the vector of slope parameters, and $z(x)$ the vector containing the covariates for x.

The model was run using the `glmer` function included in the R package `lme4` (Bates *et al.* 2015), and the Maximum Likelihood criterion was used to obtain point estimates of the parameters.

4.3.6 Landscape genomics

Landscape genomics aims at identifying genetic variants responsible for adaptation to local environmental pressures (Rellstab *et al.* 2015). The approaches developed to identify loci of putative adaptive value rely on associative models evaluating the spatial overlap between landscape features and genetic variation (e.g. Joost *et al.* 2007; Gunther & Coop 2013; Frichot & François 2015). Here, the software SAMβADA v.0.5.3 (Stucki *et al.* 2016) was used to scan the genome of autochthonous cattle from Uganda for candidate genotypes involved in the adaptive response to ECF.

4.3.6.1 Molecular dataset for landscape genomics analysis

The Nextgen project genotyped 813 georeferenced autochthonous cattle from Uganda using the medium-density BovineSNP50 BeadChip (54,596 SNPs, Illumina Inc., San Diego, CA).

This set of individuals was the target of landscape genomic analysis, and will be referred to as 'landscape genomics dataset'. The animals were sampled according to the spatial scheme described in section 4.3.5, and encompassed 503 of the individuals tested for *T. parva parva*'s infection. Quality control (QC) procedures were carried out with the software PLINK v.1.7 (Purcell *et al.* 2007). The landscape genomics dataset was limited to autosomal chromosomes and pruned for minor allele frequency (MAF) <0.01, genotype call rates <0.95, and individual call rate <0.9. Pairwise genome-wide identity-by-descent (IBD) was estimated, and one individual per pair showing IBD>0.5 was excluded from analyses to reduce the risk of spurious associations due to unreported kinship (Turner *et al.* 2011). Spatial positions of the highlighted pairs of individuals were considered prior removal, in order to avoid excluding too much individuals from nearby areas.

4.3.6.2 Population structure variables for landscape genomics analysis

Genome-environment associations may lead to false positive results especially if hidden genetic population sub-structure and habitat characteristics follow a similar spatial pattern (Rellstab *et al.* 2015). Therefore, the inclusion of population structure (e.g. ancestry coefficients from global admixture analyses or principal components from molecular data) is often recommended in the landscape genomics models to correct for spurious associations (Schoville *et al.* 2012).

ADMIXTURE v.1.3.0 (Alexander *et al.* 2009) was used to infer the putative origin of the Ugandan cattle ancestral gene pools. Prior to the analysis, the pruned landscape genomics dataset was merged with molecular data of other sanga (AI), African taurine (AT), Asian

zebuine (ASI), and European taurine (ET) populations retrieved from public databases and provided by co-authors (Supplementary 4.7.3). PLINK was used to prune the merged dataset (hereinafter 'population structure dataset') for linkage disequilibrium (LD) >0.1, using sliding windows of 50 SNPs and 10 SNPs steps (--indep-pairwise 50 10 0.1), as well as to filter for the QC thresholds reported in section 4.3.6.1.

Four ancestral gene pools were previously identified to best explain the genetic structure of the same Ugandan indigenous populations (Stucki *et al.* 2016). For this reason, this clustering solution (four clusters) was considered to represent the underlying neutral structure of the landscape genomics dataset. Due to a high degree of collinearity among two of the obtained ADMIXTURE components (|r|>0.7, Dormann *et al.* 2013), a PCA was performed trough the R function `prcomp` to provide orthogonal population structure variables for SAMβADA. After PCA analysis, the first three principal components were retained for landscape genomics analysis (see section 4.4.4).

4.3.6.3 Landscape genomics models

Given diploid species and biallelic markers, SAMβADA runs three models per locus, one for each possible genotype (i.e. AA, AB and BB). Each model estimates the probability π_i for the *i-th* individual to carry a given genotype on the logit scale, as a function of the considered environmental and population structure variables:

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{v=1}^{n} \beta_v z_{vi}$$

which is:

$$\pi_i = \frac{e^{\beta_0 + \sum_{v=1}^{n} \beta_v z_{vi}}}{1 + e^{\beta_0 + \sum_{v=1}^{n} \beta_v z_{vi}}}$$

Following Stucki *et al.* (2016), the spatial occurrence of each genotype was predicted by two competing models: a "null model" exclusively comprising the population structure variables, and an "alternative model" including the environmental variable of interest along with the population structure variables. A genotype was considered to be significantly associated with the environmental variable if the *p-value* associated with the likelihood ratio test statistic (*D*) among the "null" and "alternative" models resulted lower than the nominal significance threshold of 0.05 after a Benjamini-Hochberg (BH) correction for multiple testing (H_0: D=0, α_{BH}=0.05). The R function `p.adjust` was used to perform *p-values* corrections.

In the present study, ψ_R and γ values at the geographical location of each genotyped animal were separately tested as environmental variables into landscape genomics models. SAMβADA was allowed to run all the combinations up to tetra-variate models. For each genotype, the models comprising the three population structure variables only (i.e. the "null" models) and those comprising the three population structure variables plus either ψ_R or γ (i.e. the "alternative" models) were considered. "Null" and "alternative" models were then compared through likelihood ratio test (Supplementary 4.7.4). Logistic regressions in SAMβADA were performed on centred predictors to facilitate estimation of parameters.

4.3.7 Gene identification and local admixture analysis

Selected loci were used to identify annotated genes in the Ensembl database (Aken *et al.* 2016). Global linkage disequilibrium (LD) decay was estimated using SNeP v.1.11 (Barbato *et*

al. 2015) to define the extent of LD around loci. Genes located within 25 kbp ($r^2 \approx 0.2$) upstream and downstream a selected marker were then investigated for known biological function.

A local ancestry investigation was applied to infer the ancestral origin of genomic regions of interest (e.g. harbouring genes involved into disease tolerance). Local ancestry investigations allow to assign haplotype genomic ancestry given a set of reference populations, and have been used to infer the admixture history of closely related groups, as well as to highlight target regions of recent selection in several species (Tang *et al.* 2007; Pașaniuc *et al.* 2009). PCAdmix v.1.0 (Brisbin *et al.* 2012) was used to perform local ancestry inferences of targeted genomic regions. References were selected to represent the major gene pools observed in Uganda after population structure analysis (section 4.4.4), i.e. sanga (represented by a group of Ankole-Watusi individuals from Rwanda), zebu (Tharparkar from Pakistan), African taurine (Muturu from Nigeria) and European taurine (Hereford from British Isles).

Local ancestry analyses were not performed on the landscape genomics dataset, since the density of the markers would have not allowed to assign haplotypes with sufficient precision (Brisbin *et al.* 2012). Rather, HD genotype data (Illumina BovineHD Genotyping BeadChip) of 102 individuals collected in the same NEXTGEN sampling campaign were used and assumed to reflect the ancestry of the individuals composing landscape genomics dataset.

PCAdmix infers local ancestry for non-overlapping sliding windows, windows size being determined by a user-defined number of markers. In the present analyses, 20 SNPs per window were used, as this value allowed a window size comparable to the optimal one suggested by Brisbin *et al.* for a similar data density (Brisbin *et al.* 2012).

4.4 Results

4.4.1 *Rhipicephalus appendiculatus* distribution model

The first three principal components (PC_1, PC_2 and PC_3) of the PCA performed on the selected bioclimatic covariates explained 95% of the total amount of variance. PC_1 primarily summarized information from BIO_8, BIO_9, BIO_{10} and BIO_{11} variables. PC_2 from BIO_{16}, BIO_{17}, BIO_{18} and BIO_{19} variables, and PC3 from bio19 and bio8 variables. The model employing PC_1, PC_2 and PC_3 showed the lowest BIC value, and was thus retained for subsequent analyses (Supplementary 4.7.5).

With an estimated standardized coefficient equal to −1.799 and an odds ratio (OR) equals to 1.165, PC_3 showed the most important negative conditional effect on *R. appendiculatus* occurrences. Contrarily, PC_1 and PC_2 showed a positive conditional effect, with odds ratio equal to 2.217 and 2.275, respectively. All the considered covariates resulted significantly associated with the species spatial occurrence (Table 4.4).

Table 4.4 `Maxlike` results for *R. appendiculatus* distribution model.

Coefficient[1]	Estimate[2]	SE[3]	*p-value* (>|*z*|)[4]	OR[5]	OR_{low}[6]	OR_{up}[7]
β_0	−2.905	0.561	2.24E−07	0.055	0.018	0.164
PC_1	0.796	0.303	8.56E−03	2.217	1.224	4.014
PC_2	0.822	0.37	2.62E−02	2.275	1.102	4.698
PC_3	−1.799	0.629	4.27E−03	0.165	0.048	0.568

[1]Standardized regression coefficients, [2]Point estimates of the coefficients on the log odds scale, [3]Standard errors of the coefficients on the log odds scale, [4]*p-value* associated to the coefficients (H_0: β_i=0, α=0.05), [5]Odds ratios associated to the coefficients. Odds ratio expresses the expected change in the ratio $\psi_R/(1-\psi_R)$, for a one standard deviation increase of the concerned predictor (by holding all the other covariates fixed at a constant value). [6]Odds ratio 95% confidence interval (CI), lower bounds.

[7]Odds ratio 95% CI, upper bounds.

The selected model predicted an average ψ_R of 0.148 over the entire study area (*Md*=0.062). In particular, regions north of Lakes Kwania, Kyoga and Kojwere generally showed low habitat suitability ($0< \psi_R <0.1$). Habitat suitability increased towards Lake Victoria coasts, where ψ_R reached the highest predicted values ($0.4< \psi_R <1$). A smaller, highly suitable area was also predicted South-West of Lake Albert, at the foot of Rwenzori Mountains ($0.4< \psi_R <0.8$). A corridor of lower suitability ($0< \psi_R <0.3$) appeared to separate Lake Victoria and Rwenzori Mountains (Figure 4.2).

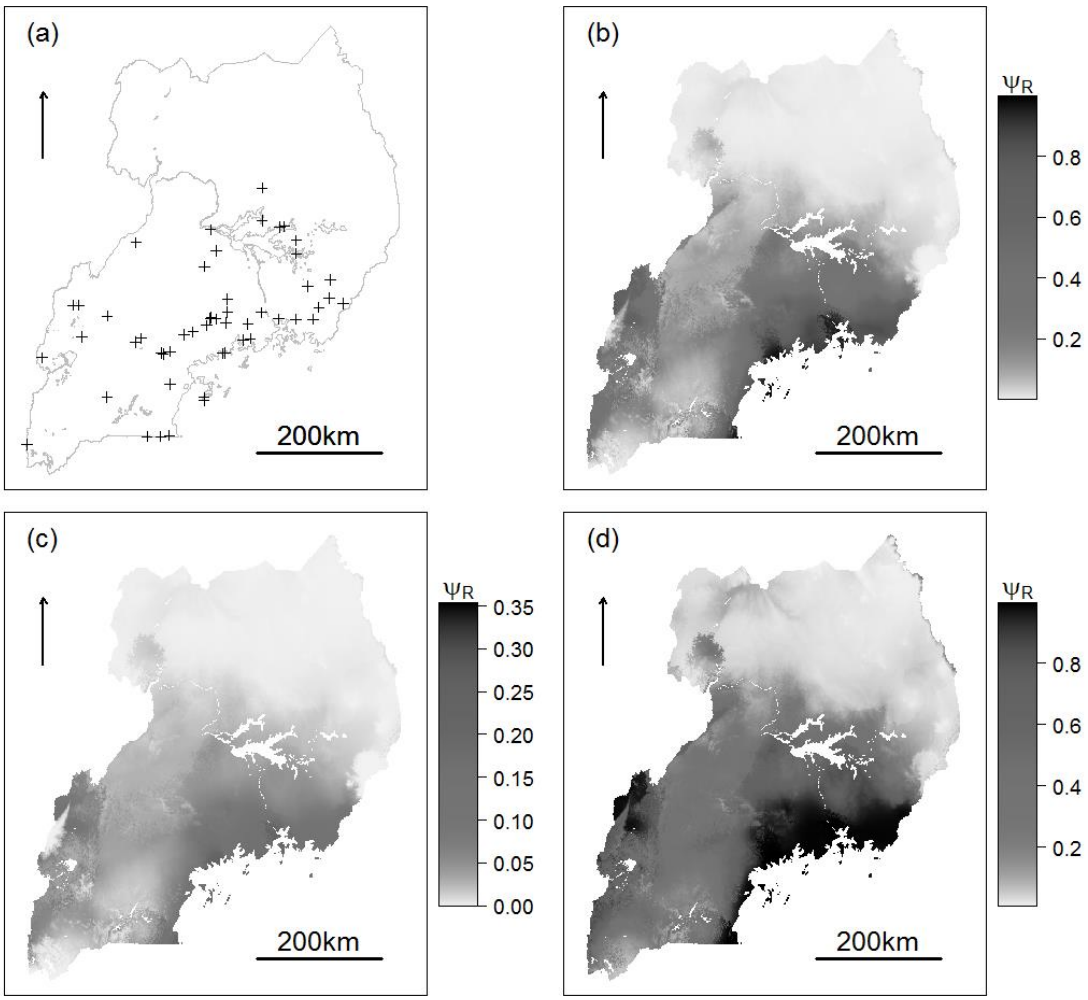

Figure 4.2 (a) *R. appendiculatus* spatial occurrences as retrieved from Cumming. 1999b. **(b)** Map of *R. appendiculatus* occurrence probability (ψ_R) as derived from the selected distribution model. Colour key corresponds to the estimated tick occurrence probability: the darker the colour, the higher the probability. **(c)** and **(d)** Lower and upper bounds of the 95% confidence intervals of ψ_R, respectively.

4.4.2 *Syncerus caffer* distribution model

The chosen set of environmental variables showed a low degree of collinearity (|r|<0.7 in all the pairwise comparisons among the predictors). One model over the 31 tested did not reach convergence and was discarded from the model-selection procedure. The model including a

linear combination of altitude, annual precipitation, average NDVI and distance from the nearest water source showed the lowest BIC value and was retained for subsequent analyses (Supplementary 4.7.6).

All the environmental covariates resulted significant (H_0: β_i=0, α=0.05), NDVI showed the greatest positive effect (OR=17.499). Conversely, distance from water (OR=0.136), altitude (OR=0.335) and precipitation (OR=0.449) showed negative relationships with buffalo occurrence (Table 4.5).

Table 4.5 `Maxlike` results for *S. caffer* distribution model.

Coefficient	Estimate	SE	*p-value* (>|*z*|)	OR[1]	OR_{low}	OR_{up}
β_0	−9.130	0.790	6.46E−31	0.000	0.000	0.001
Altitude	−1.095	0.293	1.90E−04	0.335	0.188	0.594
BIO_{12}	−0.800	0.180	9.03E−06	0.449	0.316	0.639
NDVI	2.862	0.329	3.38E−18	17.499	9.181	33.343
Wd	−1.996	0.434	4.23E−06	0.136	0.058	0.318

[1]Expected change in the ratio $\boldsymbol{\psi_S/(1-\psi_S)}$ for a one standard deviation increase of the concerned predictor.

The model predicted an average ψ_S of 0.005 over the study area (*Md*=3.49E−04). Higher occurrence probabilities ($0.2<\psi_S<0.8$) were recorded in the near proximity of the water bodies (especially along the White Nile in the North-West, the South-eastern coasts of Lake Édouard, and the coasts north of Lake George in the South-West), as well as in small patches near Katonga Game Reserve (Figure 4.3).

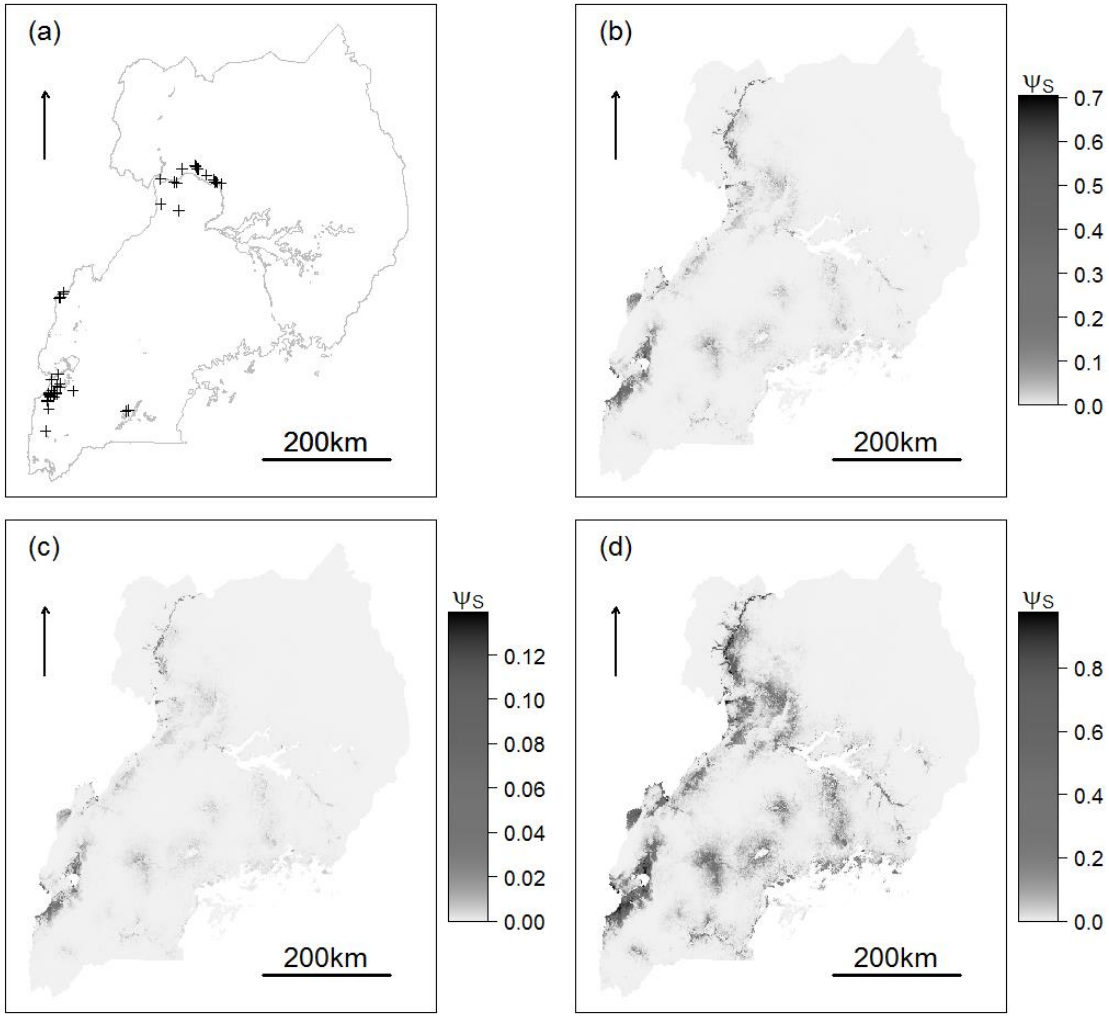

Figure 4.3 (a) *S. caffer* spatial occurrences as retrieved from GBIF, 2012. **(b)** Map of *S. caffer* occurrence probability (ψ_S) as derived from the selected distribution model. Colour key corresponds to the estimated tick occurrence probability: the darker the colour, the higher the probability. **(c)** and **(d)** Lower and upper bounds of the 95% confidence intervals of ψ_S, respectively.

4.4.3 *Theileria parva parva* infection risk model

Predictors of the model were checked for the presence of potentially influential outliers by boxplot visualization (Supplementary 4.7.7). Following inspection, ψ_R, cattle density and $\psi_S$were transformed on a $\log_{10}$ scale to reduce the observed skewness in the distributions. No worrying collinearity was observed among the predictors of the model ($|r|<0.7$).

All the explanatory variables except for cattle density showed a significant effect (H_0: $\beta_i=0$, $\alpha=0.05$). Particularly, BIO_5 (OR=0.649) resulted to have the most important conditional effect, followed by ψ_S (OR=1.279) and ψ_R (OR=0.803). With an estimated standardized coefficient of −0.219, ψ_R showed a negative association with *T. parva parva* infection (Table 4.6).

Table 4.6 Results for *T. parva parva* infection risk model

Coefficient	Estimate	SE	*p-value* (>|z|)	OR	OR_{low}	OR_{up}
β_0^*	−1.128	0.115	1.21E−22	0.324	0.258	0.406
$\log_{10}(\psi_R)^{**}$	−0.219	0.105	3.72E−02	0.803	0.654	0.987
BIO_5	−0.432	0.104	3.18E−05	0.649	0.529	0.796
$\log_{10}(Cd^{***})$	0.015	0.105	8.86E−01	1.015	0.826	1.247
$\log_{10}(\psi_S)$	0.246	0.111	2.67E−02	1.279	1.029	1.590

*Estimated population intercept. **Estimated population slope for *R. appendiculatus* effect. ***Cattle density.

The model predicted an average γ of 0.253 across Uganda (*Md*=0.235). Overall, Northern regions presented a range of probability of infection between 0.1-0.3 A similar range was observed southwards, in the region comprised between Lake Kyoga, Lake Victoria, Lake Albert and the Eastern borders with Kenya. Moving towards South-West, infection probability

increases following a positive gradient from *c.* 0.30 to *c.* 0.70 in the most southern districts (Figure 4.4).

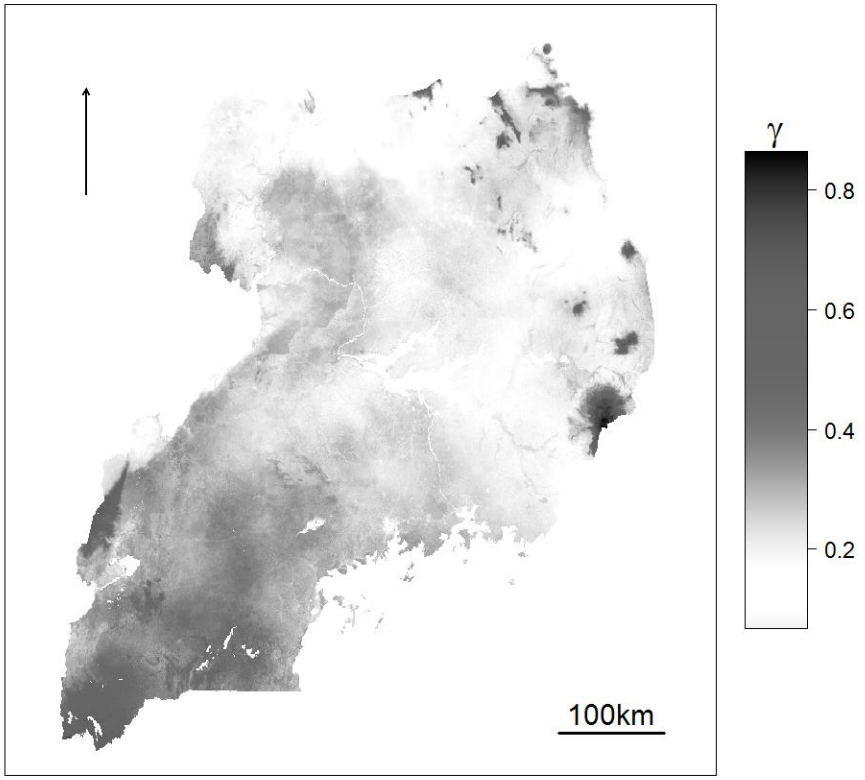

Figure 4.4 Map of the estimated *T. parva parva* risk of infection (γ) in cattle.

4.4.4 Population structure analysis

After pruning for MAF, genotype call rate and individual call rate, population structure dataset counted 12,925 SNPs and 1,355 individuals, among which 743 from Uganda, 131 and 158 composing ET and AT groups, and 195 and 128 composing AI and ASI. Sanga type represented the main gene pool shared by Ugandan individuals, showing an average of 76% (±13%) of cluster assignment (Supplementary 4.7.8). However, >20% of zebuine component

was detected in more than half of the analysed samples, with an average of 18% (±13%) over the entire Ugandan group. Cluster assignments referable to the African and European taurine components were also present, both constituting around 3% of the individual ancestries. In accordance with Stucki et al. (2016), genomic components showed a defined spatial structure, the zebu gene pool being more present in the North-East of the country, and the sanga in the central and South-West. African taurine was detectable as a background component especially in the North-West and South-West, while European introgression could be mostly identified in the South-West.

PCA explained 100% of the original variance in the four Admixture Q-scores with the first three principal components. PC_1 discriminated between sanga and zebu gene pools, PC_2 pointed out European introgression, and PC_3 showed the highest correlation with the African taurine gene pool. PC_1. PC_2 and PC_3 were included into landscape genomics models to represent genetic structure of Ugandan individuals.

4.4.5 Landscape genomics

After QC, landscape genomics dataset counted 40,886 markers and 743 individuals. Retained animals were located in 199 farms (4±1 samples/farm) and 51 cells grid (15±5 samples/cell).

Sixty-three genotypes across 41 putative adaptive loci were found to be significantly associated with *R. appendiculatus* potential distribution. Associated loci were distributed over 18 chromosomes (Figure 4.5a and Supplementary 4.7.9a). Moreover, eight genotypes across seven loci resulted significantly associated with the estimated *T. parva parva* infection risk. In

particular, four SNPs were found in a region of 103.5 kbp on chromosome 13 between 66,292 and 66,395 Mbp (Figure 4.5b and Supplementary 4.7.9b)

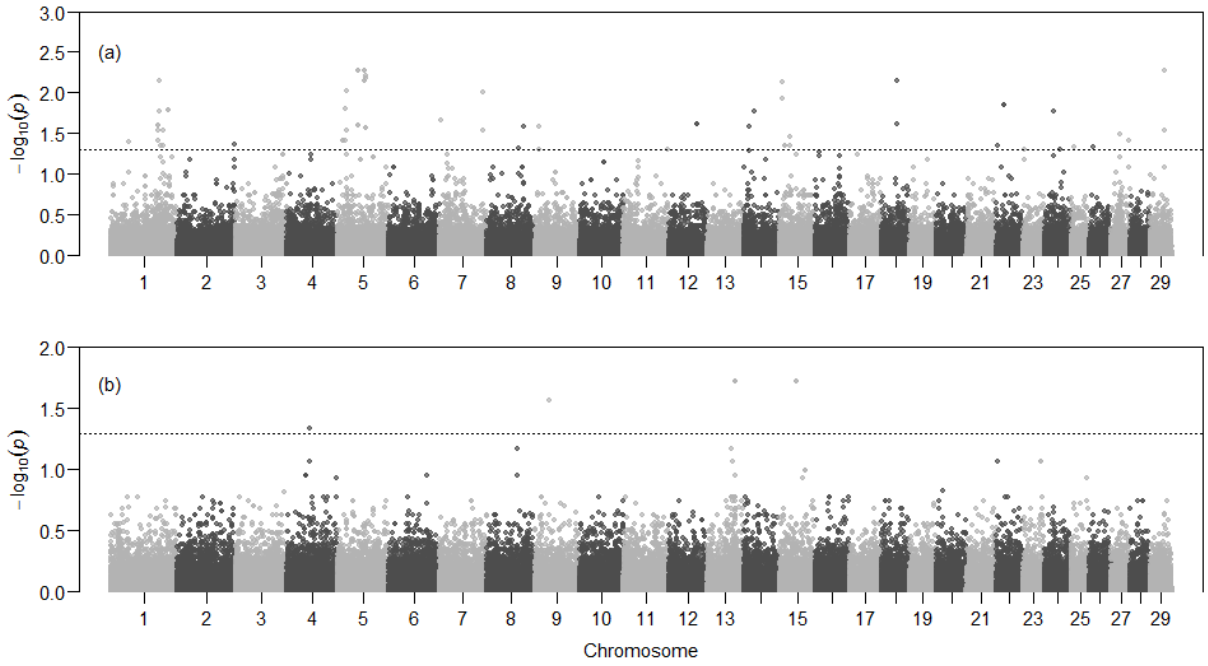

Figure 4.5 (a) Manhattan plot for the genotype-environment association study involving *R. appendiculatus* occurrence probability (Supplementary 4.7.10a). Each point represents the test statistic *p-value* referred to a single genotype. Displayed values are on $-\log_{10}$ scale after multiple testing correction. *X*-axis depicts chromosomal position of the tested markers. Nominal significance threshold (α_{BH}=0.05) is also displayed on the –log10 scale as a dotted line. **(b)** Manhattan plot for the environmental association study involving *T. parva parva* infection risk (Supplementary 4.7.10b).

4.4.6 Gene identification and local admixture analysis

Of the 41 loci significantly associated with *R. appendiculatus* distribution, 18 presented at least one annotated gene in the cattle genome within the selected window size (Table 4.7a). Locus BTA-113604-no-rs (hereafter BTA-113604) resulted to be positioned around 12.5 kbp apart from Protein kinase, cGMP-dependent, type I (PRKG1) gene on chromosome 26. This gene was already described to be involved in tick resistance mechanisms in South African

Nguni cattle (Mapholi *et al.* 2016).

Six out of the seven loci associated with *T. parva parva* infection presented at least one annotated gene within the selected window (Table 4.7b). Two SNPs (ARS-BFGL-NGS-110102 and ARS-BFGL-NGS-24867, hereafter ARS-110102 and ARS-24867, respectively) positions were within Src-like-adaptor 2 (SLA2) gene on chromosome 13. SLA2 human orthologue is known to encode the Src-like-adaptor 2, a member of the SLAP protein family involved into regulation of T and B cell-mediated immune response (Holland *et al.* 2001).

Genomic regions encompassing BTA-113604, and ARS-110102/ARS-24867 (between positions 8.331-8.614 Mbp on chromosome 26, and positions 65.837-66.649 Mbp on chromosome 13, respectively) were further investigated with local ancestry inference given their possible biological role in adaptation to ECF. Of the 204 haploid individuals investigated, 159 showed a sanga ancestry for the BTA-113604 region, 37 were assigned to the Tharparkar reference (zebuine ancestry), seven to Hereford (European ancestry) and one to Muturu (African taurine ancestry). The genomic region holding ARS-110102 and ARS-24867 had 164 haplotypes assigned to the sanga reference, 23 to the zebuine reference and two to the European *B. taurus*. No African taurine ancestry was recorded for this genomic region, and 7.3% of the individuals were assigned with a low posterior probability (<0.95)

Among the 42 haplotypes sampled in the areas with the highest predicted tick burden (grid cells around Lake Victoria), 29 presented sanga ancestry and 13 zebuine ancestry (Figure 4.6a). Further, among the 44 haplotypes sampled in areas with high *T. parva parva* infection risk (grid cells in the South-West of Uganda), 41 resulted to have sanga ancestry and three indicine ancestry (Figure 4.6b).

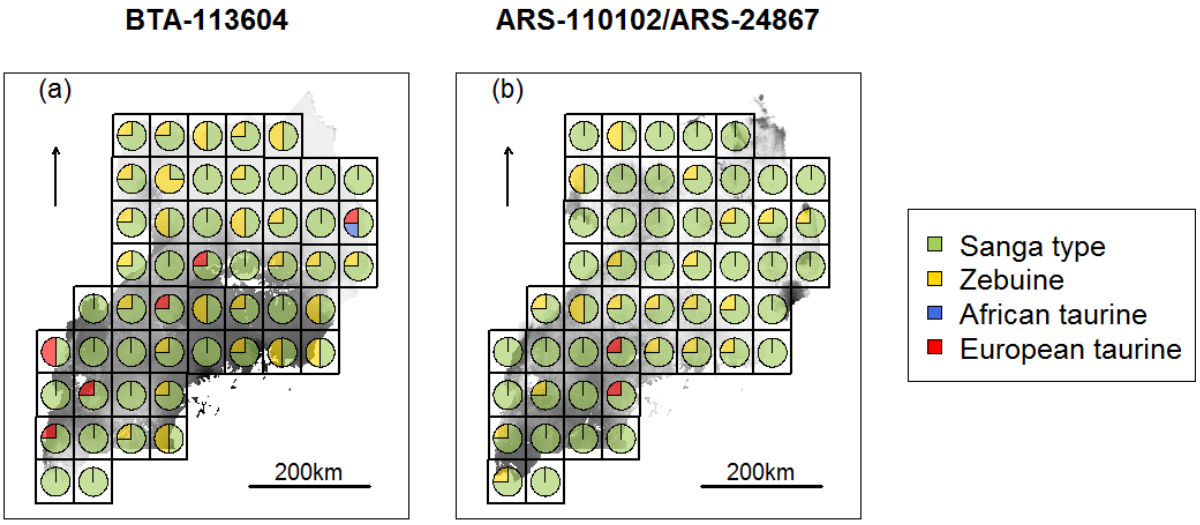

Figure 4.6 Ancestries of haploid individuals summarized per cell grid. Each pie chart refers to a specific cell and shows the proportion of haploid individuals having sanga, zebuine, African and European taurine ancestries. **(a)** Ancestries for the genomic region encompassing marker BTA-113604 on chromosome 26. Estimated *R. appendiculatus* occurrence probability is plotted in the background. **(b)** Ancestries for the genomic region encompassing markers ARS-110101 and ARS-24867 on chromosome 13. *T. parva parva* cattle infection risk is plotted in the background.

Table 4.7 Gene identification for the loci significantly associated with *R. appendicualtus* occurrence probability **(a)** and *T. parva parva* cattle infection risk **(b)** as resulted from SAMβADA analysis.

(a)

SNP ID[1]	Genotype(s)[2]	Chr.[3]	Position[4]	Annotated gene[5]	Biological function[6]
ARS-BFGL-NGS-110339	AA, AC	1	111,495,891	Uncharacterized	-
Hapmap34409-BES7_Contig244_858	AA	1	120,149,924	Glycogenin-1 (GYG1)	Energy metabolism and angiogenesis (Lancaster *et al.* 2014)
Hapmap34056-BES2_Contig421_810	AG, GG	1	138,178,130	DnaJ heat shock protein family (Hsp40) member C13 (DNAJC13)	Heat shock proteins (Kodiha *et al.* 2012)
ARS-BFGL-NGS-32909	CC, AC	5	67,846,632	5'-nucleotidase domain containing 3 (NT5DC3)	UP-regulated genes for iron content in Nelore cattle (Wellison Jarles da Silva 2015)
				Uncharacterized	-
ARS-BFGL-NGS-37845	AG, AA	5	48,633,731	Methionine sulfoxide reductase B3 (MSRB3)	Affect ear floppiness and morphology in dogs (Boyko *et al.* 2010)
BTA-46975-no-rs	CG, GG	5	68,220,538	Thioredoxin reductase 1. cytoplasmic (TXNRD1)	Milk production and oocyte developmental competence in cattle (Gilbert *et al.* 2012; Ghorbani *et al.* 2015)
Hapmap51626-BTA-73514	AA, AG	5	48,834,486	Inner nuclear membrane protein Man1 (LEMD3)	Height in pigs and cattle (Frantz *et al.* 2015)
UA-IFASA-6140	AG, AA	7	102,472,846	ST8 alpha-N-acetyl-neuraminide alpha-2.8-sialyltransferase 4 (ST8SIA4)	Metabolism of milk glycoconjugates in mammals (Song *et al.* 2016)

BTB-00292673	AA	7	4,953,801	Phosphodiesterase 4C (PDE4C)	Fertility (Glick *et al.* 2011)
				Member RAS oncogene family (RAB3A)	Calcium exocytosis in neurons (Brondyk *et al.* 1995)
				MPV17 mitochondrial inner membrane protein like 2 (MPV17L2)	Immune system (Brütting *et al*. 2016)
Hapmap31116-BTA-143121	AA	8	7,597,3285	Epoxide hydrolase 2 (EPHX2)	In vitro maturation. fertilization and culture on bovine embryos (Smith *et al.* 2009)
				L-gulonolactone oxidase (GULO)	Involved into vitamin C production in pigs (Hasan *et al.* 2004)
ARS-BFGL-NGS-104610	AG	11	104,293,559	Surfeit 6 (SURF6)	Housekeeping gene (Magoulas *et al.* 1998)
				Mediator complex subunit 22 (MED22)	Gestation length in Nelore cattle (Matos *et al*.. 2013)
				Ribosomal protein L7a (RPL7A)	Oocyte developmental competence in cattle (Gilbert *et al.* 2012)
				Uncharacterized	-
				Small nucleolar RNA (SNORD24)	May act as methylation guide for RNA targets (Kiss-László *et al.* 1996)
				Small nucleolar RNA (SNORD36)	2'-O-ribose methylation guide (Galardi *et al.* 2002)
				Small nucleolar RNA (snR47)	2'-O-methylation of large and small subunit rRNA (Samarsky & Fournier 1999)

				Small nucleolar RNA (SNORD24)	As above
				Small nucleolar RNA (SNORD36)	As above
BTB-00839408	AG. AA	22	18,978,658	Metabotropic glutamate receptor 7 precursor (GRM7)	Might be related to parasite resistance (Xu *et al.* 2016)
ARS-BFGL-NGS-39898	GG	22	1,319,636	Novel gene	-
ARS-BFGL-BAC-31319	AA	23	4,847,028	3-hydroxymethyl-3-methylglutaryl-CoA lyase like 1 (HMGCLL1)	Involved into ketogenesis (Tetens *et al.* 2015)
Hapmap51155-BTA-11643	AA	24	38,086,180	DLG associated protein 1 (DLGAP1)	Role in neurological development and behavioral disorders (Sorbolini *et al.* 2015)
Hapmap57868-rs29020458	AA	24	22,746,291	Dystrobrevin alpha (DTNA)	Formation and stability of synapses (Sjö *et al.* 2005)
				U6 spliceosomal RNA (U6)	Participate into spliceosome formation (Marz *et al.* 2008)
BTA-113604-no-rs	AA	26	8,356,096	Protein kinase. cGMP-dependent. type I (PRKG1)	Tick resistance in South African Nguni cattle (Mapholi *et al.* 2016)
ARS-BFGL-NGS-18933	GG	29	34,650,967	Opioid binding protein/cell adhesion molecule like (OPCML)	Role in opioid receptor function in humans (Smith *et al.* 1993)

(b)

SNP ID[1]	Genotype[2]	Chr.[3]	Position[4]	Annotated gene[5]	Biological function[6]
BTB-01298953	AA	4	54,930,726	Protein phosphatase 1 regulatory subunit 3A (PPP1R3A)	Glycogen synthesis in humans and mice (Savage *et al.* 2008)
BTA-33234-no-rs	GG	13	66,291,997	DLG associated protein 4 (DLGAP4)	Neuronal membrane protein (Takeuchi *et al.* 1997)
				Myosin light chain 9 (MYL9)	May participate in regulation of muscle contraction (Kumar *et al.* 1989)
ARS-BFGL-NGS-112656	AA	13	66,336,246	Myosin light chain 9 (MYL9)	As above
				TGFB induced factor homeobox 2 (TGIF2)	Transcriptional repressor (Imoto *et al.* 2000)
ARS-BFGL-NGS-110102	GG	13	66,370,867	TGFB induced factor homeobox 2 (TGIF2)	As above
				TGIF2-C20orf24 readthrough (C13H20orf24 alias RIP5)	May promote apoptosis in humans (Zha *et al.* 2004)
				Src-like-adaptor 2 (SLA2)	Downregulation of T and B cell-mediated responses (Holland *et al.* 2001)
ARS-BFGL-NGS-24867	AA	13	66,395,465	Src-like-adaptor 2 (SLA2)	As above
				NDRG family member 3 (NDRG3)	Linked to prostate cancer cells growth (Lee *et al.* 2016)

Hapmap39482-BTA-36746	CC, AC	15	40,279,014	TEA domain transcription factor 1 (TEAD1)	Transcription factor promoting apoptosis in mammals (Landin Malt *et al.* 2012)

[1]Name of the marker with associated genotype(s). [2]Associated genotype(s) from SAMβADA analysis. For estimated regression coefficients, refer to S11. [3]Name of the chromosome where the associated SNP is located. [4]Position on the chromosome in base pairs. [5]Genes falling within the selected window of 50 kbp centered on the marker position, as derived from the Ensembl database. [6]Known biological function of the annotated genes (description is provided for the found reference species).

4.5 Discussion

ECF represents a major issue for livestock health in several sub-Saharan countries (Nene *et al.* 2016), with over one million cattle per year struck by the disease, and an estimated annual economic damage comprised between 168 and 300 million USD (Norval *et al.* 1992; McLeod & Kristjanson 1999).

ECF distribution is highly correlated with the presence of its vector, the tick *R. appendiculatus*, whose occurrence is an essential precondition for *T. parva parva* infection in cattle (Olwoch *et al.* 2008). However, the present study showed that areas with a predicted poor habitat suitability for the tick present higher infection rates when compared to regions highly suitable for the ECF vector (Table 4.4), indicating that, while necessary, the presence of the vector may not be sufficient to justify *T. parva parva* infection. Here, we speculate three factors which may contribute in shaping such a counterintuitive pattern:

1) Environmental temperature (BIO_5) may play a pivotal role in shaping spatial pattern of *T. parva parva* infection in Uganda. High temperatures have been demonstrated to be more detrimental than low ones for the parasite survival at the piroplasms stage into the tick salivary glands (Young & Leitch 1981). Even short periods (around 15 days) of temperatures >28°C were reported to limit development more than equal-length periods of low temperatures (4°C) (see Table 3 in Young & Leitch 1981). Therefore, environmental temperature may affect ECF epidemiology in those areas exceeding the upper bound of the thermic optimal range for *T. parva parva* development (around 28°C, Young & Leitch 1981), by inhibiting *R. appendiculatus* transmission of the parasite. In the case of Uganda,

highly suitable areas for *R. appendiculatus* North-East of Lake Victoria can reach 30°C in the warmest month of the year (January), and exhibit a low infection risk. Conversely, moving towards South-West, temperature ranges between *c.* 8-28°C during the whole year (data not shown). In these regions, the predicted risk of infection increases, despite a concomitant decrease in habitat suitability for the tick. According to these findings, highly suitable regions for *R. appendiculatus* show temperatures above the optimal range for the parasite development in some periods of the year, a condition which could act as a limiting factor for *T. parva parva* survival, and thus affect ECF transmission dynamics.

2) The most suitable areas for ECF vector overlap a structured spatial presence of zebuine ancestry (Supplementary 4.7.8). Zebuine cattle display higher tick resistance than European *Bos taurus* (Brizuela *et al.* 1996), consequently showing a reduced tick-borne micro-organisms infection rate (Mattioli *et al.* 2000). Therefore, the concomitant occurrence of tick-resistant populations and a sub-optimal niche for the parasite might explain the low infection risk observed in *R. appendiculatus* most suitable areas. Further, indigenous cattle inhabiting areas less infested by ticks (e.g. the Southern districts) but more suited to *T. parva parva* life cycle could have not evolved tick-specific adaptations, and therefore manifest higher infection rates.
3) *R. appendiculatus* distribution model does not explicitly consider the effect of anthropogenic factors like tick control campaigns on a local and temporal basis. However, it is worth remarking that control campaigns are rarely applied properly and with efficacy in Uganda, as underlined by the Ugandan National Drug Authority, and *R. appendiculatus* might be developing drug resistance (Vudriko *et al.* 2016).

Vast areas in the North of Uganda display γ>0 despite estimated ψ_R≈0. Indeed, the negative relationship inferred between γ and ψ_R may concur in partially explaining such a result. However, infection is actually present in the North, and a cause for these positive observations may be represented by a lack of *R. appendiculatus* records in the available dataset.

Genetic adaptive response to ECF is a complex process, possibly involving adaptation to both the tick vector, and the parasite. Given the emerging ECF eco-epidemiological picture, local adaptation towards tick burden could have evolved along Lake Victoria coasts, where higher infestation rate were recorded (Fig. 4.2a). Conversely, in South-West Uganda specific adaptive responses to *T. parva parva* may have evolved due to the simultaneous presence of favourable ecological conditions for the parasite development (despite a lower tick burden), and of a less tick-resistant cattle population bearing a lower proportion of zebuine ancestry (Supplementary 4.7.8).

Tick resistance in cattle is a trait under genetic control (Marufu *et al.* 2011), with zebuine-like cattle being generally more efficient in counteracting tick infestation than *B. taurus* (Jonsson *et al.* 2014). Cutaneous inflammatory reactions triggered by the tick bite were identified to constitute the core adaptation to tick burden in cattle (Mattioli *et al.* 2000), with tick-resistant breeds showing a strong white blood cells mediated cutaneous reaction (Willadsen 1980) affecting tick attachment, salivation and engorgement and limiting inoculation of tick-borne microorganisms (Wikel & Bergrnan 1997). Therefore, adaptive mechanisms against tick infestation may play a pivotal role in limiting the effects of *T. parva parva* infection, whose clinical course is known to be parasite dose-dependent (Brossard & Wikel 1997; Nene *et al.* 2016).

Here, genomics regions across 18 different chromosomes were found to be significantly associated with ψ_R. This finding is in agreement with former research suggesting the polygenic nature of tick resistance in cattle (Mapholi *et al.* 2016). In particular, the highest number of putative loci under selection was found on BTA5 (9 loci), BTA1 (7 loci), and BTA15 (3 loci). However, none of these markers fell within or nearby an annotated gene easily attributable to tick resistance (Table 4.7). Conversely PRKG1 was identified in high LD with a marker on BTA26 significantly associated with tick occurrence probability. PRKG1 is an important mediator of vasodilation, a classical feature of inflammatory response (Sherwood & Toliver-Kinsky 2004; Surks 2007), and notably, was also reported as a candidate gene for tick resistance displaying a significant correlation with *Boophilus* infestations (Mapholi *et al.* 2016).

Genotype-environment analysis evidenced SLA2 on BTA13 as significantly associated with *T. parva parva* infection risk (both ARS-110102 and ARS-24867 markers fall within SLA2 genic region). SLA2 is involved with signal transduction in B and T cells, downregulates humoral and cell-mediated immune responses, and contributes to a correct activation and proliferation of lymphocytes (Holland *et al.* 2001; Marton *et al.* 2015; Kazi *et al.* 2015). *T. parva parva* invades cattle lymphocytes, and promotes a complex series of intra-cellular events which ultimately lead to a pathogenic clonal expansion of the parasitized cells (Baldwin *et al.* 1988; McKeever & Morrison. 1990). Such an antagonistic effect on lymphocytes proliferation would suggest the involvement of SLA2 with *T. parva parva*'s life cycle. However, further molecular and immunological investigations are needed to confirm such hypothesis.

Preliminary local ancestry analyses highlighted a preponderant indicine or Sanga origin for the candidate genomic regions under selection in the geographical areas with high tick

burden or ECF infection risk, while European taurine introgression was observed in areas at lower selection pressure. Particularly, taurine introgression from Europe appears patchy in the case of tick burden (Figure 4.6a), whilst concentrated into two nearby grid cells West of Lake Victoria in the case of ECF infection risk (Figure 4.6b). These findings suggest a possible adaptive advantage for the animals carrying gene variants evolved either in India or Africa, and point out the relevance of monitoring allochthonous introgression and conserving local genetic resources.

By excluding African *B. taurus*, local ancestry analyses point towards a possible zebuine or sanga origin for the highlighted genomic regions. However, the sample size per cell was somehow limited (on average 2±0.2 animals per grid cell), and ancestry assignations are reference-dependent (Barbato 2016). Indeed, alternative zebuine and sanga breeds might be tested to verify the reliability of the obtained assignations. Further, the concomitant existence of two ancestral components, sanga and zebuine, conferring adaptation to ECF might either suggest the evolution of local adaption in zebuine animals and the subsequent introgression into sanga, or convergent evolution between zebuine and sanga animals for the mentioned traits.

Objective limitations must be recognized to potentially affect the proposed distribution and infection models and the consequent genotype-environment association analysis. Firstly, the reduced sample sizes of *R. appendiculatus* and *S. caffer* datasets (51 and 61 occurrences, respectively) might have undermined the reliability of the predicted values for ψ_R and ψ_S. As demonstrated by Merow & Silander (2014), comparable sample size are expected to affect the estimation of the model intercept and decrease precision in ψ estimation. Further, ψ estimation might have been impacted by: (i) potentially biased

species occurrence datasets, which may not comply with `Maxlike` random sampling assumption (Merow & Silander 2014); (ii) the reliability of occurrence records, which derive from heterogeneous collections (Olwoch *et al.* 2003); (iii) a variable accuracy in point locations coordinates (see Cumming. 1999b for a detailed description of tick data reliability). However, `Maxlike` was the preferred modelling solution due to its capacity to directly estimate ψ, which is a quantity of immediate ecological meaning and interpretability. Moreover, standard errors associated to intercept estimates are not large, (around 0.6 and 0.8 on the logit scale for *R. appendiculatus* and *S. caffer* models, respectively), suggesting a precise parameter estimate.

Secondly, the reliability of epidemiological information (false positives/negatives rates in laboratory assays) was not taken into account by the proposed infection risk model (section 4.3.5). At the same time, the performed genotype-environment association study relies on the assumption that areas with a high risk of infection (i.e. endemically stable areas) are inhabited by locally ECF-adapted indigenous cattle populations. However, this assumption cannot be verified with the epidemiological data used by the present study. Indeed, no information is available on the progress of the infections, i.e. if infected individuals developed ECF or not, and, if the case, with which clinical course.

Nevertheless, the proposed approach was able to (i) detect significant associations between the eco-epidemiological predictors tested and the genetics of the analysed populations, (ii) identify genes putatively associated with EFC resistance, and (iii) advance hypotheses about their involvement with ECF endemic stability. Particularly, the significant associations observed with PRKG1 and SLA2 suggests the existence of synergic adaptive mechanisms conferring ECF tolerance: one directed towards the ECF vector *R.*

appendiculatus, and another towards the parasite *T. parva parva*. Preliminary findings on the ancestral origin of the putative genomic variants involved into ECF tolerance were also provided, suggesting a more plausible zebuine and African-sanga evolutionary origin.

To conclude, the present work provided new insights into the eco-epidemiology of ECF in Uganda, highlighted and discussed potential genetic adaptation involved in disease tolerance, and shed some light on the evolutionary origin of ECF tolerance in cattle.

4.6 Acknowledgments

I am grateful to the people involved in the European project NEXTGEN, who allowed both genotyping and epidemiological data collection used here. I also would like to thank Graeme S. Cumming who kindly provided *R. appendiculatus* occurrence dataset used in species distribution modelling.

4.7 Supplementary information

4.7.1 Bioclimatic variables used in *R. appendiculatus* distribution model

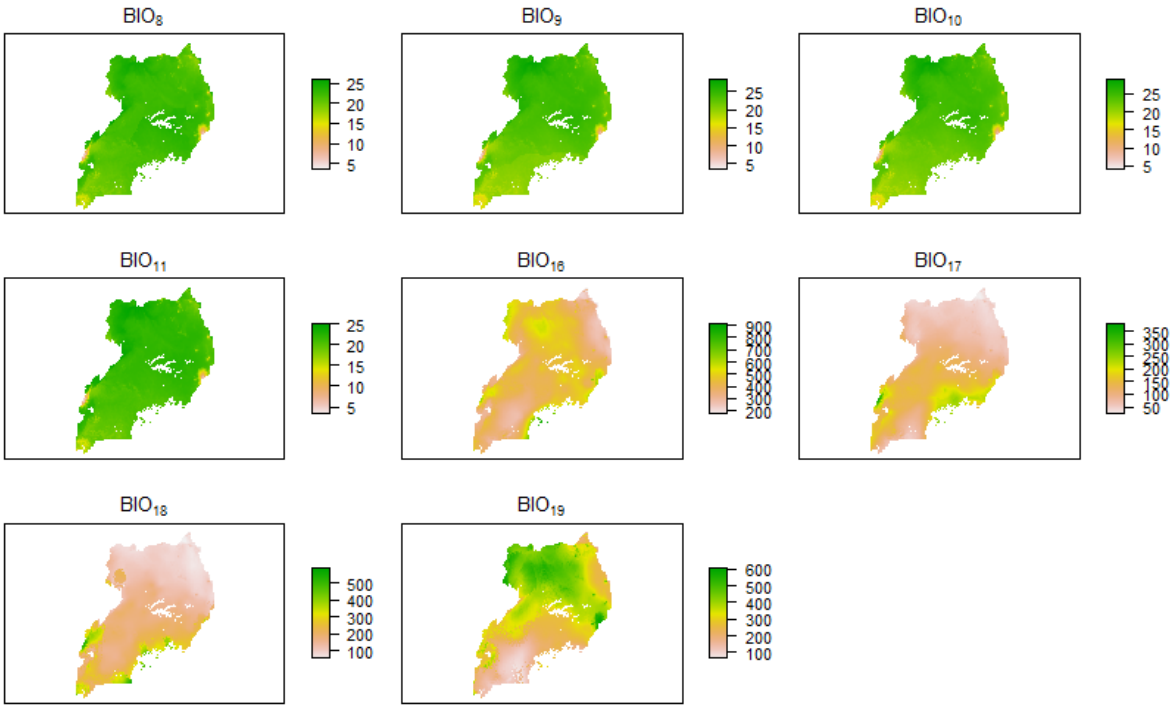

Figure 4.7 Maps of the selected bioclimatic variables used to model ψ_R over Uganda.

4.7.2 NDVI regression analysis results

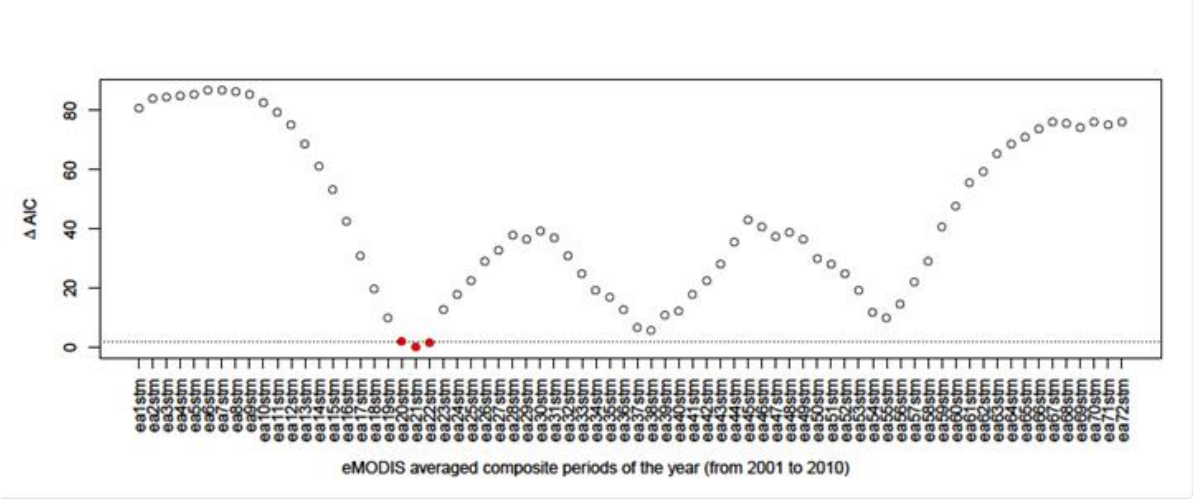

Figure 4.8 Performances of the 72 "eMODIS" annual periods (composites) in explaining the available *S. caffer* occurrences. Each annual period is averaged over the time span 2001-2010. Composite 21 (ea21stm) shows the lowest AIC, *X*-axis reports the original name of the annual periods.

4.7.3 Composition of the population structure dataset

Table 4.8 Composition of the dataset used to study population structure of Ugandan cattle. Table reports the names of the breeds (Breed name), cattle type (Type), samples size (N), geographical provenance (Provenance), and data source (Source).

Breed name	Type	Category	N	Provenience	Source
Holstein	European taurine	ET	50	Europe	Decker et al., (2009, 2014); The Bovine HapMap Consortium *et al.*, (2009); McTavish *et al.* (2013)
Jersey	European taurine	ET	31	Europe	Decker *et al.*, (2009. 2014); The Bovine HapMap Consortium *et al.* (2009); McTavish *et al.* (2013)
Hereford	European taurine	ET	50	Europe	Decker et al. (2009. 2014); The Bovine HapMap Consortium *et al.* (2009); Gautier *et al.* (2010); McTavish *et al.* (2013)
Baoule	African taurine	AT	29	Africa (Burkina Faso)	Gautier *et al.* (2009); Decker *et al.* (2014)
Lagune	African taurine	AT	30	Africa (Benin)	Gautier *et al.* (2009); Decker *et al.* (2014)
N'dama	African taurine	AT	56	Africa (Ivory Coast. Burkina Faso)	Gautier *et al.* (2009. 2010); Decker *et al.* (2014)
Somba	African taurine	AT	30	Africa (Togo)	Gautier *et al.* (2009); Decker *et al.* (2014)

Muturu	African taurine	AT	13	Africa (Nigeria)	Genotypes from T. Sonstegard. *personal communication*
-	**Sanga**	**AI**	**743**	**Africa (Uganda)**	**NextGen project**
Zebu Bororo	Sanga	AI	23	Africa (Chad)	Gautier *et al.* (2010); Decker *et al.* (2014)
Zebu Fulani	Sanga	AI	30	Africa (Benin)	Gautier *et al.* (2009); Decker *et al.* (2014)
Boran	Sanga	AI	44	Africa (Ethiopia)	McTavish *et al.* (2013); Decker *et al.* (2014)
Red Bororo	Sanga	AI	4	Africa (Nigeria)	Genotypes from T. Sonstegard. *personal communication*
Sokoto Gudali	Sanga	AI	6	Africa (Nigeria)	Genotypes from T. Sonstegard. *personal communication*
Nganda	Sanga	AI	19	Africa (Uganda)	Genotypes from T. Sonstegard and H. J. Huson. *personal communication*
Sahiwal	Sanga	AI	21	Africa (Kenya/Uganda)	Genotypes from T. Sonstegard and H. J. Huson. *personal communication*
Serere/Teso Zebu	Sanga	AI	15	Africa(Uganda)	Genotypes from T. Sonstegard and H. J. Huson. *personal communication*
Yakanaji	Sanga	AI	13	Africa (Nigeria)	Genotypes from T. Sonstegard. *personal communication*
Bunaji	Sanga	AI	4	Africa (Nigeria)	Genotypes from T. Sonstegard. *personal communication*
Karakioja	Sanga	AI	16	Africa(Uganda)	Genotypes from T. Sonstegard and H. J. Huson. *personal communication*
Sahiwal	Indicine	ASI	17	Asia (Pakistan)	Decker *et al.* (2009. 2014); McTavish *et al.* (2013)
Gir	Indicine	ASI	26	Asia (India)	Decker *et al.* (2009. 2014); The Bovine HapMap Consortium *et al.* (2009); Gautier *et al.* (2010); McTavish *et al.* (2013)
Tharparkar	Indicine	ASI	25	Asia (Pakistan)	Decker *et al.* (2014); Genotypes from T. Sonstegard. *personal communication*
Kankraj	Indicine	ASI	10	Asia (India)	Decker *et al.* (2014)
Nelore	Indicine	ASI	50	South America (Brazil)	Decker *et al.* (2009. 2014); The Bovine HapMap Consortium *et al.* (2009); Gautier *et al.* (2010); McTavish *et al.* (2013)

4.7.4 Specification of the likelihood ratio tests using SAMβADA models

Significance of associations between genotypes and environment was evaluated by means of a likelihood ratio test. “Null” and “alternative” models were compared for each genotype. Given

a specific genotype, the "null model" was always specified as

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{v=1}^{n} \beta_v s_{vi}$$

where s_{iv} represents the *i-th* observation of the *v-th* population structure variable, and the "alternative" one as

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_Z z_i + \sum_{v=1}^{n} \beta_v s_{vi}$$

where z_i is the *i-th* observation of the environmental variable Z, and β_Z the estimated regression coefficient for that variable. Such an approach allows the "null" model to be nested within the "alternative" one, being equal to the latter for $\beta_Z = 0$.

A likelihood ratio test was performed for each genotype between the "null" and the "alternative" model to test if the inclusion of the environmental variable led to a significantly improved explanation of the genotype spatial distribution. As SAMβADA returns log-likelihood (LogLik) values by default, the test was specified in the following form:

$$D = -2(\text{LogLik of the "null" model} - \text{LogLik of the "alternative" model})$$

Under the null hypothesis of *D*=0, the difference among log-likelihoods follows a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters between the "alternative" and "null" model. In the present case, *p-values* were derived from a χ^2 for one degree of freedom ("alternative" models having one parameter more than the "null" models). Estimates were done with the R function `pchisq`, by setting the appropriate value for degrees of freedom, and the option `lower` equal to `FALSE`. The latter specification was

necessary to correctly compute the probability of obtaining the observed (or more extreme) *D* values under the null hypothesis.

4.7.5 Model selection for the tested *R. appendiculatus* distribution models

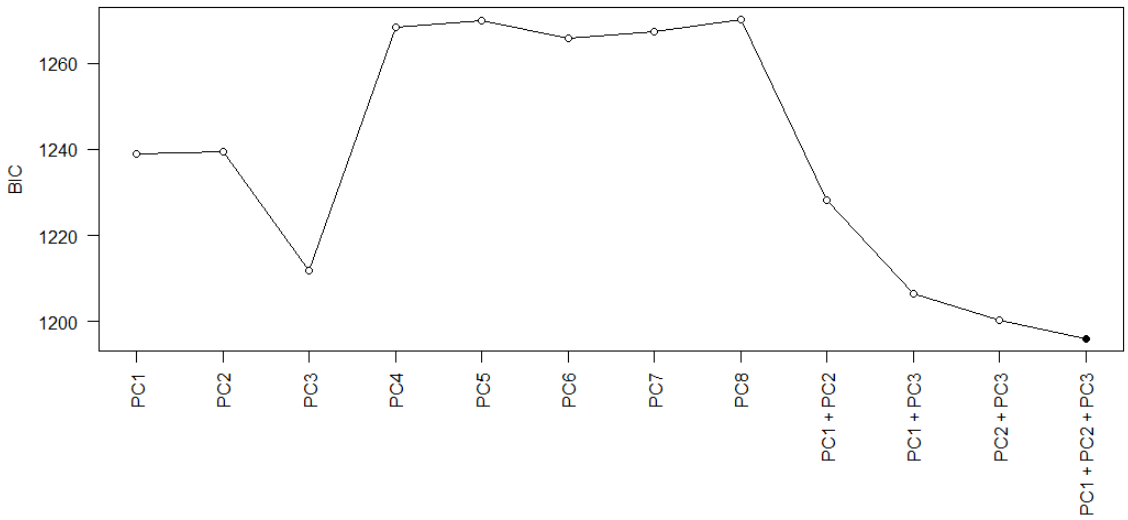

Figure 4.9 *R. appendiculatus* distribution models tested in the present study. Model structure is depicted on the *X*-axis; Bayesian information Criterion (BIC) is reported for each tested model on the *Y*-axis. The model including first, second and third principal components shows the lowest BIC value and was therefore retained to represent ψ_R spatial distribution in Uganda.

4.7.6 Model selection for the tested *S. caffer* distribution models

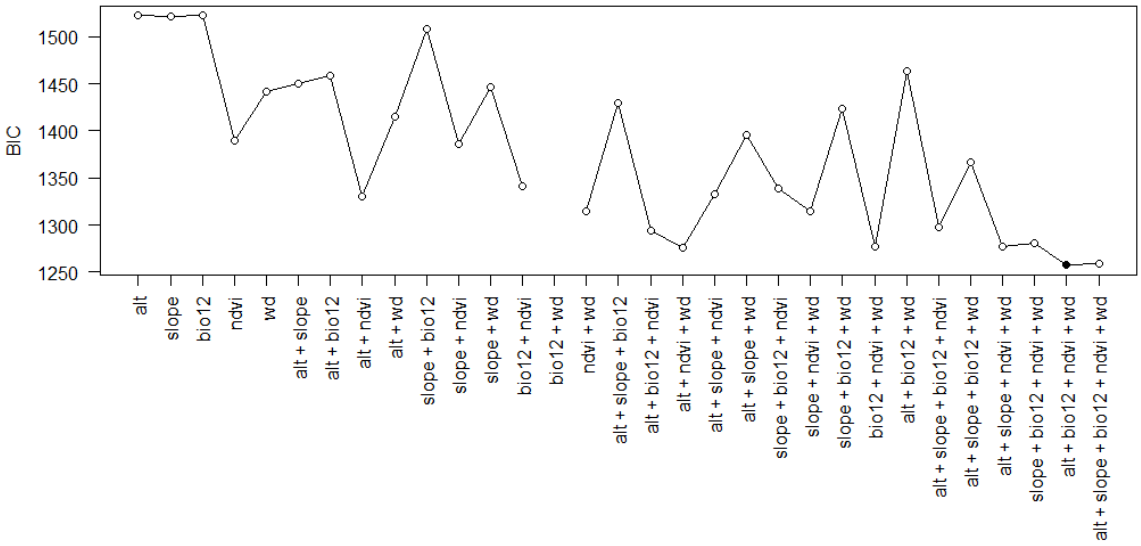

Figure 4.10 *S. caffer* distribution models tested in the present study. Model structure is depicted on the *X*-axis; Bayesian information Criterion (BIC) is reported for each tested model on the *Y*-axis. The model including altitude (alt), annual precipitation (bio12), NDVI (ndvi), and distance from water (Wd) (black point in the plot) shows the lowest BIC value and was therefore retained to represent ψ_S spatial distribution over Uganda. Model including bio12 and Wd failed to converge and does not present any associated BIC.

4.7.7 Transformation of *T. parva parva* infection risk model covariates

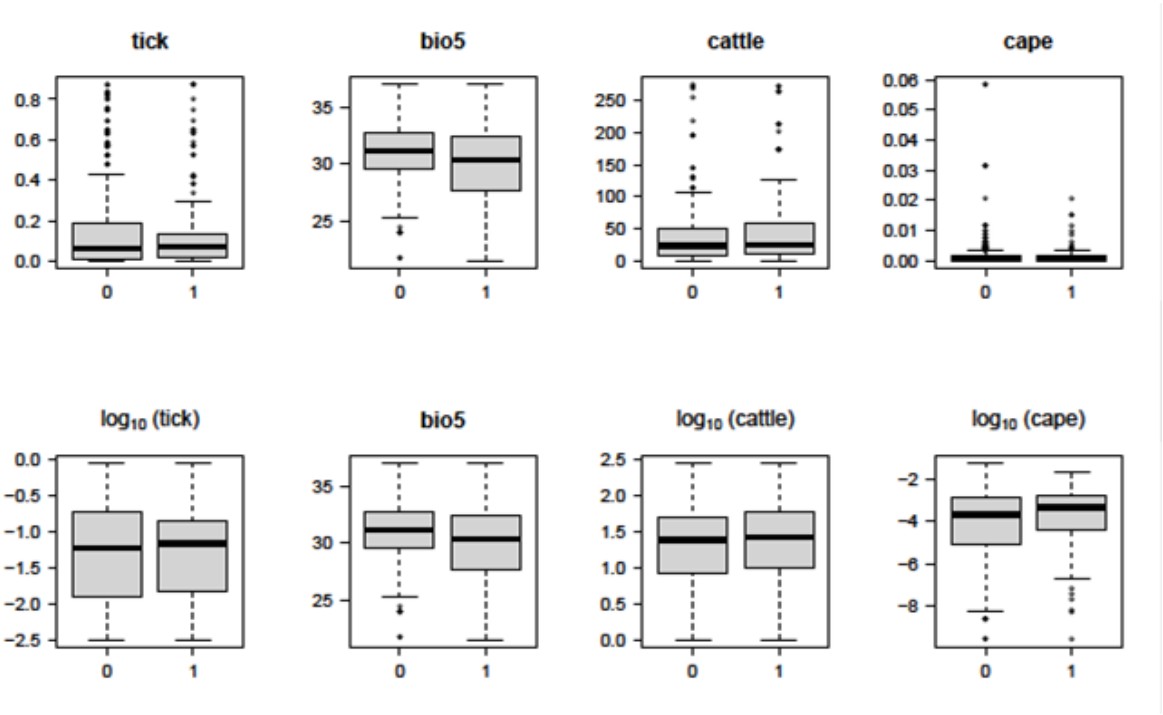

Figure 4.11 Selected predictors of γ were checked prior to modelling for the presence of outliers potentially influencing model parameters estimates. For any given predictor, the check was done separately for the groups of uninfected (0) and infected (1) animals through boxplot visualization. Outliers were defined as the values located outside 1.5 times the interquartile range above the 75% quartile and below the 25% quartile, ψ_R (here "tick"), cattle density ("cattle") and ψ_S ("cape") were transformed on the $\log_{10}$ scale to reduce a potential leverage effect due to the skewness of the distribution. Boxplots of the covariates prior and post transformation are depicted in the upper and lower panel, respectively. Independent Mann-Whitney-Wilcoxon tests were run for each predictor to test the effect of the groups "uninfected" and "infected" on the means of the distributions (H_0: $\mu_0 = \mu_1$, α=0.05). According to the tests, there was a significant difference between the means of the infected and uninfected groups for BIO_5 (*P-value*=5.203E−05) and $\log_{10}(\psi_S)$ (*P-value*=0.0234), while non-significant differences for $\log_{10}(\psi_R)$ (*P-value*=0.6951) and $\log_{10}$(cattle density) (*P-value*=0.2213).

4.7.8 Population structure analyses

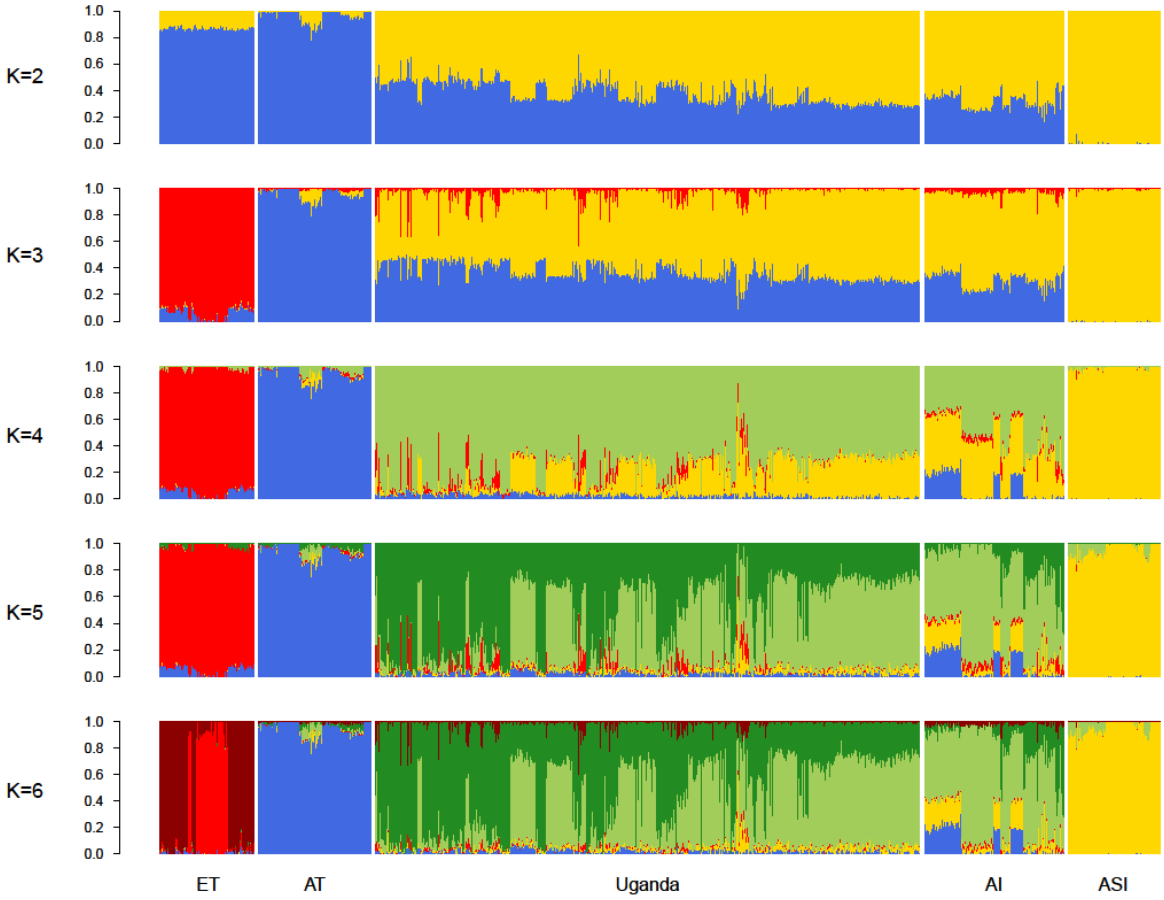

Figure 4.12 ADMIXTURE plots from two to six cluster solutions (K). At K=4, European taurine (in red), African taurine (in blue), sanga (in green) and indicine (in yellow) gene pools can be identified. Successive cluster solutions further split sanga component (at K=5), and European taurine component (at K=6), ET: European taurine breeds; AT: African taurine breeds; Uganda: indigenous Ugandan individuals under study; AI: putative sanga breeds; ASI: indicine breeds from Asia.

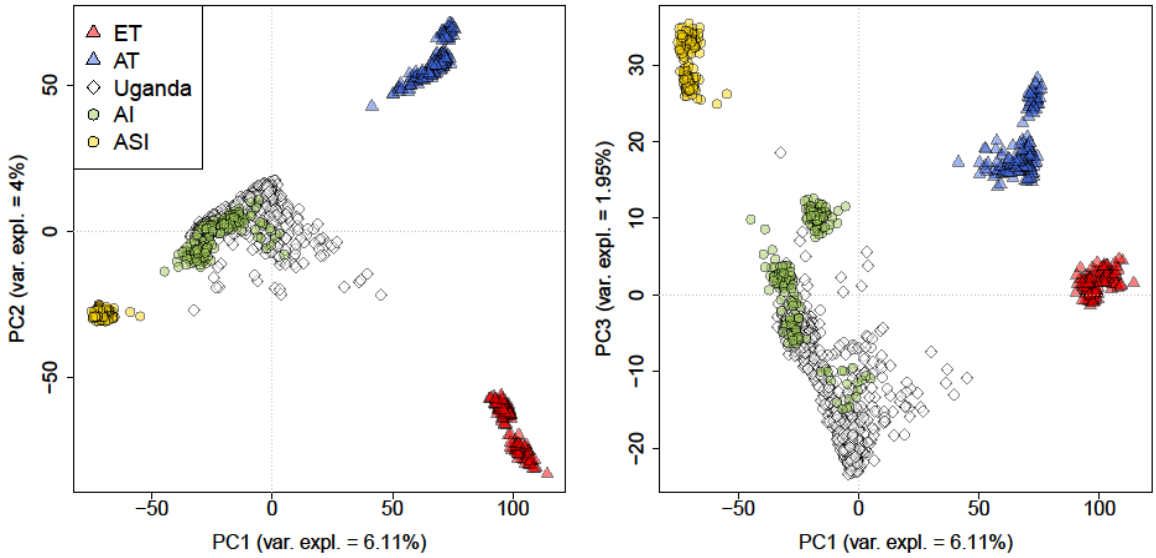

Figure 4.13 From left to right: scatterplots of the first (PC_1) *vs*. second (PC_2), first *vs*. third (PC_3) and second vs. third principal components as derived from the software FLASHPCA (Abraham & Inouye 2014). PC_1 clearly discriminates taurine from indicine breeds; PC_2 African from European taurine breeds. ET: European taurine breeds; AT: African taurine breeds; Uganda: indigenous individuals from Ugandan under study; AI: putative sanga breeds; ASI: indicine breeds from Asia.

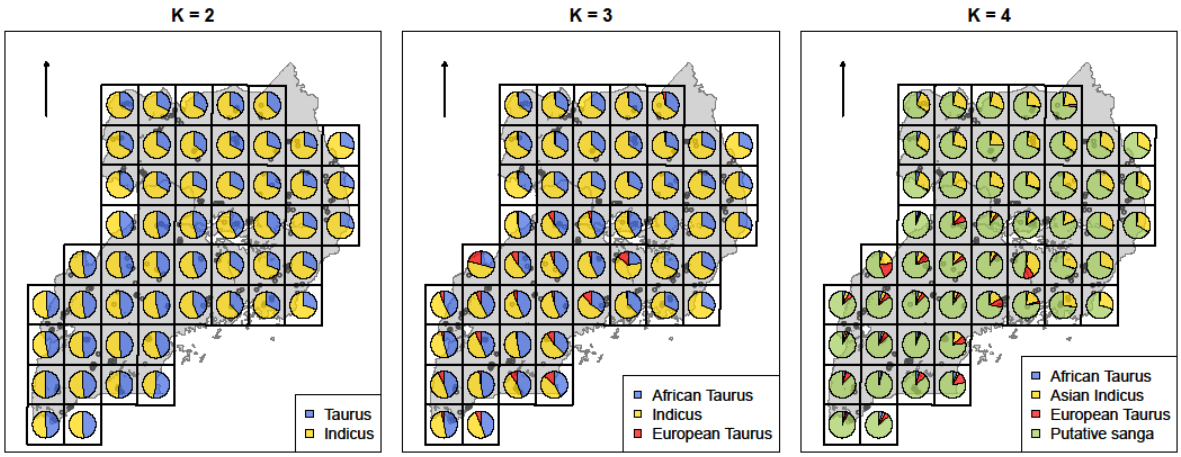

Figure 4.14 Global ancestry composition per cell across Uganda for cluster solutions from K=2 to K=4. Pie chart colours correspond to different ancestral gene pools (African taurine, Asian indicine, European taurine and sanga). At the four clusters solution (*K*=4), a spatial structure appears evident for the sanga and Asian indicine components.

4.7.9 Significant likelihood ratio tests

Table 4.9a SNPs (and related genotypes) significantly associated with *R. appendiculatus* probability of occurrence (ψ_R). Results were considered significant if *p-values* associated with the *D*-statistics (Supplementary 4.7.4) remained below the nominal threshold of 0.05 after correction for multiple testing. Associations are sorted for decreasing values of the *D*-statistics.

SNP ID (genotype)[1]	Chr.[2]	Position[3]	D[4]	P_5(BH)	β_0[6*]	ψ_R[7*]	PC1[8*]	PC2[9*]	PC3[10*]
ARS-BFGL-NGS-18933 (GG)	29	34650967	28.607	0.005	-9.272	10.365	-1.401	-0.705	-1.344
Hapmap51626-BTA-73514 (AA)	5	48834486	27.442	0.005	6.036	-6.006	0.085	0.804	-0.996
Hapmap51626-BTA-73514 (AG)	5	48834486	27.442	0.005	-6.036	6.006	-0.085	-0.804	0.996
Hapmap51479-BTA-66720 (GG)	5	64330943	26.833	0.005	4.877	-4.676	0.106	0.864	-0.941
Hapmap55537-rs29016129 (GG)	5	64380551	26.833	0.005	4.877	-4.676	0.106	0.864	-0.941
BTA-46975-no-rs (GG)	5	68220538	26.173	0.006	5.986	-5.557	0.123	0.761	-1.57
BTA-46975-no-rs (CG)	5	68220538	25.759	0.007	-6.226	5.793	-0.221	-0.719	1.657
ARS-BFGL-NGS-11580 (CC)	1	114981065	24.841	0.007	4.508	-4.133	0.04	0.78	-1.166
Hapmap51479-BTA-66720 (AG)	5	64330943	24.812	0.007	-4.953	4.641	-0.178	-0.763	0.966
Hapmap55537-rs29016129 (AG)	5	64380551	24.812	0.007	-4.953	4.641	-0.178	-0.763	0.966
ARS-BFGL-BAC-6188 (AA)	18	38850678	24.747	0.007	1.214	-2.049	0.439	-0.242	-0.185
Hapmap50589-BTA-119599 (AG)	15	7989843	24.477	0.007	-5.226	4.821	-0.226	-0.617	1.279
Hapmap36616-SCAFFOLD310212_1822 (AA)	5	23171537	23.856	0.01	5.786	-5.456	-0.283	0.611	-0.936
UA-IFASA-6140 (AA)	7	102472846	23.646	0.01	2.733	6.062	0.525	-0.028	-0.434
Hapmap50589-BTA-119599 (GG)	15	7989843	23.15	0.012	5.089	-4.606	0.177	0.66	-1.192
BTB-00839408 (AG)	22	18978658	22.586	0.014	-5.983	5.685	-0.336	-0.611	1.113
BTB-00839408 (AA)	22	18978658	22.586	0.014	5.983	-5.685	0.336	0.611	-1.113
UA-IFASA-5221 (GG)	5	18739471	22.267	0.016	4.807	-4.386	0.354	0.745	-0.85
Hapmap34056-BES2_Contig421_810 (AG)	1	138178130	21.966	0.017	-5.527	4.947	0.21	-0.618	1.091
Hapmap34056-BES2_Contig421_810 (GG)	1	138178130	21.966	0.017	5.527	-4.947	-0.21	0.618	-1.091
ARS-BFGL-NGS-11580 (AC)	1	114981065	21.685	0.017	-4.559	3.954	-0.009	-0.757	1.205
Hapmap57868-rs29020458 (AA)	24	22746291	21.675	0.017	-1.163	1.855	0.362	-0.089	-0.068
BTA-97369-no-rs (GG)	14	25887784	21.641	0.017	-0.459	-1.964	0.071	-0.116	0.252
BTB-00292673 (AA)	7	4953801	21.101	0.022	-69.144	86.381	5.517	-8.483	4.851
ARS-BFGL-BAC-6188 (CC)	18	38850678	20.782	0.024	-4.248	3.811	-0.445	0.796	-0.287
BTB-01283856 (AG)	12	65131442	20.666	0.024	-5.301	4.578	-0.001	-0.099	1.392
BTB-01283856 (GG)	12	65131442	20.666	0.024	5.301	-4.578	0.001	0.099	-1.392
BTB-01058465 (GG)	1	113745976	20.318	0.025	3.954	-3.511	0.297	0.859	-0.806
BTB-01058465 (AG)	1	113745976	20.318	0.025	-3.954	3.511	-0.297	-0.859	0.806
ARS-BFGL-NGS-37845 (AG)	5	48633731	20.308	0.025	-6.562	5.87	-0.594	-0.713	1.572
ARS-BFGL-NGS-37845 (AA)	5	48633731	20.308	0.025	6.562	-5.87	0.594	0.713	-1.572

ARS-BFGL-NGS-103237 (AA)	8	87067969	20.231	0.026	-9.864	-48.49	-0.672	-0.203	0.69
ARS-BFGL-NGS-37889 (AA)	9	10370879	20.168	0.026	4.558	-4.002	0.293	0.762	-0.821
BTB-01109852 (AG)	14	15585398	20.056	0.026	-4.562	3.955	-0.256	-0.425	1.102
ARS-BFGL-NGS-32909 (CC)	5	67846632	19.931	0.027	4.744	-4.092	0.142	0.637	-1.042
ARS-BFGL-NGS-32909 (AC)	5	67846632	19.931	0.027	-4.744	4.092	-0.142	-0.637	1.042
UA-IFASA-6140 (AG)	7	102472846	19.716	0.029	-2.708	-5.334	-0.492	0.047	0.373
Hapmap36616-SCAFFOLD310212_1822 (AG)	5	23171537	19.639	0.029	-5.793	5.087	0.331	-0.57	1.031
Hapmap50904-BTA-17187 (AA)	1	124692274	19.632	0.029	2.619	-2.313	-0.255	-0.074	-0.531
ARS-BFGL-NGS-402 (GG)	29	35698376	19.561	0.029	2.169	-2.178	-0.14	0.476	-0.448
ARS-BFGL-NGS-110339 (AA)	1	111495891	19.521	0.029	3.467	-3.013	0.029	0.676	-0.91
ARS-BFGL-NGS-11845 (AA)	27	21512601	19.276	0.032	5.262	-4.664	0.309	0.848	-0.671
ARS-BFGL-NGS-16947 (AA)	15	26629340	19.053	0.035	4.868	-3.985	-0.298	0.815	-1.086
Hapmap39895-BTA-15668 (CC)	5	13311842	18.725	0.039	5.563	-5.094	-0.131	1.103	0.437
Hapmap39895-BTA-15668 (AC)	5	13311842	18.725	0.039	-5.563	5.094	0.131	-1.103	-0.437
ARS-BFGL-NGS-110339 (AC)	1	111495891	18.669	0.039	-3.509	2.979	-0.004	-0.678	0.938
BTB-01956180 (AG)	27	43656445	18.664	0.039	-0.987	1.695	-0.024	-0.098	0.052
UA-IFASA-5221 (AG)	5	18739471	18.645	0.039	-4.779	4.093	-0.336	-0.723	0.883
ARS-BFGL-NGS-99064 (AA)	1	44813737	18.587	0.039	0.351	1.763	-0.395	0.087	-0.013
ARS-BFGL-NGS-63882 (GG)	2	135994305	18.376	0.043	6.945	-5.619	0.37	0.159	-2.585
Hapmap34409-BES7_Contig244_858 (AA)	1	120149924	18.213	0.044	4.258	-3.515	-0.082	0.758	-0.907
Hapmap39826-BTA-37247 (CC)	15	12975036	18.186	0.044	3.195	-2.786	-0.253	0.356	0.025
ARS-BFGL-NGS-39898 (GG)	22	1319636	18.165	0.044	-0.475	-1.802	0.035	-0.127	0.116
Hapmap39826-BTA-37247 (AC)	15	12975036	18.157	0.044	-3.232	2.795	0.251	-0.334	0.073
ARS-BFGL-NGS-16947 (AC)	15	26629340	18.155	0.044	-4.84	3.928	0.295	-0.786	0.991
Hapmap50904-BTA-17187 (AG)	1	124692274	18.126	0.044	-2.614	2.241	0.255	0.078	0.483
BTA-113604-no-rs (AA)	26	8356096	18.024	0.046	-7.089	6.883	-1.157	-0.404	0.252
BTA-60607-no-rs (AA)	25	6742260	17.967	0.046	-0.823	-1.996	-0.207	0.045	0.123
Hapmap31116-BTA-143121 (AA)	8	75973285	17.854	0.048	2.105	-1.974	-0.362	-0.215	-0.157
ARS-BFGL-NGS-104610 (AG)	11	104293559	17.742	0.049	-0.272	-1.7	-0.127	0.078	0.192
Hapmap51155-BTA-11643 (AA)	24	38086180	17.721	0.049	-364.868	501.77	24.71	-32.84	27.863
ARS-BFGL-NGS-37889 (AT)	9	10370879	17.695	0.049	-4.551	3.821	-0.281	-0.767	0.788
ARS-BFGL-BAC-31319 (AA)	23	4847028	17.683	0.049	-0.973	-2.1	-0.238	-0.131	0.184

[1]Name of the marker (and genotype) associated with ψ_R. [2]Chromosome where the marker is located. [3]Position of the marker on the chromosome. [4]Likelihood ratio test statistics. [5]*P-value* associated to the likelihood ratio test statistics after Benjamini-Hochberg (BH) correction for multiple testing. [6]Model intercept as estimated by SAMβADA. [7]Regression coefficient associated to the conditional effect of ψ_R on the genotype spatial occurrence. [8]Regression coefficient associated to the effect of the first principal component (a positive sign means association with the zebu gene pool). [9]Regression coefficient associated to the effect of the second principal component (a negative sign indicates association with the European taurine gene pool). [10]Regression coefficient associated to the effect of the third principal component (a negative sign indicates association with the African taurine gene pool). *Regression

coefficients are expressed on the logit scale.

Table 9b SNPs (and related genotypes) significantly associated with *T. parva parva* infection risk (γ). Associations are sorted for decreasing values of the *D*-statistics.

SNP ID (genotype)	Chr.	Position	*D*	*P*(BH)	β_0^*	γ^{1*}	PC1*	PC2*	PC3*
ARS-BFGL-NGS-112656 (AA)	13	66336246	26.507	0.019	1.799	-7.131	-0.295	0.282	-0.032
ARS-BFGL-NGS-110102 (GG)	13	66370867	24.254	0.019	1.76	-6.748	-0.263	0.286	-0.018
BTA-33234-no-rs (GG)	13	66291997	24.06	0.019	1.889	-6.881	-0.239	0.286	-0.044
ARS-BFGL-NGS-24867 (AA)	13	66395465	24.045	0.019	1.824	-6.785	-0.236	0.244	-0.121
Hapmap39482-BTA-36746 (CC)	15	40279014	24.01	0.019	5.452	-17.615	-1.162	0.495	-0.622
Hapmap39482-BTA-36746 (AC)	15	40279014	24.01	0.019	-5.452	17.615	1.162	-0.495	0.622
BTB-00384802 (AA)	9	34050782	23.05	0.027	-0.42	-6.085	0.05	-0.074	0.056
BTB-01298953 (AA)	4	54930726	21.786	0.045	1.243	6.926	0.166	0.171	-0.249

[1]Regression coefficient associated to the effect of infection probability γ on the genotype spatial distribution. *Regression coefficients are expressed on the logit scale.

4.7.10 Quantile-Quantile plots of the likelihood ratio tests

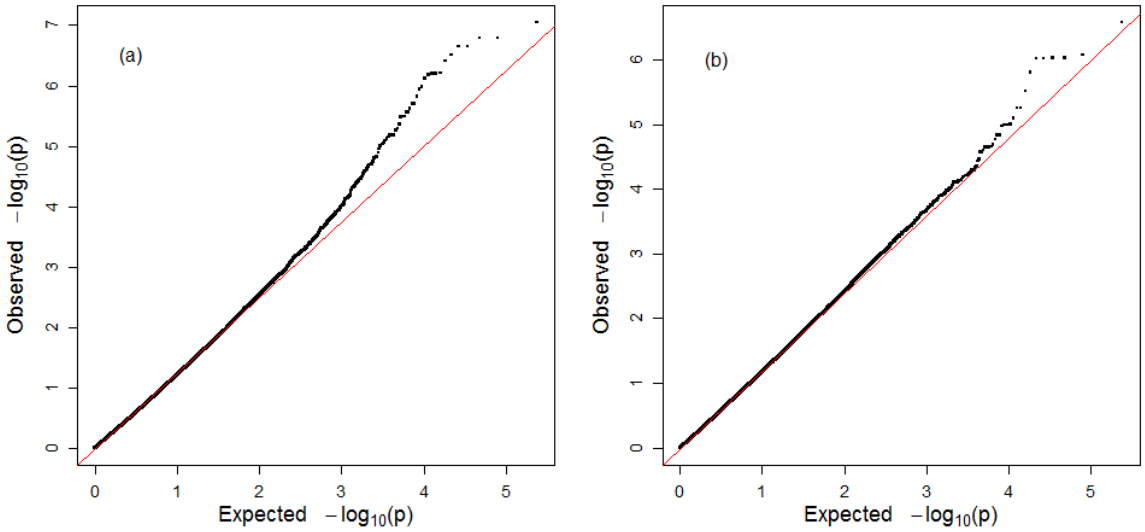

Figure 4.15 Quantile-Quantile plots of the genotype-environment association studies regarding ψ_R (**a**) and γ (**b**). Each point is relative to a single likelihood ratio test (as specified in Supplementary 4.7.4). *Y*-axis reports the sorted *p-values* associated to the test statistics (i.e. the quantiles of the observed *p-values* distribution), while *X*-axis reports the sorted *p-values* derived from a χ^2 distribution with one degree of freedom (i.e. the quantiles of the expected *p-values* distribution). The red line depicts coincidence between observed and expected quantiles, so that points away from the line identify discrepancies among the observed and expected distributions. Observed *p-values* from the ψ_R study suggest a higher divergence from the expectation then *p-values* from γ association study. *P-values* are reported prior multiple testing correction and on the $-\log_{10}$ scale.

5. General conclusions

5.1 Summary

Three main subjects have been addressed in the present thesis:

1. **Chapter 2** reviewed a number of prioritization methods addressing biodiversity crisis in natural and agricultural systems, proposed a general classification scheme for the reviewed methods, provided a decision support system in the form of a decision tree, and discussed methodological integrations which could lead to novel approaches for biological prioritization at the within-species level.
2. **Chapter 3** reported a case study where the performances of a new, species-specific SNP-chip (the Axiom® Buffalo Genotyping Array 90K) was tested to characterize water buffalo genomic diversity. This study provided genomic estimates of genetic variability, investigated population structure and phylogenetic relationships among over 30 populations worldwide, and provided hypotheses about the migrations routes following domestication events.
3. **Chapter 4** reported a case study aimed at characterizing the genetic bases underlying tolerance towards an endemic disease affecting indigenous cattle populations of sub-Saharan Africa. This study coupled statistical modelling techniques from spatial ecology (species distribution models), epidemiological modelling and landscape genomics. Two putative genes involved into local adaptation mechanisms toward the disease were identified.

5.2 Local adaptation to ECF in Uganda: general considerations, limits and future directions

Some indigenous cattle populations from Eastern Africa are able to recover from East Coast Fever (ECF) (Ndungu *et al.* 2005; Bahbahani & Hanotte 2015), which is otherwise responsible for 90-100% mortality when affecting susceptible populations (Olwoch *et al.* 2008). I specifically referred to the ability of "controlling the course of disease" (Ndungu *et al.* 2005) as a potential case of local adaptation, because (i) experimental proof shows that, for equal parasite doses, indigenous populations from ECF endemic areas survive and recover from infection in shorter times then the same breeds native to ECF-free regions (Ndungu *et al.* 2005), and (ii) host-parasite systems are known to promote local adaptation, by reciprocally exerting a strong and spatially heterogeneous selection (Kawecki & Ebert 2004). As a consequence, phenotypic differences conferring differential fitness are rarely due to phenotypic plasticity, and a limited number of major genes are expected to be involved (Kawecki & Ebert 2004).

The study was based on the molecular data provided by the NEXTGEN project, and relied on a subset of epidemiological information collected by Kabi and colleagues (2014). All the sampled individuals (including the infected ones) were phenotypically described to be "apparently healthy", thus supporting the rationale underlying the genotype-environment association study adopted in my work: the animals inhabiting areas with major risk of becoming infected are subjected to a higher selective pressure than animals living in ECF-free areas, and since they look healthy, they are expected to be disease-tolerant due to local adaptation.

The combination of species distribution modelling and landscape genomics showed the potential of identifying candidate genes for local adaptation, and could be taken into consideration for any study focusing on the interaction between species with overlapping spatial distributions. Therefore, the approach might be tested in the cases of symbiotic relationships (i.e. mutualism, parasitism and commensalism) or even competition among species in natural systems.

However, some limitations are present and deserve further consideration when looking at the results presented in **Chapter 4**, in particular:

1. The assumption "higher infection risk/presence of locally adapted populations" is hardly verifiable with the epidemiological data available, since no follow-up information exists regarding the progress of the infections (e.g. if some animals actually developed ECF and survived or not).
2. A challenge concerns how to correct the infection risk estimates with the epidemiological records' reliability. In particular, a subset of 170 paired independent trials resulted in a Kappa statistics (Lachin 2004) equal to 0.94 (95% confidence intervals: 0.88-0.99), suggesting the overall agreement between the laboratories where the paired tests were performed (Makerere University and Biosciences Eastern and Central Africa, Nairobi, respectively). Some approaches have been proposed to estimate the expected reliability between independent raters on the basis of meaningful predictors of agreement (Lipsitz *et al.* 2003). Provided relevant information is firstly retrieved about the concerned laboratories, these approaches could provide an "expected agreement" variable to be integrated as covariate in the

infection risk model.

3. A further point of concern is represented by the seasonal movements involving livestock. A transhumance takes place in Uganda during the dry season (from December to February and from June to August), when farmers migrate southwards to find fresh pastures and farm residues (Christopher Mukasa, *personal communication*). While in the South, chances exist that animals become infected and transport the parasite in the North where it can be detected (see Soudré et al., 2013 for analogies with trypanosomiasis in Burkina Faso). This transhumance-linked effect may be particularly worrying as it could induce spurious correlations with environmental conditions that are not actually associated with *T. parva parva* survival. However, recorded sampling dates suggest that the animals in the Northern grid cells were sampled in January, July, August and December 2011/2012, during the dry season. This would indicate the "Northern" infections to actually mirror local environmental features, and not to derive from the South. Nevertheless, no comprehensive information exists regarding the transhumant behaviour of the single famers, and it is difficult—with the current data—to infer if transhumance took place in years preceding NEXTGEN sampling.
4. The occurrence records at the basis of *R. appendiculatus* and *S. caffer* distribution models present small sample sizes, inhomogeneity in the records' dates, and some (hardly quantifiable) levels of spatial bias. That said, retrieving such records was not trivial, and the alternative would have been to exclude relevant predictors (i.e. ψ_R and ψ_S) from the *T. parva parva* infection risk model. Therefore, an improvement

for these models would be to retrieve and add new *R. appendiculatus* and *S. caffer* presence data. The estimation of *S. caffer* actual distribution could be further improved by accounting for the effect of the natural reserve boundaries and human presence (e.g. including variables related to human population density and proximity to agricultural fields).

5. *T. parva parva* infection risk model does not explicitly account for the potential effect of the farming system, which was proven to be associated with ECF prevalence (Rubaire-Akiiki *et al.* 2006; Gachohi *et al.* 2012). Nevertheless, any unmeasured effect acting within the sampling sites (including the farming system) should have been caught by the random intercepts estimated for each farm.

Despite these limitations, results obtained seem robust in terms of both literature findings and coherence with the parasite-host system studied. Indeed, the counterintuitive relationship between *R. appendiculatus* occurrence probability and *T. parva parva* infection risk finds support in Magona *et al.* (2008) study, where density in *R. appendiculatus* burden was associated with a reduced probability of seroconversion to *T. parva* in the South-East of Uganda. At the same time, tick resistance has been associated in several occasions with pro-inflammatory genes like TLR-5, chemokine ligand-2 and chemokine receptor-1 (Bahbahani & Hanotte 2015). In this regard, PRKG1 gene falls into such a genic category being potentially involved into the inflammatory response activated by the tick bite at the cutaneous level. Moreover, the implication of SLA2 into cellular pathways controlling and downregulating humoral and cell-mediated immune responses (Holland *et al.* 2001; Marton *et al.* 2015; Kazi *et al.* 2015) appears consistent with ECF, a disease which is able to cause an uncontrolled

proliferation of T and B cells (Baumgartner *et al.* 2003).

Validation remains a major concern of genotype-environment association studies (Rellstab *et al.* 2015). Here, the highlighted associations might be tested (i) by analysing independent populations coming from other countries (e.g. Kenya, where autochthonous cattle inhabit both ECF non-endemic and endemic areas; see Gachohi *et al.* 2012), (ii) by comparing the expression of the concerned genes in indigenous populations from areas with high tick/*T. parva parva* burden against populations from areas with low tick/*T. parva parva* burden, or (iii) by implementing reciprocal transplant experiments comparing putative tick-resistant/ECF-tolerant breeds versus exotics, as well as tick-resistant/ECF-tolerant breeds in their respective native and non-native sites (Rellstab *et al.* 2015). In the latter case, however, experimental plan might result particularly complex, and comparisons should be carefully designed before any practical implementation. Furthermore, support to the role of temperature on *T. parva parva* development might be obtained through field trials ideally comparing development rates in tick populations from the South-East and South-West of the country in different seasons of the year.

5.3 The future of conservation in livestock

Industrial livestock breeds are replacing locally adapted populations in developing countries because of increasing socio-economic pressures and their higher productive performances (Kabi *et al.* 2014; Mwai *et al.* 2015). As a consequence, the unique gene pools of indigenous populations are disappearing, leading a number of local breeds on the edge of extinction.

Next generation sequencing approaches represent a relatively new tool to address such a process of biodiversity depletion at the species level (Allendorf *et al* 2010), but promise to become the gold standard for characterizing and managing AnGR in the near future (Bruford *et al.* 2015). Therefore, I speculate that the conservation of livestock biodiversity will be more and more based on the use of genomic information, because of a number of advantages over more obsolete genotyping technologies:

1) Genomic diversity can be now characterized with increased accuracy on the basis of tens of thousands of markers, by gaining new insights into the demographic and adaptive history of the studied populations (Kristensen *et al.* 2015). Provided that the effects of ascertainment bias are adequately considered, priorities aiming at preserving the most diverse populations could be highlighted easily. The study on *B. bubalis* (**Chapter 3**) provides a good example in this direction, where two hotspots of genetic diversity were discovered to correspond to the putative domestication centres of *B. bubalis bubalis* (North-western India) and *B. bubalis carabanensis* (Thailand). The Indian (RIVPH_IN_MUR), Pakistani (RIVPK_AZK, RIVPK_KUN, RIVPK_NIL) and Thai populations (SWATH_THS, SWATH_THT) could be prioritized to preserve the species adaptive potential with regard to (i) future environmental and socio-economic change and (ii) the alarming census decline reported for several water buffalo populations worldwide (Borghese 2011).
2) Inbreeding depression, a serious threat for fitness and productivity in some livestock species, could be monitored through accurate estimation of individual relatedness (Kristensen *et al.* 2015). Therefore, focused breeding schemes can be devised to

preserve or increase genomic diversity, and recover N_e of both commercial and local breeds above the dangerous threshold of 50. At the same time, causal mutations of deleterious traits can be more easily detected, and carriers of deleterious recessive alleles identified.

3) SNP arrays are able to increase accuracy in assessing genetic uniqueness at both neutral and adaptive markers. Again, *B. bubalis* study (**Chapter 3**) provides a good example, since the 90K Affymetrix Axiom® Buffalo Genotyping Array was able to detect distinct gene pools like the indigenous Mediterranean buffalo (section 3.8.7), an ancient and locally adapted breed potentially deserving special management for conservation.
4) The capability of directly addressing adaptive variation expands the possibilities of adaptive management with regard to environmental and socio-economic change. For instance, the detection of adaptive variants, together with environmental, epidemiologic or socio-economic projections might lead to the identification of vulnerable populations deserving prioritization for conservation. Once identified, the adaptive variants might be introgressed into the vulnerable populations through targeted cross-breeding or genome-editing techniques.
5) Prioritization process might benefit from information derived from next generation sequencing approaches. Integrating Funk *et al.*'s approach (**Chapter 2**) with a genotype-environment association study (**Chapter 4**) would result in a five-steps prioritization process which might prove useful especially for those livestock breeds reared under an extensive management regime, the five steps being: (i) the

identification of candidate genes for local adaptation through the genotype-environment association study; (ii) the use of the whole set of markers available (i.e. neutral plus adaptive loci) to investigate global ancestry and identify evolutionary significant units (ESUs); (iii) the identification of the putatively neutral loci through a global F_{ST} analysis based on the highlighted ESUs; (iv) the use of the set of neutral markers to delineate management units (MUs) within (or across) the ESUs; (v) the investigation of the adaptive differentiation among MUs by relying on the SNPs highlighted in point (i); to this purpose, a global ancestry analysis or a neighbour-joining dendrogram could be employed to investigate clustering among MUs. Finally, the identified clusters would provide the basis for subsequent prioritization ranking and actions.

The indigenous cattle populations analysed in **Chapter 4** would probably benefit from this prioritization pipeline, since an allochthonous genetic introgression from Europe might affect ECF-adaptive genomic regions (section 4.7.8 and Figure 4.6) and undermine endemic stability in the whole area. Thus, the identification of tolerant clusters among defined MUs would indicate where useful gene variants for conserving endemic stability can be found, allowing genetic improvement of commercial breeds, and coping with incoming challenges imposed by environmental change.

Finally, I believe livestock conservation might be faced through a landscape perspective too. Particularly, the use of similarity measures discussed in **Chapter 2** could be explored in future research for investigating and comparing breed richness in different geographical areas, and

evidencing priority regions for livestock conservation. This approach might also be extended to several livestock species at a time, by ideally providing a multi-species approach able to evidence areas of high conservation concern for agricultural biodiversity.

6. Bibliography

Abraham G, Inouye M (2014) Fast Principal Component Analysis of Large-Scale Genome-Wide Data. *PLoS ONE*, **9**, e93766.

Ackery P, Vane-Wright R (1984) Milkweed Butterflies. *British Museum (Natural History), London*.

Aho K, Derryberry D, Peterson T (2014) Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, **95**, 631–636.

Ajmone-Marsan P, Garcia JF, Lenstra JA (2010) On the origin of cattle: how aurochs became cattle and colonized the world. *Evolutionary Anthropology: Issues, News, and Reviews*, **19**, 148–157.

Ajmone-Marsan P, The GLOBALDIV Consortium (2010) A global view of livestock biodiversity and conservation. *Animal Genetics*, **41**, 1–5.

Aken BL, Ayling S, Barrell D *et al.* (2016) The Ensembl gene annotation system. *Database*, **2016**, baw093.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome research*, **12**, 1805–1814.

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics*, **11**, 697–709.

Anderson S (2003) Animal genetic resources and sustainable livelihoods. *Ecological Economics*, **45**, 331–339.

Anderson JT, Lee C-R, Rushworth CA, Colautti RI, Mitchell-Olds T (2013) Genetic trade-offs and conditional neutrality contribute to local adaptation. *Molecular Ecology*, **22**, 699–708.

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R package for genome-wide association analysis. *Bioinformatics*, **23**, 1294–1296.

Bahbahani H, Hanotte O (2015) Genetic resistance: tolerance to vector-borne diseases and the prospects and challenges of genomics. *Rev. Sci. Tech. Off. Int. Epiz.*, **34**, 185–197.

Baldwin CL, Black SJ, Brown C *et al.* (1988) Bovine T Cells, B Cells, and Null Cells Are Transformed by the Protozoan Parasite *Theileria parva*. *Infection and Immunity*, **56**, 462–467

Bakker VJ, Doak DF, Roemer GW *et al.* (2009) Incorporating ecological drivers and uncertainty into a demographic population viability analysis for the island fox. *Ecological Monographs*, **79**, 77–108.

Barbato M (2016) Unravelling the evolutionary history and adaptation of European mouflon and some domestic sheep populations with special emphasis on the ovines of Sardinia. PhD diss. Cardiff University.

Barbato M, Orozco-terWengel P, Tapio M, Bruford MW (2015) SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Frontiers in Genetics*, **6**.

Barker JSF, Moore SS, Hetzel DJS, Evans D, Byrne K (1997) Genetic diversity of Asian water buffalo (*Bubalus bubalis*): microsatellite variation and a comparison with protein-coding loci. *Animal genetics*, **28**, 103–115.

Barve N, Barve V, Jiménez-Valverde A *et al.* (2011) The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, **222**, 1810–1819.

Baselga A (2010) Partitioning the turnover and nestedness components of beta diversity: Partitioning beta diversity. *Global Ecology and Biogeography*, **19**, 134–143.

Baselga A, Orme CDL (2012) betapart: an R package for the study of beta diversity. *Methods in Ecology and Evolution*, **3**, 808–812.

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, **67**, 1–48.

Baumgartner M, Angelisová P, Setterblad N *et al.* (2003) Constitutive exclusion of Csk from Hck-positive membrane microdomains permits Src kinase-dependent proliferation ofTheileria-transformed B lymphocytes. *Blood*, **101**, 1874–1881.

Beja-Pereira A, England PR, Ferrand N *et al.* (2004) African origins of the domestic donkey. *Science*, **304**, 1781–1781.

Bellard C, Bertelsmeier C, Leadley P, Thuiller W, Courchamp F (2012) Impacts of climate change on the future of biodiversity. *Ecology Letters*, **15**, 365–377.

Bennewitz J, Eding H, Ruane J, Simianer H (2007) *Selection of breeds for conservation. In: Utilization and Conservation of Farm Animal Genetic Resources.* Ed. by K. Oldenbrook. Wageningen Academic Publishers, Wageningen, the Netherlands.

Billiouw M, Vercruysse J, Marcotty T *et al.* (2002) *Theileria parva* epidemics: a case study in eastern Zambia. *Veterinary parasitology*, **107**, 51–63.

Blasco A (2008) Breeds in danger of extintion and biodiversity. *Revista Brasileira de Zootecnia*, **37**, 101–109.

Boettcher PJ, Tixier-Boichard M, Toro MA *et al.* (2010) Objectives, criteria and methods for using molecular genetic data in priority setting for conservation of animal genetic resources. *Animal Genetics*, **41**, 64–77.

Bökönyi S (1974) *History of Domestic Mammals in Central and Eastern Europe*. Akadémiai Kiadó, 596 pp.

Bonin A, Nicole F, Pompanon F, Miaud C, Taberlet P (2007) Population Adaptive Index: a New Method to Help Measure Intraspecific Genetic Diversity and Prioritize Populations for Conservation. *Conservation Biology*, **21**, 697–708.

Borghese A (2011) Situation and perspectives of buffalo in the World, Europe and Macedonia. *Macedonian Journal of Animal Science*, **1**, 281–296.

Borquis RRA, Baldi F, de Camargo GMF *et al.* (2014) Water buffalo genome characterization by the Illumina BovineHD BeadChip. *Genetics and Molecular Research*, **13**, 4202–4215.

Boulter N, Hall R (1999) Immunity and vaccine development in the bovine theilerioses. *Advances in Parasitology*, **44**, 41–97.

Boyko AR, Quignon P, Li L *et al.* (2010) A Simple Genetic Architecture Underlies Morphological Variation in Dogs. *PLoS Biology*, **8**, e1000451.

Bring J (1994) How to Standardize Regression Coefficients. *The American Statistician*, **48**, 209–213.

Brisbin A, Bryc K, Byrnes J *et al.* (2012) PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Human Biology*, **84**, 343–364.

Brizuela CM, Ortellado CA, Sanchez TI, Osorio O, Walker AR (1996) Formulation of integrated control of *Boophilus microplus* in Paraguay: analysis of natural infestations. *Veterinary Parasitology*, **63**, 95–108.

Brondyk WH, McKiernan CJ, Fortner KA *et al.* (1995) Interaction cloning of Rabin3, a novel protein that associates with the Ras-like GTPase Rab3A. *Molecular and cellular biology*, **15**, 1137–1143.

Brook B, O'Grady J, Chapman A *et al.* (2000) Predictive accuracy of population viability analysis in conservation biology. *Nature*, **404**, 385–387.

Brooks TM, Mittermeier RA, da Fonseca GAB *et al.* (2006) Global Biodiversity Conservation Priorities. *Science*, **313**, 58–61.

Brossard M, Wikel SK (1997) Immunology of interactions between ticks and hosts. *Medical and Veterinary Entomology*, **11**, 270–276.

Bruford MW, Bradley DG, Luikart G (2003) DNA markers reveal the complexity of livestock domestication. *Nature Reviews Genetics*, **4**, 900–910.

Bruford MW, Ginja C, Hoffmann I *et al.* (2015) Prospects and challenges for the conservation of farm animal genomic resources, 2015-2025. *Frontiers in Genetics*, **6**.

Brütting C, Emmer A, Kornhuber M, et. al. (2016) A survey of endogenous retrovirus (ERV) sequences in the vicinity of multiple sclerosis (MS)-associated single nucleotide polymorphisms (SNPs). *Molecular Biology Reports*, **43**, 827–836.

Burton JA, Hedges S, Mustari AH (2005) The taxonomic status, distribution and conservation

of the lowland anoa *Bubalus depressicornis* and mountain anoa *Bubalus quarlesi*. *Mammal Review*, **35**, 25–50.

Butchart SHM, Walpole M, Collen B *et al.* (2010) Global Biodiversity: Indicators of Recent Declines. *Science*, **328**, 1164–1168.

Caballero A, Toro MA (2002) Analysis of genetic diversity for the management of conserved subdivided populations. *Conservation Genetics*, **3**, 289–299.

Cade BS (2015) Model averaging and muddled multimodel inferences. *Ecology*, **96**, 2370–2382.

Cañón J, Alexandrino P, Bessa I *et al.* (2001) Genetic diversity measures of local European beef cattle breeds for conservation purposes. *Genetics Selection Evolution*, **33**, 1.

Cañón J, Garcia D, Garcia-Atance MA *et al.* (2006) Geographical partitioning of goat diversity in Europe and the Middle East. *Animal Genetics*, **37**, 327–334.

Carwardine J, Wilson KA, Ceballos G *et al.* (2008) Cost-effective priorities for global mammal conservation. *Proceedings of the National Academy of Sciences*, **105**, 11446–11450.

Ceballos G, Ehrlich PR, Barnosky AD *et al.* (2015) Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science Advances*, **1**, e1400253–e1400253.

Chevin L-M, Lande R, Mace GM (2010) Adaptation, Plasticity, and Extinction in a Changing Environment: Towards a Predictive Theory. *PLoS Biology*, **8**, e1000357.

CIESIN (Center for International Earth Science Information Network) Columbia University (2016) Gridded Population of the World, Version 4 (GPWv4): Land and Water Area. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC).

Clutton-Brock J (1999) *A natural history of domesticated mammals*. Cambridge University Press.

Cockrill W (1974) *The husbandry and health of the domestic buffalo*. ros.

Colli L, Joost S, Negrini R *et al.* (2014) Assessing The Spatial Dependence of Adaptive Loci in 43 European and Western Asian Goat Breeds Using AFLP Markers. *PLoS ONE*, **9**, e86668.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics*, **185**, 1411–1423.

Coulson T, Mace G, Hudson E, Possingham H (2001) The use and abuse of population viability analysis. *Trends in Ecology & Evolution*, **16**, 219–221.

Crispo E, DiBattista JD, Correa C *et al.* (2010) The evolution of phenotypic plasticity in response to anthropogenic disturbance. *Evolutionary Ecology Research*, **12**, 47–66.

Cumming GS (1998) Host preference in African ticks (Acari: Ixodida): a quantitative data set. *Bulletin of Entomological Research*, **88**, 379.

Cumming GS (1999a) Host distributions do not limit the species ranges of most African ticks (Acari: Ixodida). *Bulletin of Entomological Research*, **89**, 303–327.

Cumming GS (1999b) The evolutionary ecology of African ticks. Unpublished DPhil Thesis. University of Oxford, Oxford, UK.

Cumming GS (2002) Comparing Climate and Vegetation as Limiting Factors for Species Ranges of African Ticks. *Ecology*, **83**, 255.

De Mita S, Thuillet A-C, Gay L *et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399.

De Vos JM, Joppa LN, Gittleman JL, Stephens PR, Pimm SL (2015) Estimating the normal background rate of species extinction: Background Rate of Extinction. *Conservation Biology*, **29**, 452–462.

Decker JE, McKay SD, Rolf MM *et al.* (2014) Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle. *PLoS Genetics*, **10**, e1004254.

Decker JE, Pires JC, Conant GC *et al.* (2009) Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proceedings of the National Academy of Sciences*, **106**, 18644–18649.

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

Dikmen S, Khan FA, Huson HJ *et al.* (2014) The SLICK hair locus derived from Senepol cattle confers thermotolerance to intensively managed lactating Holstein cows. *Journal of Dairy Science*, **97**, 5508–5520.

Dormann CF, Elith J, Bacher S *et al.* (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 27–46.

Driscoll CA, Macdonald DW, O'Brien SJ (2009) From wild animals to domestic pets, an evolutionary view of domestication. *Proceedings of the National Academy of Sciences*, **106**, 9971–9978.

Eckert AJ, van Heerwaarden J, Wegrzyn JL *et al.* (2010) Patterns of Population Structure and Environmental Associations to Aridity Across the Range of Loblolly Pine (*Pinus taeda* L., Pinaceae). *Genetics*, **185**, 969–982.

Eding H, Crooijmans RPMA, Groenen MAM, Meuwissen THE (2002) Assessing the contribution of breeds to genetic diversity in conservation schemes. *Genetics Selection Evolution*, **34**, 613–633.

Eizaguirre C, Baltazar-Soares M (2014) Evolutionary conservation-evaluating the adaptive potential of species. *Evolutionary Applications*, **7**, 963–967.

Elbeltagy AR, Galal S, Abdelsalam AZ *et al.* (2008) Biodiversity in Mediterranean buffalo using two microsatellite multiplexes. *Livestock Science*, **114**, 341–346.

Elith J, Kearney M, Phillips S (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330–342.

Elith J, Leathwick JR (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

El-Kholy AF, Hassan HZ, Amin AMS, Hassanane MS (2007) Genetic diversity in Egyptian buffalo using microsatellite markers. *Arab J. Biotechnol*, **10**, 219–232.

Epstein H (1969) *Domestic Animals of China*. Commonwealth Agricultural Bureaux, Farnham Royal, Buckinghamshire, UK.

Epstein H (1971) *On the Classification of Cattle. The Origin of Domestic Animals of Africa*. African Publishing Corporation, New York.

Erwin TL (1991) An evolutionary basis for conservation strategies. *Science*, **253**, 750–752.

European Cattle Genetic Diversity Consortium (2006) Marker-assisted conservation of European cattle breeds: an evaluation. *Animal Genetics*, **37**, 475–481.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

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

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

FAO (2011) *Molecular genetic characterization of animal genetic resources. FAO Animal Production and Health Guidelines. No. 9. Rome.* FAO.

FAO (2015) *The Second Report on the State of the World's Animal Genetic Resources for Food and Agriculture*. B.D. Scherf & D. Pilling. FAO Commission on Genetic Resources for Food and Agriculture Assessments, Rome.

Felsenstein J (1989) PHYPIL - Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.

Felsenstein J (2016) PHYPIL *(Phylogeny Inference Package)*. Department of Genome Sciences, University of Washington, Seattle.

Ferrara B (1964) The present situation of buffalo breeding in Italy. *Acta med. Vet. Napoli*, **10**.

Foll M, Gaggiotti O (2008) A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics*, **180**, 977–993.

Food and Agriculture Organization of the United Nations (2012) *Report of a consultation on the definition of breed categories*. FAO, Rome.

Franklin IR, Frankham R (1998) How large must populations be to retain evolutionary potential? *Animal Conservation*, **1**, 69–73.

Frantz LAF, Mullin VE, Pionnier-Capitan M *et al.* (2016) Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science*, **352**, 1228–1231.

Frantz LAF, Schraiber JG, Madsen O *et al.* (2015) Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. *Nature Genetics*, **47**, 1141–1148.

Fraser DJ, Weir LK, Bernatchez L, Hansen MM, Taylor EB (2011) Extent and scale of local adaptation in salmonid fishes: review and meta-analysis. *Heredity*, **106**, 404–420.

Freeman AR (2006) Assessing the Relative Ages of Admixture in the Bovine Hybrid Zones of Africa and the Near East Using X Chromosome Haplotype Mosaicism. *Genetics*, **173**, 1503–1510.

Frichot E, François O (2015) LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, **6**, 925–929.

Frichot E, Schoville SD, Bouchard G, Francois O (2013) Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution*, **30**, 1687–1699.

Frison EA, Cherfas J, Hodgkin T (2011) Agricultural Biodiversity Is Essential for a Sustainable Improvement in Food and Nutrition Security. *Sustainability*, **3**, 238–253.

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, **27**, 489–496.

Gachohi JM, Kitala PM, Ngumi PN, Skilton RA (2011) Environment and farm factors associated with exposure to *Theileria parva* infection in cattle under traditional mixed farming system in Mbeere District, Kenya. *Tropical Animal Health and Production*, **43**, 271–277.

Gachohi J, Skilton R, Hansen F *et al.* (2012) Epidemiology of East Coast fever (Theileria parva infection) in Kenya: past, present and the future. *Parasit Vectors*, **5**, 194.

Galardi S, Fatica A, Bachi A *et al.* (2002) Purified Box C/D snoRNPs Are Able To Reproduce Site-Specific 2'-O-Methylation of Target RNA In Vitro. *Molecular and Cellular Biology*, **22**, 6663–6668.

Galaz V, Gars J, Moberg F, Nykvist B, Repinski C (2015) Why Ecologists Should Care about Financial Markets. *Trends in Ecology & Evolution*, **30**, 571–580.

Game ET, Kareiva P, Possingham HP (2013) Six Common Mistakes in Conservation Priority Setting. *Conservation Biology*, **27**, 480–485.

Gamfeldt L, Hillebrand H, Jonsson PR (2008) Multiple Functions Increase the Importance of Biodiversity for Overall Ecosystem Functioning. *Ecology*, **89**.

García D, Corral N, Canon J (2005) Combining Inter- and Intrapopulation Information with

the Weitzman Approach to Diversity Conservation. *Journal of Heredity*, **96**, 704–712.

Gaston KJ (2000) Global patterns in biodiversity. *Nature*, **405**.

Gautier M, Flori L, Riebler A *et al.* (2009) A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. *BMC Genomics*, **10**, 550.

Gautier M, Laloë D, Moazami-Goudarzi K (2010) Insights into the Genetic History of French Cattle from Dense SNP Data on 47 Worldwide Breeds. *PLoS ONE*, **5**, e13038.

GBIF (2012) Recommended practices for citation of the data published through the GBIF Network. Version 1.0 (Authored by Vishwas Chavan), Copenhagen: Global Biodiversity Information Facility. Pp.12, ISBN: 87-92020-36-4. Accessible at http://links.gbif.org/gbif_best_practice_data_citation_en_v1.

Ghorbani S, Tahmoorespur M, Masoudi-Nejad A, Nasiri MR, Asgari Y (2015) Analysis of the enzyme network involved in cattle milk production using graph theory. *Molecular Biology Research Communications*, **4**, 93–103.

Gilbert I, Robert C, Vigneault C, Blondin P, Sirard M-A (2012) Impact of the LH surge on granulosa cell transcript levels as markers of oocyte developmental competence in cattle. *Reproduction*, **143**, 735–747.

Ginja C, Gama LT, Cortes Ó *et al.* (2013) Analysis of conservation priorities of Iberoamerican cattle based on autosomal microsatellite markers. *Genetics Selection Evolution*, **45**, 1.

Glick G, Shirak A, Seroussi E *et al.* (2011) Fine Mapping of a QTL for Fertility on BTA7 and Its Association With a CNV in the Israeli Holsteins. *G3: Genes, Genomes, Genetics*, **1**, 65–74.

Grieves C (2015) What is Beta Diversity? *methods.blog*.

Groeneveld LF, Lenstra JA, Eding H *et al.* (2010) Genetic diversity in farm animals - a review. *Animal Genetics*, **41**, 6–31.

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

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

Gunther T, Coop G (2013) Robust Identification of Local Adaptation from Allele Frequencies. *Genetics*, **195**, 205–220.

Hanotte O, Bradley DG, Ochieng JW *et al.* (2002) African Pastoralism: Genetic Imprints of Origins and Migrations. *Science*, **296**, 336–339.

Hasan L, Vögeli P, Stoll P, et al. (2004) Intragenic deletion in the gene encoding L-gulonolactone oxidase causes vitamin C deficiency in pigs. *Mammalian Genome*, **15**, 323–333.

Hayashida K, Abe T, Weir W *et al.* (2013) Whole-Genome Sequencing of *Theileria parva*

Strains Provides Insight into Parasite Migration and Diversification in the African Continent. *DNA Research*, **20**, 209–220.

Higham C (1989) *The Archaeology of Mainland Southeast Asia*. Cambridge, Cambridge University Press, 280 pp.

Higham C (2002) *Early cultures of mainland Southeast Asia*. River Books, Bangkok.

Hijmans RJ (2016) *raster: Geographic Data Analysis and Modeling*. R package version 2.5-8. https://CRAN.R-project.org/package=raster.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hoffmann I (2010a) Climate change and the characterization, breeding and conservation of animal genetic resources. *Animal Genetics*, **41**, 32–46.

Hoffmann I (2010b) International flows of animal genetic resources - historical perspective, current status and future expectations.

Holland SJ, Liao XC, Mendenhall MK *et al.* (2001) Functional cloning of Src-like adapter protein-2 (SLAP-2), a novel inhibitor of antigen receptor signaling. *The Journal of experimental medicine*, **194**, 1263–1276.

Huson DH, Bryant D (2005) Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution*, **23**, 254–267.

Huson HJ, Kim E-S, Godfrey RW *et al.* (2014) Genome-wide association study and ancestral origins of the slick-hair coat in tropically adapted cattle. *Frontiers in Genetics*, **5**.

Iannuzzi L, Di Meo GP (2009) *Genome Mapping and Genomics in Domestic Animals*. Water buffalo, p. 19-31 In Cockett, N.E. and C. Kole (eds), Springer Verlag, Berlin, Heidelberg.

Imoto I, Pimkhaokham A, Watanabe T *et al.* (2000) Amplification and overexpression of TGIF2, a novel homeobox gene of the TALE superclass, in ovarian cancer cell lines. *Biochemical and Biophysical Research Communications*, **276**, 264–70.

Jaccard P (1912) The distribution of the flora in the alpine zone. *New phytologist*, **11**, 37–50.

Jarvis A, Reuter HI, Nelson A, Guevara E (2008) Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database.

Johnson JA, Altwegg R, Evans DM *et al.* (2016) Is there a future for genome-editing technologies in conservation? *Animal Conservation*, **19**, 97–101.

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

Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, **27**, 3070–3071.

Jombart T, Devillard S, Dufour AB, Pontier D (2008) Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*, **101**, 92–103.

Jombart T, Pontier D, Dufour AB (2009) Genetic markers in the playground of multivariate analysis. *Heredity*, **102**, 330–341.

Jonsson NN, Piper EK, Constantinoiu CC (2014) Host resistance in cattle to infestation with the cattle tick *Rhipicephalus microplus*. *Parasite Immunology*, **36**, 553–559.

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

Joseph LN, Maloney RF, Possingham HP (2009) Optimal Allocation of Resources among Threatened Species: a Project Prioritization Protocol. *Conservation Biology*, **23**, 328–338.

Justus J, Sarkar S (2002) The principle of complementarity in the design of reserve networks to conserve biodiversity: a preliminary history. *Journal of Biosciences*, **27**, 421–435.

Kabi F, Masembe C, Muwanika V, Kirunda H, Negrini R (2014) Geographic distribution of non-clinical *Theileria parva* infection among indigenous cattle populations in contrasting agro-ecological zones of Uganda: implications for control strategies. *Parasites & Vectors*, **7**.

Kabuusu RM, Alexander R, Kabuusu AM *et al.* (2013) Effect of a Wildlife-Livestock Interface on the Prevalence of Intra-Erythrocytic Hemoparasites in Cattle. *Open Journal of Veterinary Medicine*, **3**, 315–318.

Kaleff B (1942) Der Hausbüffel und seine Züchtungsbiologie im Vergleich zum Rind. Z. Tierzücht. *ZüchtBiol*, **51**.

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

Kazi JU, Kabir NN, Rönnstrand L (2015) Role of SRC-like adaptor protein (SLAP) in immune and malignant cell signaling. *Cellular and Molecular Life Sciences*, **72**, 2535–2544.

Kiage LM, Liu K -b. (2006) Late Quaternary paleoenvironmental changes in East Africa: a review of multiproxy evidence from palynology, lake sediments, and associated records. *Progress in Physical Geography*, **30**, 633–658.

Kierstein G, Vallinoto M, Silva A *et al.* (2004) Analysis of mitochondrial D-loop region casts new light on domestic water buffalo (*Bubalus bubalis*) phylogeny. *Molecular Phylogenetics and Evolution*, **30**, 308–324.

Kijas JW, Lenstra JA, Hayes B *et al.* (2012) Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. *PLoS Biology*, **10**, e1001258.

Kiss-László Z, Henry Y, Bachellerie J-P, Caizergues-Ferrer M, Kiss T (1996) Site-specific

ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell*, **85**, 1077–1088.

Kivaria FM, Heuer C, Jongejan F *et al.* (2004) Endemic stability for *Theileria parva* infections in Ankole calves of the Ankole ranching scheme, Uganda. *Onderstepoort Journal of Veterinary Research*, **71**, p–189.

Kobayashi K, Hernandez LD, Galán JE *et al.* (2002) IRAK-M is a negative regulator of Toll-like receptor signaling. *Cell*, **110**, 191–202.

Kodiha M, Frohlich M, Stochaj U (2012) Spatial Proteomics Sheds Light on the Biology of Nucleolar Chaperones. *Current Proteomics*, **9**, 186–216(31).

Koh LP, Dunn RR, Sodhi NS *et al.* (2004) Species coextinctions and the biodiversity crisis. *science*, **305**, 1632–1634.

Kotschi J (2007) Agricultural biodiversity is essential for adapting to climate change. *GAIA-Ecological Perspectives for Science and Society*, **16**, 98–101.

Kristensen TN, Hoffmann AA, Pertoldi C, Stronen AV (2015) What can livestock breeders learn from conservation genetics and vice versa? *Frontiers in Genetics*, **6**.

Kumar S, Gupta J, Kumar N *et al.* (2006) Genetic variation and relationships among eight Indian riverine buffalo breeds. *Molecular Ecology*, **15**, 593–600.

Kumar CC, Mohan SR, Zavodny PJ, Narula SK, Leibowitz PJ (1989) Characterization and differential expression of human vascular smooth muscle myosin light chain 2 isoform in nonmuscle cells. *Biochemistry*, **28**, 4027–35.

Kumar S, Nagarajan M, Sandhu JS *et al.* (2007a) Mitochondrial DNA analyses of Indian water buffalo support a distinct genetic origin of river and swamp buffalo. *Animal Genetics*, **38**, 227–232.

Kumar S, Nagarajan M, Sandhu JS, Kumar N, Behl V (2007b) Phylogeography and domestication of Indian river buffalo. *BMC Evolutionary Biology*, **7**, 186.

Lachin JM (2004) The role of measurement reliability in clinical trials. *Clinical trials*, **1**, 553–566.

Laloë D, Moazami-Goudarzi K, Lenstra JA *et al.* (2010) Spatial Trends of Genetic Variation of Domestic Ruminants in Europe. *Diversity*, **2**, 932–945.

Lancaster PA, Sharman ED, Horn GW, Krehbiel CR, Starkey JD (2014) Effect of rate of weight gain of steers during the stocker phase. III. Gene expression of adipose tissues and skeletal muscle in growing-finishing beef cattle. *Journal of animal science*, **92**, 1462–1472.

Landin Malt A, Cagliero J, Legent K *et al.* (2012) Alteration of TEAD1 Expression Levels Confers Apoptotic Resistance through the Transcriptional Up-Regulation of Livin. *PLoS ONE*, **7**, e45498.

Larson G, Dobney K, Albarella U *et al.* (2005) Worldwide Phylogeography of Wild Boar

Reveals Multiple Centers of Pig Domestication. *Science*, **307**, 1618–1621.

Larson G, Fuller DQ (2014) The Evolution of Animal Domestication. *Annual Review of Ecology, Evolution, and Systematics*, **45**, 115–136.

Larson G, Piperno DR, Allaby RG *et al.* (2014) Current perspectives and the future of domestication studies. *Proceedings of the National Academy of Sciences*, **111**, 6139–6146.

Lau CH, Drinkwater RD, Yusoff K *et al.* (1998) Genetic diversity of Asian water buffalo (*Bubalus bubalis*): mitochondrial DNA D-loop and cytochrome b sequence variation. *Animal Genetics*, **29**, 253–264.

Lee J-C, Chiang K-C, Feng T-H *et al.* (2016) The Iron Chelator, Dp44mT, Effectively Inhibits Human Oral Squamous Cell Carcinoma Cell Growth in Vitro and in Vivo. *International Journal of Molecular Sciences*, **17**, 1435.

Legendre P, Legendre L (2012) *Numerical Ecology*. Elsevier, Amsterdam, Amsterdam.

Lei CZ, Zhang W, Chen H *et al.* (2007) Independent maternal origin of Chinese swamp buffalo (*Bubalus bubalis*): Origin of Chinese buffalo. *Animal Genetics*, **38**, 97–102.

Leinonen T, McCairns RJS, O'Hara RB, Merilä J (2013) Q_{ST}-F_{ST} comparisons: evolutionary and ecological insights from genomic heterogeneity. *Nature Reviews Genetics*, **14**, 179–190.

Lipsitz SR, Parzen M, Fitzmaurice GM, Klar N (2003) A two-stage logistic regression model for analyzing inter-rater agreement. *Psychometrika*, **68**, 289–298.

Lisha MA, Staab PR, Rose LE, Metzler D (2013) Why to account for finite sites in population genetic studies and how to do this with JAATHA 2.0. *Ecology and Evolution*, **3**, 3647–3662.

Littlejohn MD, Henty KM, Tiplady K *et al.* (2014) Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle. *Nature Communications*, **5**, 5861.

Liu L, Chen X, Jiang L (2004) A study of Neolithic water buffalo remains from Zhejiang, China. *Bulletin of the Indo-Pacific Prehistory Association: The Taipei Papers*, **24**, 113–120.

Lowe WH, Allendorf FW (2010) What can genetics tell us about population connectivity? *Molecular Ecology*, **19**, 3038–3051.

Luikart G, Gielly L, Excoffier L *et al.* (2001) Multiple maternal origins and weak phylogeographic structure in domestic goats. *Proceedings of the National Academy of Sciences*, **98**, 5927–5932.

Mace GM, Norris K, Fitter AH (2012) Biodiversity and ecosystem services: a multilayered relationship. *Trends in Ecology & Evolution*, **27**, 19–26.

Macgregor R (1941) The domestic buffalo. *Vet. Rec.*, **53**, 443–450.

MacHugh DE, Shriver MD, Loftus, RT, Cunningham P, Bradley DG (1997) Microsatellite DNA Variation and the Evolution, Domestication and Phylogeography of Taurine and Zebu Cattle (*Bos taurus* and *Bos indicus*). *Genetics*, **146**, 1071–1086.

Magee DA, MacHugh DE, Edwards CJ (2014) Interrogation of modern and ancient genomes reveals the complex domestic history of cattle. *Animal Frontiers*, **4**, 7–22.

Magona JW, Walubengo J, Olaho-Mukani W *et al.* (2008) Clinical features associated with seroconversion to *Anaplasma marginale*, *Babesia bigemina* and *Theileria parva* infections in African cattle under natural tick challenge. *Veterinary Parasitology*, **155**, 273–280.

Magona JW, Walubengo J, Olaho-Mukani W *et al.* (2011) Spatial variation of tick abundance and seroconversion rates of indigenous cattle to *Anaplasma marginale*, *Babesia bigemina* and *Theileria parva* infections in Uganda. *Experimental and Applied Acarology*, **55**, 203–213.

Magoulas C, Zatsepina OV, Jordan PW, et al. (1998) The SURF-6 protein is a component of the nucleolar matrix and has a high binding capacity for nucleic acids in vitro. *Eur. J. Cell Biol.*, **75**, 174–83.

Manel S, Joost S, Epperson BK *et al.* (2010) Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology*, **19**, 3760–3772.

Manson IL (1974) *Species, types and breeds. In: Cockrill, W. R. (1974). The husbandry and health of the domestic buffalo. Food and Agriculture Organization of the United Nations, 1-47.*

Mapholi NO, Maiwashe A, Matika O *et al.* (2016) Genome-wide association study of tick resistance in South African Nguni cattle. *Ticks and Tick-borne Diseases*, **7**, 487–497.

Marques JRF, Martínez AM, Costa MR *et al.* (2011) Genetic diversity of Brazilian buffaloes (*Bubalus bubalis*) using DNA microsatellites. *Archivos de zootecnia*, **60**, 1213–1221.

Marris E (2007) Conservation priorities: What to let go. *Nature*, **450**, 152–155.

Marton N, Baricza E, Érsek B, Buzás EI, Nagy G (2015) The Emerging and Diverse Roles of Src-Like Adaptor Proteins in Health and Disease. *Mediators of Inflammation*, **2015**, 1–9.

Marufu MC, Qokweni L, Chimonyo M, Dzama K (2011) Relationships between tick counts and coat characteristics in Nguni and Bonsmara cattle reared on semiarid rangelands in South Africa. *Ticks and Tick-borne Diseases*, **2**, 172–177.

Marz M, Kirsten T, Stadler PF (2008) Evolution of Spliceosomal snRNA Genes in Metazoan Animals. *Journal of Molecular Evolution*, **67**, 594–607.

Master LL (1991) Assessing Threats and Setting Priorities for Conservation. *Conservation Biology*, **5**, 559–563.

Matawa F, Murwira A, Schmidt KS (2012) Explaining elephant (*Loxodonta africana*) and buffalo (*Syncerus caffer*) spatial distribution in the Zambezi Valley using maximum

entropy modelling. *Ecological Modelling*, **242**, 189–197.

Matos MC, Utsunomiya YT, Santana do Carmo A *et al.* (2013) *Meta-analysis of genome-wide association studies for gestation length in Nellore cattle - Preliminary results. In: X Simpósio Brasileiro de Melhoramento Animal, Uberaba*.

Mattioli RC, Pandey VS, Murray M, Fitzpatrick JL (2000) Immunogenetic influences on tick resistance in African cattle with particular reference to trypanotolerant N'Dama (*Bos taurus*) and trypanosusceptible Gobra zebu (*Bos indicus*) cattle. *Acta Tropica*, **75**, 263–277.

May RM (1990) Taxonomy as destiny. *Nature*, **347(6289)**, 129–130.

McKeever DJ, Morrison WI (1990) Theileria parva: the nature of the immune response and its significance for immunoprophylaxis. *Scientific and Technical Review of the Office International des Epizooties (Paris)*, **9**, 405–421.

McLeod A, Kristjanson R (1999) *Impact of ticks and associated diseases on cattlein Asia, Australia and Africa. ILRI and eSYS report to ACIAR*. International Livestock Research Institute, Nairobi, Kenya.

McTavish EJ, Decker JE, Schnabel RD, Taylor JF, Hillis DM (2013) New World cattle show ancestry from multiple independent domestication events. *Proceedings of the National Academy of Sciences*, **110**, E1398–E1406.

Merow C, Silander JA (2014) A comparison of MAXLIKE and MAXENT for modelling species distributions. *Methods in Ecology and Evolution*, **5**, 215–225.

Michelizzi VN, Wu X, Dodson MV *et al.* (2011) A global view of 54,001 single nucleotide polymorphisms (SNPs) on the Illumina BovineSNP50 BeadChip and their transferability to water buffalo. *Int J Biol Sci*, **7**, 18–27.

Midgley GF, Hannah L, Millar D, Thuiller W, Booth A (2003) Developing regional and species-level assessments of climate change impacts on biodiversity in the Cape Floristic Region. *Biological Conservation*, **112**, 87–97.

Mishra BP, Dubey PK, Prakash B *et al.* (2015) Genetic analysis of river, swamp and hybrid buffaloes of north-east India throw new light on phylogeography of water buffalo *(Bubalus bubalis)*. *Journal of Animal Breeding and Genetics*, **132**, 454–466.

Mitton J, Linhart Y, Hamrick J, Beckman J (1977) Observations on genetic structure and mating system of ponderosa pine in Colorado front range. *Theoretical and Applied Genetics*, **51**, 5–13.

Moilanen A, Franco AM., Early RI *et al.* (2005) Prioritizing multiple-use landscapes for conservation: methods for large multi-species planning problems. *Proceedings of the Royal Society B: Biological Sciences*, **272**, 1885–1891.

Moilanen A, Leathwick J, Elith J (2008) A method for spatial freshwater conservation prioritization. *Freshwater Biology*, **53**, 577–592.

Moioli B, Georgoudis A, Napolitano F *et al.* (2001) Genetic diversity between Italian and

Greek buffalo populations. *Animal Genetic Resources Information*, **29**, 31–40.

Mokany K, Westcott DA, Prasad S, Ford AJ, Metcalfe DJ (2014) Identifying Priority Areas for Conservation and Management in Diverse Tropical Forests. *PLoS ONE*, **9**, e89084.

Morris W, Doak D (2002) *Quantitative conservation biology: theory and practice of population viability analysis*. Sinauer Associates, Inc. Publishers Sunderland, Massachusetts U.S.A.

Muhanguzi D, Picozzi K, Hatendorf J *et al.* (2014) Prevalence and spatial distribution of *Theileria parva* in cattle under crop-livestock farming systems in Tororo District, Eastern Uganda. *Parasites & vectors*, **7**, 91.

Muscarella R, Galante PJ, Soley-Guardia M *et al.* (2014) ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for MAXENT ecological niche models. *Methods in Ecology and Evolution*, **5**, 1198–1205.

Mwai O, Hanotte O, Kwon Y-J, Cho S (2015) Invited Review - African Indigenous Cattle: Unique Genetic Resources in a Rapidly Changing World. *Asian-Australasian Journal of Animal Sciences*, **28**, 911–921.

Myers N (1988) Threatened biotas: "Hot Spots" in tropical forests. *The Environmentalist*, **8**, 1–20.

Naduvilezhath L, Rose LE, Metzler D (2011) JAATHA: a fast composite-likelihood approach to estimate demographic parameters. *Molecular Ecology*, **20**, 2709–2723.

Nagarajan M, Nimisha K, Kumar S (2015) Mitochondrial DNA Variability of Domestic River Buffalo (*Bubalus bubalis*) Populations: Genetic Evidence for Domestication of River Buffalo in Indian Subcontinent. *Genome Biology and Evolution*, **7**, 1252–1259.

Naidoo R, Du Preez P, Stuart-Hill G, Jago M, Wegmann M (2012) Home on the Range: Factors Explaining Partial Migration of African Buffalo in a Tropical Environment. *PLoS ONE*, **7**, e36527.

Ndungu SG, Brown CGD, Dolan TT (2005) In vivo comparison of susceptibility between *Bos indicus* and *Bos taurus* cattle types to *Theileria parva* infection. *The Onderstepoort journal of veterinary research*, **72**, 13.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, **70**, 3321–3323.

Nene V, Kiara H, Lacasta A *et al.* (2016) The biology of *Theileria parva* and control of East Coast fever - Current status and future trends. *Ticks and Tick-borne Diseases*, **7**, 549–564.

Nicoloso L, Bomba L, Colli L *et al.* (2015) Genetic diversity of Italian goat breeds assessed with a medium-density SNP chip. *Genetics Selection Evolution*, **47**.

Nielsen EE, Hemmer-Hansen J, Poulsen NA *et al.* (2009) Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus*

morhua). *BMC Evolutionary Biology*, **9**, 276.

Nobis MP, Normand S (2014) KISSMig - a simple model for R to account for limited migration in analyses of species distributions. *Ecography*, **37**, 1282–1287.

Norval RAI, Perry BD, Young AS (1992) *The epidemiology of theileriosis in Africa*. ILRI (aka ILCA and ILRAD).

Notter DR (1999) The importance of genetic diversity in livestock populations of the future. *Journal of animal science*, **77**, 61–69.

Ollivier L, Foulley J-L (2005) Aggregate diversity: New approach combining within- and between-breed genetic diversity. *Livestock Production Science*, **95**, 247–254.

Olsen MT, Andersen LW, Dietz R *et al.* (2014) Integrating genetic data and population viability analyses for the identification of harbour seal (*Phoca vitulina*) populations and management units. *Molecular Ecology*, **23**, 815–831.

Olson DM, Dinerstein E (2002) The Global 200: Priority Ecoregions for Global Conservation. *Annals of the Missouri Botanical Garden*, **89**, 199–224.

Olwoch JM, Rautenbach C de W, Erasmus BFN, Engelbrecht FA, Van Jaarsveld AS (2003) Simulating tick distributions over sub-Saharan Africa: the use of observed and simulated climate surfaces. *Journal of Biogeography*, **30**, 1221–1232.

Olwoch JM, Reyers B, Engelbrecht FA, Erasmus BFN (2008) Climate change and the tick-borne disease, Theileriosis (East Coast fever) in sub-Saharan Africa. *Journal of Arid Environments*, **72**, 108–120.

O'Reilly D, von den Driesch A, Voeun V (2006) Archaeology and Archaeozoology of Phum Snay: an Iron Age Cemetery in Northwest Cambodia. *Asian Perspectives*, **45**, 188–211.

Oura CAL, Tait A, Asiimwe B, Lubega GW, Weir W (2011) Haemoparasite prevalence and *Theileria parva* strain diversity in Cape buffalo (*Syncerus caffer*) in Uganda. *Veterinary Parasitology*, **175**, 212–219.

Pariset L, Joost S, Gargani M, Valentini A (2012) *Landscape genomics in livestock*. InTech.

Patel AK, Meadow RH (1998) *The exploitation of wild and domestic water buffalo in prehistoric northwestern south Asia. In: Buitenhuis HL, Bartosiewicz L, Choyke AM, editors. Archaeology of the near east: Proceedings of the third international symposium on the archeozoology of the southwestern Asia and adjacent areas.* 2–5 September 1996, Budapest, Hungary. Groningen (The Netherlands): ARC-Publications. p. 180–199.

Peterson AT, Ortega-Huerta MA, Bartley J *et al.* (2002) Future projections for Mexican faunas under global climate change scenarios. *Nature*, **416**.

Petit RJ, Abdelhamid EM, Odile P (1998) Identifying populations for conservation on the basis of genetic markers. *Conservation Biology*, **12**, 844–855.

Pettorelli N, Ryan S, Mueller T *et al.* (2011) The Normalized Difference Vegetation Index (NDVI): unforeseen successes in animal ecology. *Climate Research*, **46**, 15–27.

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

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

Poncet BN, Herrmann D, Gugerli F *et al.* (2010) Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. *Molecular Ecology*, **19**, 2896–2907.

Primack RB, Ralls K (1995) *A primer of conservation biology*. ^ eSunderland Sunderland: Sinauer Associates.

Pritchard JK, Di Rienzo A (2010) Adaptation - not by sweeps alone. *Nature Reviews Genetics*, **11**, 665–667.

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

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

QGIS Development Team (2016) *QGIS Geographic Information System*. Open Source Geospatial Foundation Project.

Quintela M, Johansson MP, Kristjánsson BK, Barreiro R, Laurila A (2014) AFLPs and Mitochondrial Haplotypes Reveal Local Adaptation to Extreme Thermal Environments in a Freshwater Gastropod. *PLoS ONE*, **9**, e101821.

R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Razgour O, Taggart J, Manel S *et al.* An integrated framework to identify wildlife populations under threat from climate change. Submitted.

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

Reist-Marti SB, Abdulai A, Simianer H (2006) Optimum allocation of conservation funds and choice of conservation programs for a set of African cattle breeds. *Genetics Selection Evolution*, **38**, 99–126.

Reist-Marti SB, Simianer H, Gibson J, Hanotte O, Rege JEO (2003) Weitzman's approach and conservation of breed diversity: an application to African cattle breeds. *Conservation Biology*, **17**, 1299–1311.

Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R (2015) A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, **24**,

4348–4370.

Reyes FF (1948) The growth and development of Philippine and Murrah carabaos and their crossbreed. *Philipp. Agric*, **31**, 271–278.

Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, **105**, 767–779.

Ritz LR, Glowatzki-Mullis M-L, MacHugh DE, Gaillard C (2000) Phylogenetic analysis of the tribe Bovini using microsatellites. *Animal Genetics*, **31**, 178–185.

Robinson TP, Wint GRW, Conchedda G *et al.* (2014) Mapping the global distribution of livestock. *PLoS ONE*, **9**, e96084.

Rockman MV (2012) The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution*, **66**, 1–17.

Royle JA, Chandler RB, Yackulic C, Nichols JD (2012) Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, **3**, 545–554.

Rubaire-Akiiki CM, Okello-Onen J, Musunga D *et al.* (2006) Effect of agro-ecological zone and grazing system on incidence of East Coast Fever in calves in Mbale and Sironko Districts of Eastern Uganda. *Preventive Veterinary Medicine*, **75**, 251–266.

Sæther SA, Fiske P, Kålås JA *et al.* (2007) Inferring local adaptation from Q_{ST}-F_{ST} comparisons: neutral genetic and quantitative trait variation in European populations of great snipe. *Journal of Evolutionary Biology*, **20**, 1563–1576.

Saif R, Babar ME, Awan AR *et al.* (2012) DNA fingerprinting of Pakistani buffalo breeds (Nili-Ravi, Kundi) using microsatellite and cytochrome b gene markers. *Molecular Biology Reports*, **39**, 851–856.

Samarsky DA, Fournier MJ (1999) A comprehensive database for the small nucleolar RNAs from *Saccharomyces cerevisiae*. *Nucleic acids research*, **27**, 161–164.

Sarkar S (2014) Conservation Biology. *The Stanford Encyclopedia of Philosophy*.

Savage DB, Zhai L, Ravikumar B *et al.* (2008) A Prevalent Variant in PPP1R3A Impairs Glycogen Synthesis and Reduces Muscle Glycogen Content in Humans and Mice (LC Groop, Ed,). *PLoS Medicine*, **5**, e27.

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics*, **14**, 807–820.

Scherf BD (Ed. . (2000) *World watch list for domestic animal diversity*.

Schoville SD, Bonin A, François O *et al.* (2012) Adaptive Genetic Variation on the Landscape: Methods and Cases. *Annual Review of Ecology, Evolution, and Systematics*, **43**, 23–43.

Sebastián-González E, Sánchez-Zapata JA, Botella F *et al.* (2011) Linking cost efficiency evaluation with population viability analysis to prioritize wetland bird conservation

actions. *Biological conservation*, **144**, 2354–2361.

Secretariat of the Convention on Biological Diversity. Handbook of the Convention on Biological Diversity Including its Cartagena Protocol on Biosafety (2005)

Sherwood ER, Toliver-Kinsky T (2004) Mechanisms of the inflammatory response. *Best Practice & Research Clinical Anaesthesiology*, **18**, 385–405.

Sidky AR (1951) The buffalo in Egypt. I. General study and improvement work. Cairo, Ministry of Agriculture: 19.

Simianer H (2005) Decision making in livestock conservation. *Ecological Economics*, **53**, 559–572.

Simianer H, Marti S., Gibson J, Hanotte O, Rege JE. (2003) An approach to the optimal allocation of conservation funds to minimize loss of genetic diversity between livestock breeds. *Ecological Economics*, **45**, 377–392.

Simpson G (1943) Mammals and the nature of continents. *American Journal of Science*, **241**, 1–31.

Singh JS (2002) The biodiversity crisis: a multifaceted review. *Current Science*, **82**, 638–647.

Singh J, Gill JS, Kwatra MS, Sharma KK (1993) Treatment of theileriosis in crossbred cattle in the Punjab. *Tropical Animal Health and Production*, **25**, 75–78.

Sivakumar T, Hayashida K, Sugimoto C, Yokoyama N (2014) Evolution and genetic diversity of *Theileria*. *Infection, Genetics and Evolution*, **27**, 250–263.

Sjö A, Magnusson K, Peterson KJ (2005) Association of α-Dystrobrevin with Reorganizing Tight Junctions. *The Journal of Membrane Biology*, **203**, 21–30.

Skrbinšek T, Jelenčič M, Waits LP *et al.* (2012) Using a reference population yardstick to calibrate and compare genetic diversity reported in different studies: an example from the brown bear. *Heredity*, **109**, 299–305.

Smith MW, Clark SP, Hutchinson JS *et al.* (1993) A sequence-tagged site map of human chromosome 11. *Genomics*, **17**, 699–725.

Smith SL, Everts RE, Sung L-Y *et al.* (2009) Gene Expression Profiling of Single Bovine Embryos UncoversSignificant Effects of In Vitro Maturation, Fertilization and Culture. *Molecular Reproduction & Development*, **76**, 38–47.

Song S, Ou-Yang Y, Huo J *et al.* (2016) Molecular cloning, sequence characterization, and tissue expression analysis of three water buffalo (*Bubalus bubalis*) genes - ST6GAL1, ST8SIA4 and SLC35C1. *Archives Animal Breeding*, **59**, 363–372.

Sorbolini S, Marras G, Gaspa G *et al.* (2015) Detection of selection signatures in Piemontese and Marchigiana cattle, two breeds with similar production aptitudes but different selection histories. *Genetics Selection Evolution*, **47**.

Sørensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation

on Danish commons. *Kongelige Danske Videnskabernes Selskabs Biologiske skrifter*, **5**, 1–34.

Sork VL, Aitken SN, Dyer RJ *et al.* (2013) Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. *Tree Genetics & Genomes*, **9**, 901–911.

Soudré A, Ouédraogo-Koné S, Wurzinger M *et al.* (2013) Trypanosomosis: a priority disease in tsetse-challenged areas of Burkina Faso. *Tropical Animal Health and Production*, **45**, 497–503.

Spector S (2002) Biogeographic crossroads as priority areas for biodiversity conservation. *Conservation Biology*, **16**, 1480–1487.

Spitze K (1993) Population Structure in *Daphnia obtusa*: Quantitative Genetic and Allozymic Variation. *Genetics*, **135**, 367–374.

Stockwell CA, Hendry AP, Kinnison MT (2003) Contemporary evolution meets conservation biology. *Trends in Ecology & Evolution*, **18**, 94–101.

Stubben C, Milligan B, others (2007) Estimating and analyzing demographic models using the popbio package in R. *Journal of Statistical Software*, **22**, 1–23.

Stucki S, Orozco-terWengel P, Forester BR *et al.* (2016) High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources*.

Surks HK (2007) cGMP-Dependent Protein Kinase I and Smooth Muscle Relaxation: A Tale of Two Isoforms. *Circulation Research*, **101**, 1078–1080.

Swets DL, Reed BC, Rowland JD, Marko SE (1999) *A weighted least-squares approach to temporal NDVI smoothing. In: Bethesda, Maryland, American Society for Photogrammetry and Remote Sensing*. Portland, Oregon.

Taberlet P, Valentini A, Rezaei HR *et al.* (2008) Are cattle, sheep, and goats endangered species? *Molecular Ecology*, **17**, 275–284.

Takeuchi M, Hata Y, Hirao K *et al.* (1997) A family of PSD-95/SAP90-associated proteins localized at postsynaptic density. *The Journal of Biological Chemistry*, **272**, 11943–51.

Tang H, Choudhry S, Mei R *et al.* (2007) Recent Genetic Selection in the Ancestral Admixture of Puerto Ricans. *The American Journal of Human Genetics*, **81**, 626–633.

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

Tapio I, VäRv S, Bennewitz J *et al.* (2006) Prioritization for Conservation of Northern European Cattle Breeds Based on Analysis of Microsatellite Data. *Conservation Biology*, **20**, 1768–1779.

Taylor HR, Gemmell NJ (2016) Emerging Technologies to Conserve Biodiversity: Further Opportunities via Genomics. Response to Pimm et al. *Trends in Ecology & Evolution*,

31, 171–172.

Tenaillon MI, Tiffin PL (2008) The quest for adaptive evolution: a theoretical challenge in a maze of data. *Current Opinion in Plant Biology*, **11**, 110–115.

Tetens J, Heuer C, Heyer I *et al.* (2015) Polymorphisms within the APOBR gene are highly associated with milk levels of prognostic ketosis biomarkers in dairy cows. *Physiological Genomics*, **47**, 129–137.

The Bovine HapMap Consortium, Gibbs RA, Taylor JF *et al.* (2009) Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science*, **324**, 528–532.

Thomas CD, Franco AM, Hill JK (2006) Range retractions and extinction in the face of climate warming. *Trends in Ecology & Evolution*, **21**, 415–416.

Thompson LG, Mosley-Thompson E, Davis ME *et al.* (2002) Kilimanjaro Ice Core Records: Evidence of Holocene Climate Change in Tropical Africa. *Science*, **298**, 589–593.

Thuiller W, Georges D, Engler R, Breiner F (2016) *biomod2: Ensemble Platform for Species Distribution Modeling*. R package version 3.3-7. https://CRAN.R-project.org/package=biomod2.

Thuiller W, Lafourcade B, Engler R, Araújo MB (2009) BIOMOD - a platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369–373.

Tian Y, Wu J, Smith AT *et al.* (2011) Population viability of the Siberian Tiger in a changing landscape: Going, going and gone? *Ecological Modelling*, **222**, 3166–3180.

Toro MA, Caballero A (2005) Characterization and conservation of genetic diversity in subdivided populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 1367–1378.

Traill L, Bradshaw C, Brook B (2007) Minimum viable population size: A meta-analysis of 30 years of published estimates. *Biological Conservation*, **139**, 159–166.

Troy CS, MacHugh DE, Bailey JF *et al.* (2001) Genetic evidence for Near-Eastern origins of European cattle. *Nature*, **410**, 1088–1091.

Turner S, Armstrong LL, Bradford Y *et al.* (2011) *Quality Control Procedures for Genome-Wide Association Studies. In: Current Protocols in Human Genetics* (eds Haines JL, Korf BR, Morton CC, et al.). John Wiley & Sons, Inc., Hoboken, NJ, USA.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV (2010) Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nature Genetics*, **42**, 260–263.

Ünal EÖ, Soysal Mİ, Yüncü E, Dağtaş ND, Togan İ (2014) Microsatellite based genetic diversity among the three water buffalo (*Bubalus bubalis*) populations in Turkey. *Archiv fürTierzucht-Archives of animal breeding*, **57**.

Urbina-Cardona JN, Flores-Villela O (2010) Ecological-Niche Modeling and Prioritization of

Conservation-Area Networks for Mexican Herpetofauna. *Conservation Biology*, **24**, 1031–1041.

Vahidi SMF, Tarang AR, Anbaran MF *et al.* (2014) Investigation of the genetic diversity of domestic *Capra hircus* breeds reared within an early goat domestication area in Iran. *Genetics Selection Evolution*, **46**, 27.

Vane-Wright RI, Humphries CJ, Williams PH (1991) What to protect?-Systematics and the agony of choice. *Biological Conservation*, **55**, 235–254.

Vijh RK, Tantia MS, Mishra B, Bharani Kumar ST (2008) Genetic relationship and diversity analysis of Indian water buffalo (*Bubalus bubalis*). *Journal of Animal Science*, **86**, 1495–1502.

de Villemereuil P, Gaggiotti OE (2015) A new F_{ST}-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*, **6**, 1248–1258.

Vitousek PM, Mooney HA, Lubchenco J, Melillo JM (1997) Human Domination of Earth's Ecosystems. *Science*, **277**, 494–499.

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, **47**, 97–120.

Volkmann L, Martyn I, Moulton V, Spillner A, Mooers AO (2014) Prioritizing Populations for Conservation Using Phylogenetic Networks. *PLoS ONE*, **9**, e88945.

Vudriko P, Okwee-Acai J, Tayebwa DS *et al.* (2016) Emergence of multi-acaricide resistant *Rhipicephalus* ticks and its implication on chemical tick control in Uganda. *Parasites & Vectors*, **9**.

Wang J (2005) Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 1395–1409.

Webber BL, Raghu S, Edwards OR (2015) Opinion: Is CRISPR-based gene drive a biocontrol silver bullet or global conservation threat? *Proceedings of the National Academy of Sciences*, **112**, 10565–10567.

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–70.

Weitzman ML (1992) On diversity. *The Quarterly Journal of Economics*, **107**, 363–405.

Weitzman ML (1993) What to preserve? An application of diversity theory to crane conservation. *Quarterly Journal of Economics*, **108**, 157–183.

Weitzman ML (1998) The Noah's ark problem. *Econometrica*, **66**, 1279–1298.

Wellison Jarles da Silva D (2015) Expressão gênica diferencial relacionada ao contenúdo de ferro no músculo em animais Nelore. Universidade Federal De São Carlos, Centro de Ciências Biológicas e da Saude, Programa de pós-graduação em genética evolutiva e biologia molecular, São Paulo.

Wellmann R, Bennewitz J, Meuwissen THE (2014) A unified approach to characterize and conserve adaptive and neutral genetic diversity in subdivided populations. *Genetics Research*, **96**.

Whittaker RH (1960) Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, **30**, 279–338.

Whittaker RH (1972) Evolution and measurement of species diversity. *Taxon*, **21**, 213–251.

Wikel SK, Bergrnan D (1997) Tick-Host Immunology: Significant Advances and Challenging Opportunities. *Parasitology Today*, **13**, 383–389.

Willadsen P (1980) *Immediate hypersensitivity to* Boophilus microplus*: factors affecting hypersensitivity, and their relevance in the resistance of cattle to ticks. In: Proceedings of a Symposium held at the 56th Annual Conference of Australian Veterinary Association*. Johnston, L.A.Y., Cooper, M.G. (Eds.), Townsville, 14–18 May 1976, Sydney, Australia, 60-62.

Wilson KA, McBride MF, Bode M, Possingham HP (2006) Prioritizing global conservation efforts. *Nature*, **440**, 337–340.

Winnie JA, Cross P, Getz W (2008) Habitat quality and heterogeneity influence distribution and behavior in African buffalo (*Syncerus caffer*). *Ecology*, **89**, 1457–1468.

Wright S (1965) The interpretation of population structure by *F*-statistics with special regard to systems of mating. *Evolution*, **19**.

Xu L, Hou Y, Bickhart DM *et al.* (2016) Population-genetic properties of differentiated copy number variations in cattle. *Scientific Reports*, **6**, 23161.

Yang DY, Liu L, Chen X, Speller CF (2008) Wild or domesticated: DNA analysis of ancient water buffalo remains from north China. *Journal of Archaeological Science*, **35**, 2778–2785.

Yeaman S, Otto SP (2011) Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift. *Evolution*, **65**, 2123–2129.

Yindee M, Vlamings BH, Wajjwalku W *et al.* (2010) Y-chromosomal variation confirms independent domestications of swamp and river buffalo. *Animal Genetics*.

Young AS, Leitch BL (1981) Epidemiology of East Coast fever: some effects of temperature on the development of *Theileria parva* in the tick vector *Rhipicephalus appendiculatus*. *Parasitology*, **83**, 199–211.

Yue X-P, Li R, Xie W-M *et al.* (2013) Phylogeography and Domestication of Chinese Swamp Buffalo. *PLoS ONE*, **8**, e56552.

Zeuner FE (1963) *A History of Domesticated Animals*. London: Hutchinson. New York: Harper & Row.

Zha J, Zhou Q, Xu L-G *et al.* (2004) RIP5 is a RIP-homologous inducer of cell death. *Biochemical and Biophysical Research Communications*, **319**, 298–303.

Zhang Y, Lu Y, Yindee M *et al.* (2016) Strong and stable geographic differentiation of swamp buffalo maternal and paternal lineages indicates domestication in the China/Indochina border region. *Molecular Ecology*, **25**, 1530–1550.

Zhang Y, Sun D, Yu Y, Zhang Y (2007) Genetic diversity and differentiation of Chinese domestic buffalo based on 30 microsatellite markers. *Animal Genetics*, **38**, 569–575.

Zhang Y, Vankan D, Zhang Y, Barker JSF (2011) Genetic differentiation of water buffalo (*Bubalus bubalis*) populations in China, Nepal and south-east Asia: inferences on the region of domestication of the swamp buffalo. *Animal Genetics*, **42**, 366–377.

7. Final report

7.1 First year

7.1.1 Mandatory courses attended and exams completed (January–September 2014)

1. *Sustainable Animal Production*—Instructor: Prof. Paolo Ajmone Marsan (8 h);
2. *Sustainable Crop Production*—Instructor: Prof. Stefano Amaducci (8 h);
3. *Statistics and Data Management*—Instructor: Prof. Enrico Fabrizi (25 h);
4. *Human nutrition*—Instructor: Prof. Giancarlo Carrara (10 h);
5. *Diritto Internazionale ed Europeo del commercio dei prodotti agroalimentari*—Instructor: Prof. Francesco Bestagno (10 h);
6. *Diritto europeo multi-livello e disciplina agroalimentare: le fonti della materia fra ordinamento statale e integrazione giuridica continentale*—Instructor: Prof. Dino Rinoldi (10 h);
7. *Agricultural and food policies of the European Union*—Instructor: Prof. Paolo Sckokai (10 h);
8. *Food Technologies and Sustainability*—Instructor: Prof.ssa Giorgia Spigno (10 h);
9. *Basic Management and Management of Knowledge*—Instructor: Prof. Emanuele Vendramini (15 h);
10. *English course*—Instructor: Prof Nicoletta Gueli (60 h).

7.1.2 Mandatory seminars attended

1. *A tavola con le religioni. Le regole religiose alimentari ed il loro impatto nella vita quotidiana*. Lecturer: Massimo Salani.
2. *Sportiva-mente Movimento, alimentazione e sostenibilità per vivere meglio!* Lecturer: Dr. Francesco Confalonieri.
3. *RI-CIBIAMO chi ama il cibo non lo spreca! Un tuffo nella blue economy!* Lecturer: Prof. Paolo Rizzi.
4. *Innovative Tools for Sustainable management of Vineyards in IPM.* Lecturer: Dr.

Tito Caffi.

5. *Growing Grapes in a Climate Change Scenario: the New Challenge*. Lecturer: Prof. Stefano Poni
6. *Creation and development of innovative food products: EcoTrophelia and the case study of SOcrock*. Lecturer: Dr. Roberta Dordoni.

7.1.3 Research activity

7.1.3.1 Study of local adaptation to East Coast Fever in indigenous cattle population from Uganda

I started my PhD research activity in the context of the European Project NEXTGEN ("Next generation methods to preserve farm animal biodiversity by optimizing present and future breeding options" – EU FP7-KBBE-2009-01-01-03, http://nextgen.epfl.ch/). The project targeted indigenous livestock populations of Iranian and Moroccan sheep and goats, as well as Ugandan cattle. In each case, a vast set of animals was characterized by means of next-generation methods, following specific sampling schemes aimed at covering all the environmental conditions experienced by the concerned populations. An unprecedented amount of genetic data was produced to characterize (i) genetic diversity, (ii) local adaptation to climate and diseases, and (iii) conserve biodiversity of the studied populations.

Within this context, I began my activity focusing on the study of local adaptation shown by indigenous cattle population from Uganda towards endemic diseases, in particular East Coast Fever. Much of the work during the first PhD year was devoted to the study of the literature on (i) the epidemiology of East Coast Fever, and on (ii) the spatial analysis methods (e.g. species distribution modelling) which would have been necessary to characterize the geographical occurrence of the disease vector and the parasite over the study area, i.e. Uganda.

7.1.3.2 The study of *Bubalus bubalis* diversity

Staring from September 2014, I started to be involved also in the study of *Bubalus bublis* diversity and evolutionary history. The investigation was based on the new, species-specific, 90K Affymetrix Axiom® Buffalo Genotyping Array, as developed by the International Buffalo

Consortium. The Institute of Zootechnics of the Università Cattolica del S. Cuore participated as a partner to the consortium, being in charge of describing worldwide patterns of domestic buffalo genetic diversity. Genetic data from 'river' and 'swamp' buffaloes populations from several countries, spanning from the probable domestication centres (Pakistan and South-eastern Asia) to China, Middle East, Africa, Europe and South America were collected. Much of the work during the first PhD year was devoted to metadata collection (e.g. historical information concerning the studied populations, coordinates of the sampled populations) and the genomic dataset construction.

7.2 Second year

7.2.1 Mandatory courses attended and exams completed

Research ethics—Instructor: Prof Mariachiara Tallacchini.

7.2.2 Freely chosen courses

Spatial analysis of ecological data using R for ecologists and epidemiologists. Instructor: Prof. Jason Matthiopoulos, The Scottish Centre for Ecology and the Natural Environment, University of Glasgow, Loch Lomond, Glasgow, Scotland. Pr~statistics: Delivering ecology based courses and workshops (Including a total of 40 contact hours redeemable as 2 Open University points in the United Kingdom which are transferable as 2 ECTS in Europe).

7.2.3 Research activity

7.2.3.1 The study of local adaptation to East Coast Fever in indigenous cattle population from Uganda

From April to July 2015, I continued my research activity on local adaption to East Coast

Fever under the supervision of Prof. Stéphane Joost at the Laboratory of Geographic Information System (LASIG), École Polytechnique Féderale de Lausanne (Switzerland), where I had the opportunity to deepen my knowledge on spatial modelling and landscape genomics. My stay at LASIG produced three contributions to international congresses. Here, I report the title, co-authors and abstract for each contribution:

1. Poster presentation at the *XXI Congress of the Animal Science and Production Association (ASPA)*. The University of Milan, June 9─12, 2015.

Modelling the spatial distribution of *Theileria parva* (Theiler 1904), causative agent of East Coast Fever disease in cattle

Elia Vajana[1], Licia Colli[1], Marco Milanesi[1], Lorenzo Bomba[1], Riccardo Negrini[1, 2], Stefano Capomaccio[1], Elisa Eufemi[1], Raffaele Mazza[2], Alessandra Stella[3], Stephane Joost[4], Sylvie Stucki[4], Pierre Taberlet[5], François Pompanon[5], Fred Kabi[6], Vincent Muwanika[6], Charles Masembe[6], Paolo Ajmone-Marsan[1], The NEXTGEN Consortium[7]

(1) Istituto di Zootecnica e Centro di Ricerca BioDNA, Facoltà di Scienze Agrarie, Alimentari ed Ambientali, Università Cattolica del S. Cuore, Piacenza, Italy (2) AIA─Associazione Italiana Allevatori, Rome. Italy (3) IBBA-CNR and FPTP - Fondazione Parco Tecnologico Padano. Lodi. Italy (4) Laboratory of Geographic Information Systems (LASIG), School of Architecture, Civil and environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland (5) LECA─Lab. d'Ecologie Alpine, UJF-CNRS, Grenoble, France (6) Institute of Environment & Natural Resources, Makerere University, Kampala, Uganda (7) EU funded project, http://nextgen.epfl.ch.

Theileria parva is a protozoan emo-parasite, which affects *Bos taurus* and *Bos indicus* cattle populations causing East Coast Fever disease, one of the most relevant cattle plagues in sub-Saharan Africa causing the death of $\sim 1.1 \cdot 10^6$ animals per year and an annual loss of $\sim 168 \cdot 10^6$ USD, *T. parva* occurrence is bound to three conditions: i) the presence of susceptible bovine host populations; ii) the presence of its main tick vector *Rhipicephalus appendiculatus*; iii) suitable ecological conditions for the survival of both the vector and the parasite in all their developmental stages. While the environmental drivers affecting the vector occurrence have been extensively investigated, studies focusing solely on the conditions determining the presence of the parasite are still lacking. The present study aims therefore at investigating the ecological conditions needed to maintain the parasite-vector-host biological system. In the course of the EU-funded project Nextgen, 590 cattle blood samples from 204 georeferenced locations covering the whole Ugandan country have been tested for the presence/absence of *T. parva* DNA. The values of 19 bioclimatic variables and topographic data (altitude, aspect and slope) for each sampling site were derived from WorldClim (Global Climate Data) and Shuttle Radar Topography Mission (SRTM) databases. A classification tree model approach was used to test bioclimatic and topographic variables together with geographical coordinates. This analysis revealed latitude as the main geographical driver for *T. parva* occurrence in Uganda, with potential interactions among temperature seasonality, temperature annual range and precipitations of the wettest month in the southern regions (latitude≤ -0.15). For central-northern regions, instead, mean diurnal range, territory aspect and slope were the variables influencing most the presence of the parasite. This preliminary work represents a first step for the development of a full probabilistic model for *T. parva* occurrence in sub-Saharan Africa.

2. Oral and poster presentation at the *XIX Evolutionary Biology Meeting*, Marseilles, September 15─18, 2015.

Effect of climate change on the spatial distribution of genomic variants involved in the resistance to East Coast Fever in Ugandan cattle

Estelle Rochat[1*], Elia Vajana[2*], Licia Colli[2], Charles Masembe[3], Riccardo Negrini2, Paolo Ajmone-Marsan[2], Stéphane Joost[1] and the NEXTGEN Consortium

(1) Laboratory of Geographic Information Systems (LASIG), School of Architecture, Civil and environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland (2) Institute of Zootechnics and BioDNA Research Centre, Faculty of Agricultural, Food and Environmental Sciences, Università Cattolica del S, Cuore, Piacenza, Italy (3) Institute of Environment & Natural Resources, Makerere University, Kampala, Uganda

These authors contributed equally to this work

East Coast Fever (ECF) is a major livestock disease caused by *Theileria parva* Theiler, 1904, an emo-parasite protozoan transmitted by the tick *Rhipicephalus appendiculatus* Neumann, 1901. This disease provokes high mortality in cattle populations of East and Central Africa, especially in exotic breeds and crossbreds (Olwoch et al., 2008). Here, we use landscape genomics (Joost *et al.* 2007) to highlight genomic regions likely involved into tolerance/resistance mechanisms against ECF, and we introduce *SPatial Area of Genotype probability* (*SPAG*) to delimit territories where favourable genotypes are predicted to be present.

Between 2010 and 2012, the NEXTGEN project (nextgen.epfl.ch) carried out the geo-referencing and genotyping (54K SNPs) of 803 Ugandan cattle, among which 496 were tested for *T. parva* presence. Moreover, 532 additional *R. appendiculatus* occurrences were obtained from a published database (Cumming. 1998). Current and future values of 19 bioclimatic variables were also retrieved from the WorldClim database (www.worldclim.org/).

In order to evaluate the selective pressure of the parasite, we used MAXENT (Phillips *et al.* 2006); (Muscarella *et al.* 2014) and a mixed logistic regression (Bates *et al.* 2015) to model and map the ecological niches of both *T. parva* and *R.* appendiculatus. Then, we used a correlative approach (Stucki et al., 2014) to detect genotypes positively associated with the resulting probabilities of presence and built the corresponding *SPAG*. Finally, we considered bioclimatic predictors representing two different climate change scenarios for 2070—one moderate and one severe—to forecast the simultaneous shift of both *SPAG* and vector/pathogen niches.

While suitable ecological conditions for *T. parva* are predicted to remain constant, the best environment for the vector is predicted around Lake Victoria. However, when considering future conditions, parasite occurrence is expected to decrease because of the contraction of suitable environments for the tick in both scenarios.

Landscape genomics' analyses revealed several markers significantly associated with a high probability of presence of the tick and of the parasite. Among them, we found the marker ARS-BFGL-NGS-113888, whose heterozygous genotype AG showed a positive association. Interestingly, this marker is located close to the gene IRAK-M, an essential component of the Toll-like receptors involved in the immune response against pathogens (Kobayashi *et al.* 2002). If the implication of this gene into resistance mechanisms against ECF is confirmed, the corresponding *SPAG* (Figure 7.1) represents either areas where the variant of interest shows a high probability to exist now, or areas where ecological characteristics are the most favorable to induce its presence under future climatic conditions.

Beyond the results presented here, the combined use of *SPAG* and niche maps could help identifying critical geographical regions that do not present the favourable genetic variant in the present, but where a parasite is likely to expand its range in the future. This may represent a valuable tool to support the identification of current resistant populations and to direct future targeted crossbreeding schemes.

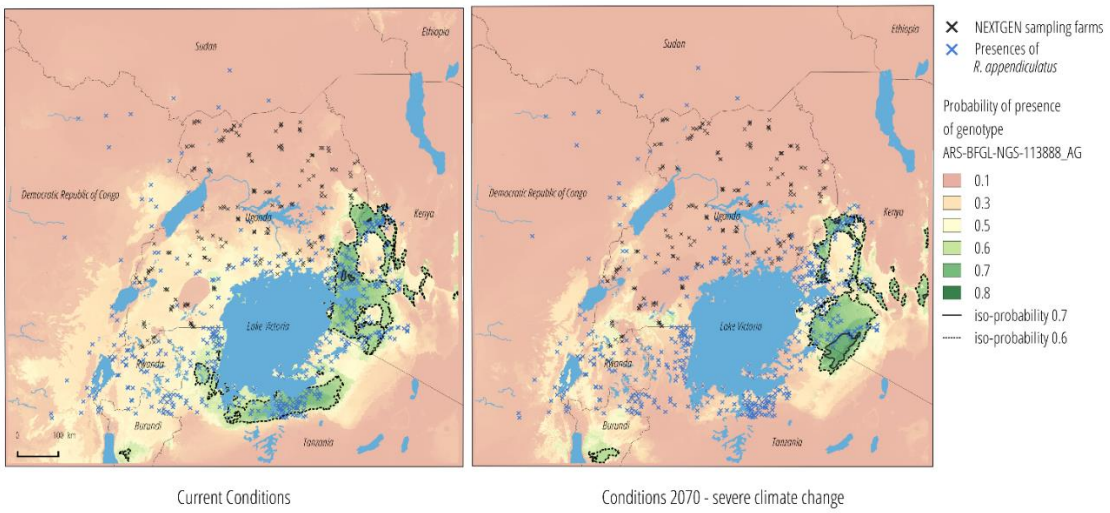

Figure 7.1 *SPatial Area of Genotype probability* (*SPAG*) for the genotype AG of the SNP "ARS-BFGL-NGS-113888" (ARS-11), highlighting areas where this genotype shows a high probability to be present (Current Conditions), and where it may be distributed in the future (Conditions 2070). As the presence of ARS-11_AG is positively correlated with the presence of the tick *R. appendiculatus* (α= 0.01; Efron pseudo R^2 = 0.074), we can estimate the probability of presence of this genotype also in regions without sampling points and thus without genetic data. At present, the areas of high probability of presence of ARS-11_AG are mainly observed in the North-East and the South of Lake Victoria. However, when considering environmental conditions in 2070 (assuming severe climate change), these areas are expected to be mainly restricted to the North-East of Lake Victoria, where favorable conditions for the presence of *R. appendiculatus* are supposed to be maintained.

3. Poster presentation at the *XXIV International Plant & Animal Genome*, San Diego, California, USA, January 9─13, 2016.

Spatial areas of genotype probability of cattle genomic variants involved in the resistance to East Coast Fever: a tool to predict future disease-vulnerable geographical regions

Elia Vajana[1], Estelle Rochat[2], Licia Colli[1], Charles Masembe[3], Riccardo Negrini[1], Paolo Ajmone-Marsan[1], Stéphane Joost[2] and the NEXTGEN Consortium

(1) Institute of Zootechnics and BioDNA Research Centre, Faculty of Agricultural, Food and Environmental Sciences, Università Cattolica del S, Cuore, Piacenza, Italy (2) Laboratory of Geographic Information Systems (LASIG), School of Architecture, Civil and environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland (3) Institute of Environment & Natural Resources, Makerere University, Kampala, Uganda

These authors contributed equally to this work

East Coast Fever (ECF) is a livestock disease caused by *Theileria parva*, a protozoan transmitted by the vector tick *Rhipicephalus appendiculatus*. This disease causes high mortality in cattle populations of Central and Eastern Africa, especially in exotic breeds. Here, we highlight genomic regions likely involved into tolerance/resistance mechanisms against ECF, and we introduce the estimation of their Spatial Area of Genotype Probability (SPAG) to delimit areas where the concerned genotypes are predicted to be present.

During the NEXTGEN project, 803 Ugandan cattle were geo-referenced and genotyped (54K SNPs), while 532 tick occurrences were retrieved from a published database. To get a proxy of the parasite selective pressure, we used WorldClim bioclimatic variables to model vector ecological niche. Landscape genomics

models were then used to detect cattle genotypes associated with vector probability of presence, and to estimate their SPAGs. Finally, climate change scenarios for 2070 were considered to compare the predicted shift in the vector niche with the estimated current SPAG.
The analysis revealed two main areas of presence of possibly resistance-related genotypes, one South and one East of Lake Victoria. Climate change will probably shift tick niche southwards in the Eastern regions of Lake Victoria, inducing a critical area that currently does not show the candidate genotypes, but where disease will likely spread in the future.
The combined use of SPAGs and niche maps could therefore facilitate the identification of regions of concern and to direct future targeted breeding schemes.

7.2.3.2 The study of *Bubalus bubalis* diversity

I collaborated in performing several of the analyses reported in **Chapter 3**, particularly those concerning population structure, admixture and migration events.

7.2.3.1 Review on prioritization methods in conservation biology

From October 2015 to February 2016, I have been hosted by Prof. Michael W. Bruford's Laboratory, at Cardiff School of Biosciences, Division Organisms and Environment, Cardiff University. Originally, the objective of my stay was to develop a new adaptive index for prioritizing populations for conservation. However, my research target changed given the complexity of the topic and the vast amount of literature dedicated to this issue. Under the supervision of Prof. Michael W. Bruford and Dr. Pablo Orozco-terWengel, I started reviewing the literature on the available prioritization methods in conservation biology, with the aim of proposing an original conceptual framework/decision tool to help decision-makers in conservation biology in selecting the most appropriate methodologies given case-specific requirements. The new framework aimed at being valid for both livestock and wildlife conservation, unraveling methodological gaps in current literature, and envisaging possible new prioritization methods based on genomic data.

7.3 Third year

7.3.1 Freely chosen courses

Introduction to Bayesian statistics with R (Introduzione alla statistica Bayesiana con R). Instructor: Prof. Stefano Leonardi, Dipartimento di Scienze Chimiche, della Vita e della Sostenibilità Ambientale, Università di Parma, Parma, Italy, July 6─8, 2016. *c.* 24 hours.

7.3.2 Congresses attended

Congenomics 2016—Conference on conservation genomics, May 3─6, 2016, CIBIO-InBIO, Campus Agrário de Vairão, University of Porto, Portugal.

7.3.3 Research activity

7.3.3.1 The study of local adaptation to East Coast Fever in indigenous cattle population from Uganda

I finalized the study on local adaptation to East Coast Fever in Uganda. **Chapter 4** represents the result of my work: I performed the statistical analyses presented in the chapter (except for local ancestry and linkage disequilibrium estimates, for which I was assisted by Dr. Mario Barbato, and gene identification analyses, for which I was assisted by Dr. Marcello del Corvo), and wrote the first draft of the document.

7.3.3.2 Review on prioritization methods in conservation biology

I finalized the literature review on prioritization methods in conservation biology and wrote the manuscript. **Chapter 2** represents the result of my work: I reviewed around 30 methods, proposed a general classification scheme in form of decision tree, and highlighted some methodological integrations which might provide the basis for future research in the field of conservation genomics.

7.3.3.3 The study of *Bubalus bubalis* diversity

I contributed to finalize the analyses agreed with my supervisors: in particular, I performed analyses aimed at quantifying ascertainment bias in the dataset, population structure analyses, and part of the TREEMIX analyses.

7.3.3.4 Collaborations

Patrone V, Vajana E, Minuti A, Callegari ML, Federico A, Loguercio C, Dallio M, Tolone S, Docimo L, Morelli L, 2016. Postoperative Changes in Fecal Bacterial Communities and Fermentation Products in Obese Patients Undergoing Bilio-Intestinal Bypass. *Frontiers in Microbiology* **7**, doi: fmicb.2016.00200.

Here, I collaborated in the statistical analysis of the paper by developing customized R scripts. I also collaborated in the drafting of the manuscript, with special emphasis to those sections reporting my work.