Spring 5-2011

Pathogen-Mediated Evolution of Immunogenetic Variation in Plains Zebra (Equus quagga) of Southern Africa

Pauline L. Kamath

Pathogen-mediated evolution of immunogenetic variation in plains zebra (*Equus quagga*) of southern Africa

by

Pauline Lalitha Kamath

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor in Philosophy

in

Environmental Science, Policy and Management

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Wayne M. Getz, Chair
Professor Rauri C. K. Bowie
Professor Justin S. Brashares

Spring 2011

Pathogen-mediated evolution of immunogenetic variation in plains zebra (*Equus quagga*) of southern Africa

by Pauline Lalitha Kamath

Abstract

Pathogen-mediated evolution of immunogenetic variation in plains zebra (*Equus quagga*) of southern Africa

by

Pauline Lalitha Kamath

Doctor of Philosophy in Environmental Science, Policy and Management

University of California, Berkeley

Professor Wayne M. Getz, Chair

Investigating patterns of variability in functional protein-coding genes is fundamental to identifying the basis for population and species adaptation and ultimately, for predicting evolutionary potential in the face of environmental change. The Major Histocompatibility Complex (MHC), a family of immune genes, has been one of the most emphasized gene systems for studying selection and adaptation in vertebrates due to its significance in pathogen recognition and consequently, in eliciting host immune response. Pathogen evasion of host resistance is thought to be the primary mechanism preserving extreme levels of MHC polymorphism and shaping immunogenetic patterns across host populations and species. In this thesis, I examined the evolution of two equine MHC genes, *DRA* and *DQA*, over the history of the genus *Equus* and across free-ranging plains zebra (*E. quagga*) populations of southern Africa: Etosha National Park (ENP), Namibia and Kruger National Park (KNP), South Africa. Furthermore, I evaluated the relationships between the *DRA* locus and parasite intensity in *E. quagga* of ENP, to elucidate the mechanisms by which parasites have shaped diversity at the MHC.

In equids, the full extent of diversity and selection on the MHC in wild populations is unknown. Therefore, in this study, I molecularly characterized MHC diversity and selection across equid species to shed light on its mode of evolution in *Equus* and to identify specific sites under positive selection. Both the *DRA* and *DQA* exhibited a high degree of polymorphism and more intriguingly, greater allelic diversity was observed at the *DRA* than has previously been shown in any other vertebrate taxon. Global selection analyses of both loci indicated that the majority of codon sites are under purifying selection which may be explained by functional constraints on the protein. However, maximum likelihood based codon models of selection, allowing for heterogeneity in selection across codons, suggested that selective pressures varied across sites. Furthermore, at the *DQA* locus, all sites predicted to be under positive selection were antigen binding sites, implying that a few selected amino acid residues may play a significant role in equid immune function. Observations of trans-species polymorphisms and elevated genetic diversity were concordant with the hypothesis that balancing selection is acting on these genes.

Over the past half century, the role of neutral versus selective processes in shaping genetic diversity has been at the center of an ongoing dialogue among evolutionary biologists. To determine the relative influence of demography versus selection on the *DRA* and *DQA* loci, I contrasted diversity patterns of neutral and MHC data across the *E. quagga* populations of ENP and KNP. Neutrality tests, along with observations of elevated diversity and low differentiation across populations relative to nuclear intron data, provided further evidence for balancing selection at these loci among *E. quagga* populations. However, at the *DRA* locus, differentiation was comparable to results at microsatellite loci. Furthermore, zebra in ENP exhibited reduced levels of diversity relative to KNP due to a highly skewed allele frequency distribution that could not be explained by demography. These findings were indicative of spatially heterogeneous selection and suggested directional selection and local adaptation at the *DRA* locus.

There still remains a great deal of discussion over the mechanisms by which pathogens preserve immune gene diversity. The leading hypotheses that have been predominantly considered are: (i) heterozygote advantage (i.e. overdominant selection), (ii) rare allele advantage (i.e. frequency-dependent selection), and (iii) spatiotemporally fluctuating selection. An increasing number of studies have investigated MHC-parasite relationships to reconcile this debate, with conflicting results. To elucidate the mechanism driving the population-level patterns of diversity at the *DRA* locus, I examined relationships between this locus and both gastrointestinal (GI) and ectoparasite intensity in plains zebra of ENP. I discovered antagonistic pleiotropic effects of particular *DRA* alleles, with rare alleles predicting increased GI parasitism and common alleles associated with higher tick burdens. These results supported a frequency-dependent process and because maladaptive 'susceptibility alleles' were found at reduced frequencies, suggested that GI parasites exert strong selective pressure at this locus. Furthermore, heterozygote advantage also played a role in decreasing GI parasite burden, but only when a common allele was paired with a more divergent allele, implying that frequency-dependent and overdominant selection are acting in synchrony. These results indicated that an immunogenetic tradeoff may modulate resistance/susceptibility to parasites in this system, such that with MHC-based resistance to GI parasitism, a fitness cost is incurred to the host in the form of increased ectoparasite susceptibility. It is also suggested that these selective mechanisms are not mutually exclusive.

In conclusion, these results provided species and population-level evidence for selection on the equid MHC, and highlighted the complexity in which selection operates in natural systems. In addition to heterogeneity in selective pressures at the molecular-level (across a gene region), selection likely varies spatiotemporally across populations due to fluctuations in pathogen regimes. Furthermore, pleiotropic effects of multiple pathogens can obscure our ability to understand adaptive processes. Given the level of complexity in which selection operates, I emphasize the necessity of incorporating multiple lines of evidence, using both neutral and adaptive data, to illuminate how selection operates. Finally, I also highlight the importance of considering the selective effects of multiple pathogens on host immunogenetics to better understand MHC function and adaptation.

For my husband and family,
who have been my foundation

Table of Contents

Acknowledgements

There are a plethora of people that have supported me through my years of graduate study at Berkeley, from those who have helped in the field and lab to those who have contributed to my overall intellectual and personal growth. First of all, I am tremendously grateful to my advisor, Wayne Getz, for taking me in as an orphaned graduate student, giving me the opportunity to work in an amazing study system and being supportive throughout the years. I have highly valued his jovial personality, sense of humor and exemplary mentorship; I particularly admire his "open door" approach to graduate advising and encouragement of intellectual independence which allowed me to develop my own research ideas. My dissertation committee members, Rauri Bowie and Justin Brashares, have also had a profound influence on my development as a scientist. I thank Rauri for being a mentor– welcoming me to join his lab group meetings and providing me with the opportunity to interact with the MVZ community, providing access to lab resources, and most importantly for the thoughtful discussions that have aided in the progress of my research. Justin, as my committee member, professor, and teaching mentor has been extremely influential in my growth from a student to an educator; I am extremely grateful for the opportunities I had to teach under his guidance as a graduate student instructor. I also thank Craig Moritz, Claire Kremen, Richard Dodd and Walt Koenig for serving on my orals committee and guiding me through this "rite of passage" at Berkeley.

The opportunity to work on a large collaborative project (a.k.a. "Team Anthrax") has been especially rewarding, with people from diverse scientific backgrounds working towards a common goal. I specifically want to thank the people involved on this project for this incredible experience, from the writing of the EEID grant proposal to data collection and analyses: Wayne Getz, Justin Brashares, Russell Vance, Wendy Turner, Holly Ganz, Steve Bellan, Carrie Cizauskas, Miriam Tsyaluk, Leo Polansky, Royi Zidon, Orr Spiegel, Martina Küsters and Zepee Havarua.

I feel extremely fortunate to have had the incredible community of colleagues and friends at Berkeley that I have been able to interact with and learn from. The members of the Getz Lab have not only provided me with valuable intellectual support, but have also been a second family to me, making the lab a welcoming and fun environment. In particular, I would like to thank Wendy Turner for being an inspirational lab-mate, collaborator, and Pilates buddy. I am grateful for her help in getting my bearings in the field and for providing me with parasite data and thoughtful comments that were critical to a chapter of this thesis. Holly Ganz has been a role model to me and I want to thank her for generously giving her time to help me collect fecal samples in Kruger, providing me with helpful advice in the lab and on grant proposals/manuscripts. I thank Steve Bellan and Karen Weinbaum for imparting their wisdom on statistical modeling in R and valuable discussions regarding GLMM, GAMM, AIC, QIC, GEE, lmer, and all other statistical modeling acronyms! Kari Roesch Goodman, Margarita Hadjistylli, Emily Rubidge (i.e. the "Pop Gen Gals") and Anna Sellas have given continual guidance in every aspect of my genetic analyses, and I seriously could not have persevered without

their support. I thank Felipe Zapata for discussions on phylogenetic analyses, particularly in dealing with clone data. Juan Parra provided me with invaluable comments on my NSF DDIG grant proposal. Healy Hamilton, Mary Beth Rew, Anna Sellas and Emily Rubidge helped me through my early graduate years. I also thank Erica Spotswood, Cole Burton, Mike Wasserman, Tim DeChant, Shirley Huang, Emily Adams and her furry companions (Luna and Yeti) for their personal support. Kari Roesch Goodman, Orien Richmond, Suzie McIlroy and Kasey Fowler-Finn (a.k.a. "The Muffins") were a constant source of strength and inspiration, and I will take with me many wonderful memories of our time together. Most of all, these people have all provided me with invaluable friendships that have made this an unforgettable experience.

Many other people at Berkeley and beyond have contributed to the laboratory work and genetic analyses that were critical for the progress of this research. I am thankful for the opportunity to interact with the Museum of Vertebrate Zoology (MVZ) community. I also am grateful to Lydia Smith (MVZ) and Brian Ort (ESPM genetics facility) for being wonderful lab managers and for facilitating my work in the lab. Patrick O'Grady generously allowed me to use some of his lab's equipment when I was in need. I thank Miriam Tsyaluk for her help with microsatellite loci genotyping. My undergraduate assistants, Elyse DeFranco, Kaitlin Hall and Leonard Goo, were a delight to work with and tremendously helpful to this project by putting in many tedious hours of pipetting and setting up PCRs, among other invaluable work. Clint Epps provided me an essential protocol for extracting DNA from fecal samples. I thank Ivan Horak for his identification of tick species, Wendy Turner for gastrointestinal parasite quantification, and Martina Küsters for tick abundance estimates that provided the parasite data that was central to this research. I would also like to acknowledge Martina for her dedication to the anthrax project, her contribution as a collaborator to a chapter of this thesis, and her help in dealing with difficult permitting issues, for which I am extremely grateful. I also greatly appreciate the support from the Bowie Lab who welcomed me into their lab group meetings, for their intriguing discussions on phylogeography and population genetics, guidance in my genetic data analysis, and their thoughtful comments that have helped me to improve my proposals and manuscripts.

The Ministry of Environment and Tourism, Namibia, and South African National Parks provided the permits and support necessary for the sample collection phase of this research. In particular, members of the scientific and management staff at the Etosha Ecological Institute (EEI) made significant contributions to the sample collection phase of this research. I would especially like to thank Werner Kilian for his support and advice on this project and Wilfred Versfeld for providing me with samples. The following EEI staff members were extremely helpful in zebra captures and other aspects of sample collection: Werner Kilian, Wilfred Versfeld, Shayne Kötting, Birgit Kötting, Gabriel Shatumbu, Martin Kasona, Johanes Kapner, Nigel Berriman and Monika Shikango. The veterinarians were critical to these captures, making all of this work possible: Otwin Aschenborn, Mark Jago and Nad Brain. I thank Velly, my game guard in Kruger, for earnestly protecting me from danger while I collected samples. I also want to thank Peter Buss for managing my SANParks proposal and helping me with the arrangements for my

field collecting trips in Kruger. I am grateful to Amanda Bastos for comments and for providing the letter of support for my EEID grant proposal research.

I am extremely appreciative for the financial support to conduct this research, which was generously provided by the NIH Ecology and Evolution of Infectious Disease Program, NSF Doctoral Dissertation Improvement Grant, Carolyn Meek Scholarship, Edna and Yoshinori "Joe" Tanada Award, Ruth Hiebert Memorial Scholarship, and an ESPM departmental block grant. I would also like to acknowledge the ESPM administrative staff who have been very helpful through this process─ particularly Richard Battrick, Theodosia "Doty" Valrey, Cassia Stepak and Raul Torres.

Finally, I would like to thank my family for being my foundation─ they have supported me on this journey and I know they will be by my side as the adventure continues. I thank the multitude of people in my extended family, particularly my cousins here in the Bay Area who I feel fortunate to have had close by during this time period. My big furball, Lewy, is a constant source of comfort to me. My siblings, Tara and Ben, have provided me with their love and support, and also helped me through this process by inspiring me to travel and experience the world. I thank my parents for their love and encouragement throughout my life, which has made me who I am today. Last, and most importantly, I thank my husband, Olivier, who enthusiastically helped with sample collection, listened to countless practice talks and edited proposals, and has provided me with love, patience and support throughout this entire experience.

Chapter 1. Introduction

> *"Natural selection is not the wind which propels the vessel, but the rudder which, by friction, now on this side and now on that, shapes the course"*
>
> — *Asa Gray*

For over a half century, a primary aim in evolutionary biology has been to explain the processes and mechanisms shaping the patterns of genetic variation across populations and species. Since the discovery of high levels of enzyme diversity in human populations (Harris 1966), there has been an ongoing debate surrounding the mechanisms by which this unexplained variability could be maintained over time. Kimura's neutral theory (1983) postulated that most of the molecular variation in the genome is primarily caused by random fixation, through genetic drift of selectively neutral mutations (Kimura 1983). Neutral theory is also congruent with the idea that most mutations in functional protein-coding genes are deleterious, and hence are eliminated through the process of purifying selection. However, an interesting anomaly to this was observed at genes involved in immune function, where diversity was discovered to be at least two times higher than the genomic average (Gaudieri *et al.* 2000), a phenomenon which could not be explained without selection. Balancing selection, or long-term stability of polymorphisms, has now long been postulated as the primary means by which this high degree of genetic diversity could be maintained. Although neutral theory has been valuable to advancements in understanding evolution, it is the variability in functional genes that is directly responsible for adaptation. With increasing environmental change, examining the patterns of functional gene variation and the mechanisms by which they are driven is critical for understanding the evolutionary potential of a population or species.

Major Histocompatibility Complex (MHC)

The Major Histocompatibility Complex (MHC) is a multigene family that is central to gnathostome (i.e. jawed vertebrates) immune response and has been one of the best studied genetic systems in evolutionary biology. These genes encode glycoprotein receptor molecules that are responsible for recognizing and binding pathogenic antigens, then presenting these to specialized host immune cells (e.g. cytotoxic T-cells or T-helper cells) to trigger a cascade of host immune responses (Klein 1986). The genes of this complex are usually found clustered together in the genome, and encode molecules that fall into two main sub-families (Class I and II) that differ in their immunological function, specifically with respect to the broad pathogen/parasite groups that they recognize. Class I molecules are expressed on the surface of almost all nucleated cells, with the exception of sperm cells and some nerve cells, and are known to bind and present endogenous antigens derived from intracellular parasites (e.g. viruses). Class II molecules are found on specific antigen-presenting cells, such as macrophages and

lymphocytes, and bind and present exogenous peptides derived from extracellular parasites/pathogens (i.e. bacteria, macro-parasites). There is also a MHC class III sub-unit which is located between class I and II genes, but these genes encode immune complement components (e.g. cytokines) and, therefore, are not obviously related in structure or function to class I or II genes.

As MHC molecules are receptors, their general structure consists of a transmembrane domain, which anchors the molecules to the host cell surface, and an antigen recognition region, or peptide binding groove, where specific antigen binding sites (ABS) are located. These ABS are the amino acids known to directly bind foreign peptides and therefore, this particular region has received the most attention in disease and adaptation studies. Effective immune response not only requires specific binding of ABS to foreign peptides, but also to T-cells. Although the interaction of ABS and pathogenic peptides is thought to be the most selective of the processing stages, some MHC molecules are capable of binding multiple peptides with similar amino acids at their anchor sites (Altuvia & Margalit 2004).

Exceptional levels of MHC polymorphism have been reported across vertebrate taxa, and the interaction between pathogenic peptides and the host ABS are believed to be the primary mechanism by which this diversity is promoted (Doherty & Zinkernagel 1975; Hedrick 1994; Hughes & Hughes 1995; Hedrick & Kim 1998). In particular, as pathogens evade host resistance and evolve increased virulence, molecular evolution of host specificity at ABS must ensue. Thus, the resulting high level of diversity is driven by this 'evolutionary arms race' (Van Valen 1973), and believed to occur through pathogen-mediated balancing selection (see reviews by Meyer & Thomson 2001; Bernatchez & Landry 2003; Sommer 2005; Piertney & Oliver 2006). The role of parasites in shaping MHC diversity has been supported by observations that human populations with increased pathogen diversity also exhibit higher MHC diversity (Prugnolle *et al.* 2005). In addition, convincing evidence has been found from observations of specific MHC allele associations with parasite infections in humans (Hill *et al.* 1991) as well as wildlife (Paterson *et al.* 1998; Langefors *et al.* 2001; Froeschke & Sommer 2005; Meyer-Lucht & Sommer 2005; Schad *et al.* 2005).

Beyond its clear significance to pathogen resistance, the MHC has been implicated to effect other biological traits such as mate choice, kin recognition and maternal-fetal interactions (reviewed by Edwards & Hedrick 1998; Bernatchez & Landry 2003; Piertney & Oliver 2006). In particular, elevated MHC heterozygosity has been found to be involved with sexual selection, enhancing reproductive success in terms of fecundity, mating success, or both. For example, Sauermann et al, 2001, found that in a free-ranging population of rhesus macaques (*Macaca mulatta*), males with a heterozygous MHC genotype sired more offspring than homozygotes (Sauermann *et al.* 2001). In European brown hares (*Lepus europaeus*), heterozygosity at one MHC locus had a significant effect on female reproduction; homozygosity within sub-populations increased sterility and depressed fecundity (Smith *et al.* 2010). However, a simplistic explanation for these findings may be that these relationships are an indirect consequence of the proximate role of the MHC in conferring resistance to disease, such that higher reproductive success is a by-product of an individual having higher survivability. In other words, it would be expected that an individual with fewer pathogens should be healthier, have higher survivability, and be able to devote more resources to mating and reproduction. However,

a recent study on house mice (*Mus musculus musculus*) revealed that MHC heterozygosity was associated with enhanced mating and reproductive success even in the absence of any survival effects (Thoß *et al.* 2011). Although a more proximate role for the MHC in pathogen recognition is the most parsimonious explanation given its well-known function, the investigation of its role in sexual selection requires further attention.

Tests for selection and empirical evidence for balancing selection at the MHC

There are multiple lines of genetic data that have been put forth as evidence for balancing selection at the MHC. These include, but are not limited to: (i) high allelic diversity, (ii) trans-species polymorphisms, (iii) higher rates of nonsynonymous (d_N) relative to synonymous (d_S) mutations than expected under neutral evolution, (iv) heterozygous excess within populations, (v) uniform allele frequency distributions, and (vi) lower differentiation across populations than observed at neutral genetic loci. In general, it is necessary to combine multiple approaches, across both species and populations, in order to elucidate the selective processes at play.

As gene flow and drift (influenced by demography and migration) are also predominant evolutionary forces shaping the diversity and partitioning of genetic variation among species and populations, many tests for selection contrast the variation at MHC loci with theoretical predictions under the assumptions of neutrality. Further, comparing MHC data with empirical data from non-functional neutral loci provides a valuable null hypothesis and control for what proportion of the observed variation is generated by demographic processes. Therefore, contrasting observations at adaptive and neutral data can be extremely informative for studying selection at the MHC. Several selection tests have been developed (reviewed by Ford 2002; Bernatchez & Landry 2003; Garrigan & Hedrick 2003; Piertney & Oliver 2006), and some of those most widely used will be reviewed in the following section. These tests have been discussed in the context of approaches that are designed to examine long-term selective pressures through analyses using sequence-based polymorphism data as well as those designed to reflect recent selection within and among populations based on allelic and genotypic frequencies. A summary of the methods discussed and their interpretations can be found in Table 1.

High levels of diversity

As mentioned previously, extreme polymorphism at the MHC has been attributed to balancing selection acting on this gene family. This high level of MHC diversity has even been preserved in genetically depauperate populations; for example, the San Nicolas island fox (*Urocyon littoralis dickeyi*) population has undergone long-term population declines, but yet have high levels of MHC diversity (Aguilar *et al.* 2004). Analogous results are found in fragmented gray mouse lemur (*Microcebus murinus*) populations of Madagascar, with high allelic and gene diversity relative to diversity observed at neutral microsatellite loci (Schad *et al.* 2004). In several further examples, MHC diversity has been found to exceed that at neutral loci, although in some instances this tendency has

Table 1. Summary of evidence from genetic data used to assess selection

Line of evidence	Level of analysis	Observation	Interpretation
Levels of diversity	any	equal to neutral data	neutral evolution
		low compared to neutral data	purifying/ directional selection
		high compared to neutral data	balancing selection
Gene trees	among species	same as species trees	neutral evolution
		trans-species polymorphism	balancing selection
d_N:d_S ratio	among species	$d_N = d_S$	neutral evolution
		$d_N < d_S$	purifying selection
		$d_N > d_S$	positive selection
Tajima's D / Fu's F_S	within species/ populations	not significant	neutral evolution
		significantly negative	purifying selection OR population expansion
		significantly positive	balancing selection OR population bottleneck
Population differentiation	among populations	same as neutral loci	neutral processes (drift, gene flow) play greater role
		higher than at neutral loci	spatially heterogeneous, directional selection/ local adaptation
		lower than at neutral loci	balancing selection
Allele frequency distributions	within populations	normal	neutral evolution
		skewed	directional selection
		even	balancing selection
Ewens-Watterson Test	within populations	$F_{OBS} = F_{EXP}$	neutral evolution
		$F_{OBS} > F_{EXP}$	directional selection
		$F_{OBS} < F_{EXP}$	balancing selection
Genotype proportions[1]	within populations	heterozygotes deficiency	purifying/directional selection OR inbreeding
		heterozygote excess	overdominant balancing selection OR outbreeding

d_N = rate of non-synonymous mutation; d_S = rate of synonymous mutation; F_{OBS} = observed homozygosity, F_{EXP} = expected homozygosity based on an allele frequency distribution under mutation-drift equilibrium, given a sample of the same size with the same number of alleles.
[1]*Deviations from Hardy-Weinberg equilibrium can also reflect population subdivision, assortative mating, overlapping generations or null alleles.*

been contradicted. For example, limited polymorphisms have been observed in the MHC of fin (*Balaenoptera physalus*) and sei whales (*B. borealis*) (Trowsdale *et al.* 1989). Also, in bighorn sheep (*Ovis canadensis*) populations, microsatellite heterozygosity was found to be higher than that of MHC loci (Boyce *et al.* 1997). This observation led to the authors' conclusion that balancing selection, if operating at all, is weak and masked by other more significant micro-evolutionary forces, specifically significant population bottlenecks. In addition, lack of diversity at particular MHC loci has been observed and speculated to be the result of functional constraints (e.g. the *DRA* locus; Chu *et al.* 1994).

Species-level selection:

Trans-species polymorphisms

Balancing selection is expected to result in the maintenance of polymorphisms over long time periods. Under coalescent theory, neutral mutations are not expected to persist across multiple speciation events (existing only ~4Ne generations). Balancing selection, however, has been theoretically shown to slow down this mutational process and generate deep allelic lineages (Takahata 1990). This phenomenon may result in allele divergences that pre-date species divergences, causing alleles to be more closely related between species than within species ('trans-species polymorphisms') and thus, detectable by discordance between gene versus species trees (Figueroa *et al.* 1988). Trans-species polymorphisms have proven to be widespread at the MHC (for a few examples see Van den Bussche *et al.* 1999; Ottova *et al.* 2005; Bryja *et al.* 2006; Cutrera & Lacey 2007; Mona *et al.* 2008) with even identical alleles shared among related species (e.g. Otting *et al.* 2002). Alternatively, high allelic divergence at the MHC within species could also be due to inter-locus recombination and gene conversion (Martinsohn *et al.* 1999) confounding conclusions about the origin of MHC alleles. Thus, the interpretation of these observations should be approached with caution and tests for recombination simultaneously employed.

d_N/ d_S ratio test

One of the most commonly applied methods for detecting selection over longer time periods is the d_N/d_S ratio test (Hill & Hastie 1987; Hughes & Nei 1988). This test is based on the fact that synonymous mutations (d_S), or mutations that do not result in a change at the amino acid level, are effectively neutral. In contrast, non-synonymous mutations (d_N) ultimately resulting in a protein change are, thus, subject to selection pressures. Therefore, by comparing rates of d_N to d_S one can determine whether mutations are deleterious ($d_N < d_S$), advantageous ($d_N > d_S$), or neutral ($d_N = d_S$), reflecting purifying selection, positive selection or neutral evolution, respectively. This test is useful because it does not make any assumptions about population structure or equilibrium (Nielsen 2001). However, the d_N/d_S test is often applied to full gene sequences and thus, reflects the average global selection pressure over all sites. The problem with this approach is that often selection is heterogeneous across gene regions with positive selection acting on single isolated sites. Small positively selected changes may have large consequences for gene function and ultimately fitness, with conservation of the majority of sites required to

maintain protein structure and integrity. This implies that these whole-gene tests may not be sensitive enough to detect fine-scale selection and may sometimes lead to misleading conclusions. To address this, codon site-specific models have proven to be useful for elucidating how rates of evolution vary across a gene region and for pin-pointing particular sites under selection (Yang & Bielawski 2000; Yang *et al.* 2000) such as those that specifically interact and recognize foreign peptides. In particular, site-specific methods have been valuable for identifying elevated d_N/d_S ratios at ABS and have suggested substantially differing rates of evolution across the MHC (Hughes & Hughes 1995).

Tajima's D and Fu's F_S

Another sequence-based approach useful for examining selection over the history of a species is Tajima's *D* test (Tajima 1989). Again, this test is based on mutational data, so reflects both historical and recent selective processes and is useful for investigating the strength of selection over longer time scales. But, as opposed to the d_N/d_S method, it is often used to test for selection within populations. The test statistic, *D*, is based on the comparison of two estimators of the neutral diversity parameter ($4N_e\mu$): the first is based on the average pair-wise sequence divergence (θ) and the second on the number of segregating sites (*S*). Both of these estimators are expected to produce the same value under the expectation of neutrality, and therefore *D* measures whether the ratio is significantly different from this expectation. Fu's F_S test (Fu 1997) is similar to Tajima's *D*, but instead measures the probability of the observed number of alleles given θ. Significantly positive values of the *D* or F_S statistic are expected to reflect balancing selection and negative values, directional purifying selection. However, positive or negative results may also reflect the demographic events of a population bottleneck or population expansion, respectively. Therefore, these tests should be interpreted in contrast to data from neutral loci.

Population-level selection:

Differentiation among populations

Exploring the differentiation (F_{ST}; Wright 1951) between populations over space and time can provide valuable information regarding the nature of selection at the MHC. However, partitioning of MHC variation can be confounded by demographic processes, such as changes in population size or migration. Therefore, contrasts with neutral data (e.g. mtDNA, microsatellites, single nucleotide polymorphisms, nuclear introns) are necessary to control for these processes. Under balancing selection, higher MHC diversity within and lower differentiation between populations relative to neutral loci is expected, as polymorphisms should be retained over time, even in scenarios of restricted migration (Schierup *et al.* 2000). This observation has been discovered in many natural studies; for example, among small isolated alpine chamois (*Rupicapra rupicapra***)** populations of the Italian Alps, Mona *et al.* (2008) reported very low MHC gene differentiation relative to that observed at a mitochondrial DNA (mtDNA) locus, despite the expectation that these populations would be strongly subject to the effects of genetic drift. In another example, lower differentiation was observed at two MHC genes than at the mtDNA locus in Malagasy jumping rats (*Hypogeomys antimena*) (Sommer 2003).

However, in this study, it is also possible that the elevated mtDNA divergence could have also been due to sex-biased dispersal, as females are known to exhibit a high degree of philopatry.

However, many studies have alternatively found equivalent or lower levels of differentiation in neutral versus MHC loci (reviewed in Bernatchez & Landry 2003). One explanation for comparable differentiation among MHC and neutral loci is that migration and drift may play a larger role in shaping MHC variation on a micro-evolutionary time scale, particularly in populations with smaller effective population sizes that are subject to stronger drift effects. For example, in tuatara (*Sphenodon spp*.) island populations, MHC and microsatellite loci were both highly differentiated, suggesting that the combined effects of population bottlenecks and isolation played a larger role than selection at shaping MHC variation (Miller *et al.* 2010). Scenarios where differentiation at MHC loci has been reported to be greater than neutral loci have been attributed to local adaptation that varies over geographic time scales. For example, Landry & Bernatchez (2001) found higher differentiation at the MHC relative to neutral data between different spawning sites of Atlantic salmon (*Salmo salar*) within rivers and attributed this to local adaptation at these sites. Interestingly, their data over larger geographic areas (i.e. between rivers) revealed equivalent levels of differentiation and suggested that demographic processes override the effects of selection. Miller *et al*, 2001, similarly reported high MHC differentiation relative to microsatellites in sockeye salmon (*Oncorhynchus nerka*) and concluded this to be a result of temporal or spatial heterogeneity in pathogen pressures among localities. A study of Soay sheep (*Ovis aries*), contrasting MHC and neutral differentiation over a 13 year period, found that relative differentiation varied temporally; with MHC data comparable to neutral data for a large portion of the time, but lower than neutral data for four years of their study (Charbonnel & Pemberton 2005). Their results provided an intriguing example of temporal variation in selection within a natural population over a very short time scale, most likely due to fluctuating parasite pressures.

Allele-frequency distributions and genotypic proportions

Inspection of the allele frequency distributions and genotypic proportions within populations has been suggested to be valuable for characterizing selection. For instance, balancing selection is expected to preserve existing diversity through the selection of rare advantageous alleles and/or selection of heterozygote genotypes (discussed further below). Therefore, the allele frequency distribution within a population is predicted to be more even than that expected under neutrality. Alternatively, if a particular allele or mutation is directionally being selected for, a skew in this distribution may be expected. Therefore, evaluation of allele frequency distributions is not only valuable for understanding if balancing selection is occurring, but also for investigating potential variability in selection pressures within and among populations and/or species (e.g. Cutrera *et al.* 2010).

The Ewens-Watterson (E-W) homozygosity test of neutrality (Ewens 1972; Watterson 1978) is a widely used method for testing for balancing selection within populations based on allelic patterns. It is founded on the expectation that under balancing selection, rarer alleles will be present at higher frequencies than expected under neutrality, thereby resulting in more even allele frequency distributions (described above). Therefore, balancing selection should be detectable by comparing the observed

homozygosity (F_{OBS}) to that theoretically expected under neutral evolution (F_{EXP}). Expected homozygosity is derived from a sampling distribution given the same sample size and number of alleles under the assumption of mutation-drift equilibrium. At the MHC, homozygosity is expected to be lower than neutral expectations given the even allele frequency distributions. The E-W test has been applied to natural populations and in some cases has indicated significant balancing selection at the MHC (Paterson *et al.* 1998; Landry & Bernatchez 2001; Miller *et al.* 2001; Hambuch & Lacey 2002). However, Ejsmond *et al* (2010) through simulation suggested that the E-W test may be limited in its ability to detect balancing selection, depending on the specific balancing selection mechanism at play (see below). They showed that a frequency-dependent mechanism resulted in unstable allele frequency distributions that were sometimes theoretically indistinguishable from neutral expectations, and therefore concluded that E-W test results can be unreliable. In contrast to this, under a heterozygote mechanism the E-W test was shown to easily detect balancing selection due to consistently even allele frequency distributions.

Deviations from Hardy-Weinberg equilibrium

Within populations, deviations from expected genotypic proportions under Hardy-Weinberg (H-W) equilibrium (Hardy 1908; Weinberg 1908) may provide insight into selection at the generational level. For example, heterozygote deficiency, which can be measured by the inbreeding coefficient (F_{IS}), may be indicative of purifying or directional selection. Conversely, heterozygote excess (or homozygote deficiency) at MHC loci may be expected under the hypothesis of 'over-dominant' balancing selection (Doherty & Zinkernagel 1975), which claims that heterozygote genotypes have higher fitness than homozygotes (discussed below). Whereas heterozygote excess has been observed in some human-MHC studies (e.g. Black & Salzano 1981; Chen *et al.* 1999), in general, few data support heterozygote excess at the MHC. This approach is best suited for addressing this particular type of balancing selection, and therefore many studies may have failed to detect heterozygote excess due to the likelihood of a more complicated selective mechanism acting on these loci. Further, deviations from H-W equilibrium may be considered to be a weak test for selection because of the multitude of other diverse interpretations, including non-random mating, population sub-division or overlapping generations. Therefore, tests for selection based on deviations from H-W should be employed along with contrasts to neutral loci, as well as in conjunction with more sensitive selection tests.

Pathogen-driven selective mechanisms for balanced polymorphisms

Several hypotheses on how pathogens drive patterns of variation at the MHC have been at the forefront of continual discussion, and have been examined by both theoretical and empirical studies. Three primary mechanisms for explaining balancing selection are generally proposed: (1) heterozygote advantage, (2) frequency-dependent selection and (3) fluctuating selection (reviewed in Potts & Slev 1995; Apanius *et al.* 1997; Sommer 2005; Piertney & Oliver 2006).

Heterozygote advantage

The 'heterozygote advantage' hypothesis is based on the idea that individuals with heterozygous genotypes (two different MHC alleles at a locus) are capable of responding to a broader range of pathogens in their environment than homozygotes with only one MHC variant. Thus, under this mechanism, it is postulated that heterozygotes should benefit from increased pathogen resistance and have higher fitness relative to homozygotes (Doherty & Zinkernagel 1975). In evaluating this mechanism, it is important to distinguish between several different ways to view heterozygote advantage. First, it may be thought of in the broad sense where heterozygotes are generally more fit due to the masking of susceptibility alleles by dominant alleles of higher fitness value, also referred to as 'dominance'. This specifically requires that heterozygotes are at least as fit as the most resistant homozygote or, in a population sense, are more fit than the average of the parental homozygotes, a mechanism which has been shown to be an inefficient means to promote and maintain MHC diversity (Gould *et al.* 2004; Sommer 2005). In contrast, under the model of 'overdominance', a specific heterozygote genotype is predicted to have higher fitness than either homozygote of its parental alleles. 'Overdominance' has been thought to be an appropriate model when considering that hosts are most likely challenged by multiple pathogens or parasites (Hughes & Nei 1992), as heterozygotes have higher variability at ABS and, therefore, a greater probability of recognizing a diversity of pathogen-derived peptides (Hughes & Nei 1989; Takahata & Nei 1990).

To more specifically refine these hypotheses, the 'overdominant' mode of selection can also be considered as occurring through: (i) 'symmetric overdominance' (Takahata 1990), a simplistic model where all heterozygotes in a population are predicted to be of equivalently higher fitness, or (ii) 'divergent allele advantage' (Wakeland *et al.* 1990), a more molecularly explicit model suggesting that heterozygote fitness level is dependent on the degree of mutational differences between the two alleles in the genotype. Experimental co-infection studies have found convincing evidence to support the mechanism of heterozygote advantage in terms of both 'dominance' (Penn *et al.* 2002) and 'symmetric overdominance' (McClelland *et al.* 2003). Although few examples exist, there is some empirical evidence from natural studies to support the latter (but see Oliver *et al.* 2009b). This paucity of support from natural studies is possibly due to that fact that appropriate testing of the overdominance theory should be done in the context of multiple parasites and using MHC systems with limited diversity that have sufficient power for statistical analyses, as demonstrated in Olivier *et al*., 2009b. However, this type of data is often not available or applicable to the majority of study systems. Finally, the divergent allele hypothesis has probably received even less attention than that of 'symmetric overdominance'. One exception is a theoretical analysis that used data from a class II MHC gene in deer mice (*Peromyscus maniculatus*) to confirm that the divergent allele advantage better explained the observed MHC gene genealogy than a more simplistic symmetric overdominant model (Richman *et al.* 2001). However, to my knowledge, there have not been any studies that have examined this hypothesis and the direct role it plays in conferring pathogen resistance in wild populations.

Frequency-dependent selection

The second key balancing mechanism proposed is frequency-dependent selection, which asserts that the fitness advantage of any particular allele varies and is reliant on its frequency in the population (Kojima 1971; Takahata & Nei 1990). This type of selection is driven by the 'evolutionary arms race' (Van Valen 1973) between host and pathogen, and assumes that rare alleles are advantageous to the host as pathogens have not yet evolved a means of evasion. Thus, as rare alleles are favored and driven up in frequency, the relative fitness of common alleles is reduced, causing them to subsequently decrease in frequency. Again, as rare alleles become common, pathogens evolve to avoid host recognition, bringing us back to the beginning of this cycle. This type of cyclical interplay between host and pathogen has been shown to be theoretically capable of driving a balanced polymorphisms and maintaining high levels of diversity at the MHC (Takahata & Nei 1990). Evidence for this theory has largely been supported by observations of specific alleles associated with increased resistance or, alternatively, susceptibility, although this is considered to be indirect inference (Sommer 2005). Many natural studies have found MHC allele-pathogen associations, and suggested that this was evidence for frequency-dependent selection: in brook charr (Croisetiere *et al.* 2008), Atlantic salmon (Langefors *et al.* 2001), Malagasy mouse lemurs (Schad *et al.* 2005) and striped mice (Froeschke & Sommer 2005).

Spatiotemporally fluctuating selection

Because of the time-lag inherent to the host-pathogen interaction, as hosts evolve resistance and pathogens respond by evading host recognition, selection pressure is expected to fluctuate over time in any given environment as well as vary across environments. Such heterogeneity over space and time seem to be more realistic given that external conditions (e.g. temperature or rainfall) mediating host-pathogen interactions may also vary spatiotemporally. Therefore the third major balancing selection hypothesis that has been receiving increased attention is that of 'diversifying selection over space and time'(Hedrick 2002). Several natural studies have demonstrated selection on the MHC that varies spatially across populations by contrasting patterns of neutral and MHC variation across populations (e.g. Landry & Bernatchez 2001; Miller *et al.* 2001; Alcaide *et al.* 2008; Loiseau *et al.* 2009), but few have linked this to spatial differences in parasite prevalence or intensity (but see exception below). One reason may be that few studies have had the appropriate tools to examine this hypothesis over time due to limitations in the scope of most studies (Apanius *et al.* 1997). One study, however, used time series data from Great reed warblers (*Acrocephalus arundinaceus*) to demonstrate that MHC alleles exhibited greater changes in frequency over time than neutral microsatellite alleles (Westerdahl *et al.* 2004). Another study also examined between-year MHC data from a water vole (*Arvicola terrestris*) meta-population, and found evidence that the strength of selection varied across sub-populations and years sampled (Oliver *et al.* 2009a). However, both studies solely investigated temporal patterns in genetic data without associated parasite data, and thus were not able to link their observations directly to possible parasite-related mechanisms. Charbonnel and Pemberton (2005) discovered both temporal and spatial fluctuations of MHC diversity in the well-studied Soay sheep (*Ovis aries*) population of St. Kilda. Their study, in contrast to the others, was able to discuss their results in the context of previous data that described spatial variation in gastrointestinal (GI) nematode fecal egg counts (FEC)

across hefts (spatial sub-units of the population) (Wilson *et al.* 2003). Consequently, they discovered stronger differentiation among hefts at MHC loci than at neutral loci and, more intriguingly, temporal MHC divergence within the heft with the highest parasite intensity.

Empirical identification of parasite-mediated mechanisms

Distinguishing among the various parasite-mediated selective mechanisms has proven to be an exceptional challenge (Spurgin & Richardson 2010). Firstly, the inspection of MHC allele frequency distributions alone does not allow for identifying frequency-dependence from heterozygote advantage, as simulation models have showed they are oftentimes identical in appearance (Ejsmond *et al.* 2010). Furthermore, spatiotemporally fluctuating selective pressures can obscure interpretations, thereby confounding conclusions if only a portion of the story is being assessed (i.e. if there is an unknown pathogen or locus playing a significant role). Another problem in making a distinction between selective mechanisms is that the theories of heterozygote advantage and frequency-dependence are most likely not mutually exclusive, and could occur at the same time (e.g. Froeschke & Sommer 2005; Evans & Neff 2009; Oppelt *et al.* 2010). For example, Froeschke & Sommer (2005) suggested both mechanisms are important in striped mice (*Rhabdomys pumilio*), after finding specific alleles associated with GI nematode resistance and that heterozygotes had significantly lower parasite loads. Similarly, Oppelt *et al.* (2010) found an association between a specific MHC allele and intestinal coccidian resistance in European rabbits (*Oryctolagus cuniculus*), but only when this particular allele was found in heterozygous genotypes.

Some of the difficulty in interpretation may also be due to these mechanisms operating on different scales, driven by both the intensity of a specific pathogen and pathogen species richness, thereby confounding conclusions. For example, particular alleles may be responsible for specific pathogenic antigen recognition, modulated in a frequency-dependent manner based on pathogen pressure, but at the same time heterozygosity may also be important for conferring resistance to multiple parasites. Few studies, however, have addressed the effects of multiple parasites and/or parasite communities on MHC diversity patterns. In one example, Evans and Neff, 2009, examined the relationships between MHC diversity and bacterial community infections in Chinook salmon (*Oncorhynchus tshawytscha*), finding strong evidence for heterozygote advantage across populations, but also that MHC alleles were associated with infection susceptibility to specific bacterial parasites. They attributed these results to the complexity of the bacterial community influencing diverse host immune adaptations. Ditchkoff *et al.* (2005) also provided evidence from white-tailed deer (*Odocoileus virginianus*) for an immune gene tradeoff between GI nematodes and ectoparasite ticks, with alleles from different evolutionary clades conferring resistance to different parasites and deer with alleles from both lineages having moderate resistance to both parasite types (Ditchkoff *et al.* 2005).

Interestingly, many of these studies have revealed perplexing findings of MHC 'susceptibility' alleles, or alleles associated with increased rather than decreased resistance to pathogens or parasites. 'Susceptibility alleles' have been observed in wildlife populations, including that of fish (Langefors *et al.* 2001; Croisetiere *et al.* 2008;

Evans & Neff 2009), birds (Loiseau *et al.* 2008) and mammals (Paterson *et al.* 1998; Froeschke & Sommer 2005; Meyer-Lucht & Sommer 2005; Schad *et al.* 2005; Schwensow *et al.* 2007). These observations are confusing as maladaptive allele/genotypes are expected to be eliminated from the population. But, again, if viewed in the context of a multi-parasite system, it is possible that MHC loci exhibit antagonistic pleiotropic effects. For example, one allele at a locus may be effective for fighting one parasite, but ineffective for another parasite or adaptive function (reviewed by Slev & Potts 2002). The most classic example of this is sickle-cell anemia which results from a single-point mutation in the hemoglobin gene. While conferring decreased fitness with respect to this disease, individuals with one copy of the mutant hemoglobin allele also have increased resistance to malaria (Allison 1954). Alternatively, the observation of 'susceptibility alleles' may be the result of pathogen evasion of host alleles expected during the frequency-dependent process, as described previously (Slev & Potts 2002). However, because this process is one of constant flux, study results may only reflect a snapshot of a particular point in time.

Study system background

Equine Lymphocyte Antigen (ELA) genes

Given the long history of investigation on MHC structure, function and selective mechanisms, this gene system is an extremely valuable candidate for studying adaptation in natural populations (Hedrick 1994; Bernatchez & Landry 2003). MHC genes of the family Equidae, also called the Equine Lymphocyte Antigen (ELA), have been structurally described (Gustafson *et al.* 2003) and are believed to be conserved in general function to that of humans and other vertebrates (Madden 1995). Although homologous to the human MHC (HLA), the ELA is unique in some aspects of its architecture. For example, in situ hybridization revealed that the majority of ELA genes are clustered together on chromosome 20 (Mäkinen *et al.* 1989), however, a homologous copy of the ELA-*DQA* locus was localized to chromosome 5 (Bailey *et al.* 2000). This physical separation of MHC genes over chromosomes of the genome is a phenomenon that has not been observed in any other vertebrate taxon, with the exception of MHC genes in zebra finch (*Taeniopygia guttata*), which were recently reported to be distributed over four chromosomes (Balakrishnan *et al.* 2010). In addition, a large duplication was found in the region encompassing the ELA class I and class III boundary and is suggested to have functional significance due to its conservation through Perissodactyl (i.e. odd-toed ungulate) evolution (Brinkmeyer-Langford *et al.* 2010).

Another unique feature of the ELA is that the class II *DRA* locus has been shown to exhibit greater allelic diversity in Equidae than has ever been observed in any other taxon. For example, in just the exon 2 region (which contains the functional ABS), 19 unique alleles were discovered over all equid species, with five to six alleles observed in each of the following equid species: domestic horse (*E. callabus*), donkey (*E. asinus*), plains zebra (*E. quagga/ E. burchelli*) and mountain zebra (*E. zebra*) (Albright-Fraser *et al.* 1996; Brown *et al.* 2004; Luis *et al.* 2005; Janova *et al.* 2009). Interestingly, humans as well as the majority of other vertebrate species exhibit little to no variation at this locus (e.g. Wagner *et al.* 1995; Yuhki & O'Brien 1997; Takada *et al.* 1998). The lack of

diversity at the *DRA* over most species is a paradox and, in contrast, the unusually high variation observed in equids is intriguing and remains unexplained.

The *DRA* and *DQA* are two equine MHC genes that encode the α-chain of class II molecules and are predicted to have antigen-presenting functions that are similar to homologous HLA genes. While the diversity in these genes has previously been investigated, these studies have primarily focused on domestic or captive equids (Albright-Fraser *et al.* 1996; Fraser & Bailey 1998; Brown *et al.* 2004; Janova *et al.* 2009). To my knowledge, no studies have examined diversity, selection or the functional significance of ELA genes in wild equid populations.

Pathogen pressures in plains zebra (E. quagga) of southern Africa

Plains zebra (*E. quagga*, formerly *E. burchelli*), the most geographically widespread wild equid species, provide an excellent natural system in which to study the evolution of ELA immune genes and elucidate the mechanisms by which parasites shape patterns of adaptive variation. In particular, Etosha National Park (ENP), Namibia and Kruger National Park (KNP), South Africa, support large zebra populations in semi-arid grassland savannas, yet these populations are exposed to different pathogen pressures. In ENP, *E. quagga* are the main host species of anthrax, a deadly disease caused by the bacterial pathogen, *Bacillus anthracis*, and severe outbreaks occur annually in this population (Turnbull *et al.* 1989; Lindeque & Turnbull 1994). In contrast, *E. quagga* of KNP are only mildly affected by anthrax, and sporadic outbreaks occur on an approximate decadal cycle (de Vos 1990). Plains zebra are also appreciably infected by gastrointestinal (GI) nematodes, found in both populations at nearly 100% prevalence (Krecek *et al.* 1987a; Krecek *et al.* 1987b; Turner & Getz 2010). However the GI nematode communities were shown to be variable across populations with 32 nematode species discovered in KNP compared with the 21 species in ENP (Matthee *et al.* 2004). In particular, Matthee *et al*, 2004, found that this difference was specifically due to fewer 'large strongyles' species (Family: Strongylinae) in zebra of ENP, and attributed this to potential climatic differences; specifically, ENP has lower mean annual rainfall (<500mm/yr) than KNP (550-650 mm/yr) (Scialdo-Krecek *et al.* 1983). Congruent results have been observed in the diversity of hard-bodied tick species across populations and previous studies have identified 5 Ixodidae *spp.* in ENP (Horak *et al.* 1992) and 7 Ixodidae *spp.* in plains zebra of KNP (Horak *et al.* 1984b). The differences in pathogen pressures across ENP and KNP zebra populations provides a unique system with which to examine how selection acts to shape MHC variation on a geographic scale. Furthermore, the previous and ongoing extensive research on parasitism within plains zebra of ENP presents an excellent opportunity to examine and identify the parasite-mediated adaptive mechanisms driving these immunogenetic diversity patterns.

Thesis research: predictions and objectives

The primary objective of this research was to investigate the evolution of equid MHC genes (ELA) at both macro- and micro-evolutionary time scales in order to illuminate the role of selection on the ELA over the genus *Equus* and between two natural plains zebra (*E. quagga*) populations (ENP and KNP). I specifically examined the exon 2 regions of

two ELA genes, *DRA* and *DQA*, which were predicted to encompass the functional ABS based on HLA equivalents. As class II MHC molecules, these genes were expected to be important for recognizing bacterial and macro-parasite peptides and therefore, to be under selection across equid species and populations. At the population level, the considerable difference in endo- and ectoparasite richness was hypothesized to generate geographic variability in selective pressures and local adaptation, thereby affecting ELA gene differentiation and diversity across host populations. Specifically, I expected that ELA diversity would be higher in KNP zebra where parasite diversity was also higher. Whereas more limited parasitism in ENP zebra, along with intense selective pressure by anthrax, may result in lower ELA diversity. Conversely, balancing selection may be occurring to retain comparable diversity patterns at these genes across populations due to particular similarities in their pathogen communities and given the specific function of the candidate ELA genes examined here. Finally, this research aimed to link the observed selective patterns in *E. quagga* of ENP with possible parasite-mediated mechanisms (as described above) by examining the relationships between immune gene diversity and parasite intensity, using data from multiple parasite types (GI and ectoparasites). These results were expected to reveal that variation in immune genes predicts parasitism and that by examining host immunogenetic effects on these two very different parasite types would reveal multiple selective mechanisms at play. Finally, these results were collectively expected to underscore the adaptive significance of the ELA in *E. quagga*.

This research has been conducted in three parts, each of which is an individual chapter in this thesis:

1. Adaptive evolution of the Major Histocompatibility Complex genes, *DRA* and *DQA*, in the genus *Equus*

2. Relative influence of demography and selection on immune gene variation in plains zebra (*Equus quagga*) populations of southern Africa

3. Parasite-mediated selection and immunogenetic tradeoffs in plains zebra (*Equus quagga*) of Etosha National Park, Namibia.

Chapter 2. Adaptive evolution of the Major Histocompatibility Complex genes, DRA and DQA, in the genus Equus

"Natural selection is a mechanism for generating an exceedingly high degree of improbability"

– Ronald Aylmer Fisher, 1931

Abstract

Major Histocompatibility Complex (MHC) genes are central to vertebrate immune response and believed to be under balancing selection by pathogens. This hypothesis has been supported by observations of extremely high polymorphism, elevated nonsynonymous to synonymous base pair substitution rates and trans-species polymorphisms at these loci. In equids, the organization and variability of this gene family has been described, however the full extent of diversity and selection is unknown. As selection is not expected to act uniformly on a functional gene, maximum likelihood codon-based models of selection that allow heterogeneity in selection across codon positions can be valuable for examining MHC gene evolution and the molecular basis for species adaptations. I investigated the evolution of two class II MHC genes of the Equine Lymphocyte Antigen (ELA), *DRA* and *DQA*, in the genus Equus with the addition of novel alleles identified in plains zebra (E. quagga, formerly E. burchelli). I found that both genes exhibited a high degree of polymorphism and inter-specific sharing of allele lineages. To my knowledge, *DRA* allelic diversity was discovered to be higher than has ever been observed in vertebrates. Evidence was also found to support a duplication of the *DQA* locus. Selection analyses, evaluated in terms of relative rates of nonsynonymous to synonymous mutations (d_N/d_S) averaged over the gene region, indicated that the majority of codon sites were conserved and under purifying selection ($d_N < d_S$). However, the most likely evolutionary codon models allowed for variable rates of selection across codon sites at both loci and, at the *DQA*, supported the hypothesis of positive selection acting on specific sites. Observations of elevated genetic diversity and trans-species polymorphisms supported the conclusion that balancing selection may be acting on these loci. Furthermore, at the *DQA*, positive selection was occurring at antigen binding sites, suggesting that a few selected residues may play a significant role in equid immune function. Future studies in natural equid populations will be valuable for understanding the functional significance of the uniquely diverse *DRA* locus and for elucidating the mechanism maintaining diversity at these MHC loci.

Background

Genes of the Major Histocompatibility Complex (MHC) are ideal candidates for investigating the influence of selection in promoting patterns of genetic diversity (Hedrick & Kim 1998; Meyer & Thomson 2001), due to their ecological significance. This multi-gene family has been widely demonstrated to play a fundamental role in gnathostome (i.e. jawed vertebrate) immune response by modulation of resistance to parasites and pathogens (Doherty & Zinkernagel 1975; Potts & Slev 1995; Hedrick & Kim 1998). More specifically, class I and II MHC genes encode cell-surface glycoproteins that recognize foreign antigen molecules and, subsequently, present them to T-lymphocytes to initiate an immune system response in the host (Klein 1986). The MHC is known to be the most polymorphic gene region in vertebrates and, in humans, exhibits levels of nucleotide diversity that are two times higher than the genomic average (Gaudieri *et al.* 2000). Evidence from studies of natural populations suggests that this elevated genetic diversity is driven and maintained by exposure to pathogens and parasites in the environment. For example, studies on sheep (Buitkamp *et al.* 1996; Paterson *et al.* 1998), mice (Meyer-Lucht & Sommer 2005), voles (Kloch *et al.* 2010) and lemurs (Schad *et al.* 2005) have found relationships between gastrointestinal parasites and MHC diversity or associations between specific alleles and infection levels. This vital role for the MHC in pathogen recognition has been the subject of much investigation. Further study of selection at the molecular level, however, is imperative to facilitate understanding of the mechanistic basis for adaptation in natural systems.

The MHC is believed to be under strong selective balancing pressure (*reviewed in* (Piertney & Oliver 2006) under the key hypothesized mechanisms of negative frequency-dependent (Kojima 1971; Takahata & Nei 1990) and overdominant selection (Doherty & Zinkernagel 1975; Hughes & Nei 1988; Hughes & Nei 1989). Balancing selection is often supported by three lines of evidence: (1) high levels of polymorphism, (2) higher rates of nonsynonymous (d_N) to synonymous (d_S) nucleotide substitutions than what would be expected under neutral evolution (Hughes & Nei 1988; Hughes & Nei 1989) and (3) trans-species polymorphisms with alleles among species maintained over longer evolutionary time than those observed at neutral loci (Klein *et al.* 1993). In support of the latter observation, MHC allelic lineages of some mammals are thought to be millions of years old and allele divergences often pre-date species divergences (Figueroa *et al.* 1988). As a result, alleles from different species may be more closely related than alleles within a species (Nei & Rooney 2005). MHC trans-specific diversity has been demonstrated in many natural systems, including fish (Ottova *et al.* 2005), rodents (Figueroa *et al.* 1988; Edwards *et al.* 1997; Bryja *et al.* 2006; Cutrera & Lacey 2007; Kundu & Faulkes 2007), ungulates (Van den Bussche *et al.* 1999; Hedrick *et al.* 2000b), carnivores (Hedrick *et al.* 2000a; Seddon & Ellegren 2002) and primates (Otting *et al.* 2002). The persistence of highly divergent alleles over time may be explained by the hypothesis that increased diversity confers a fitness advantage to the host with an ability to recognize a broader spectrum of pathogens (Wakeland *et al.* 1990).

The extent to which selection is responsible for the observed mode of MHC evolution requires an in-depth look at patterns of variation occurring across the gene. The nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$) has been widely used as a measure of selective pressure on a gene (*reviewed in* Yang & Bielawski 2000).

Whereas ratios larger than one indicate a fitness advantage for mutations resulting in an amino acid change (i.e. positive selection), ratios smaller than one suggest selection against deleterious mutations (i.e. purifying selection). Within a MHC molecule only a limited proportion of amino acids have been found to be involved in antigen recognition and binding (Hughes & Nei 1988) and, thus, d_N/d_S estimates averaged across the gene can be misleading. Site-specific selection analyses have proven to be useful for elucidating how rates of evolution can vary across a gene region and for pin-pointing particular sites under selection (Yang & Bielawski 2000; Yang *et al.* 2000), such as those that specifically interact and recognize foreign peptides. Site-specific methods have found elevated d_N/d_S ratios at these antigen binding sites (ABS), suggesting substantially differing rates of evolution across the MHC (Hughes & Hughes 1995).

MHC genes of the family Equidae, also called the Equine Lymphocyte Antigen (ELA), are similar in organization to those of humans, with adjacent class I, II and III regions (Gustafson *et al.* 2003), and their structure and overall function are believed to be conserved (Madden 1995). Despite these similarities, the evolution of the ELA has been shown to differ in some ways from other species. For example, the most striking observation is that the ELA comprises at least two homologues of the class II *DQA* locus distributed on two different chromosomes, a phenomenon which has never been observed in any other mammalian species (Fraser & Bailey 1998). In situ hybridization studies have localized the ELA to chromosome 20q14-q22 (Ansari *et al.* 1988; Mäkinen *et al.* 1989), except for the single *DQA* homologue which was localized to chromosome 5 (Bailey *et al.* 2000). Further examination of differences in the ELA revealed that the class II *DRA* locus, exon 2, has greater allelic variation in Equidae than in most other taxa (Bailey 1994; Brown *et al.* 2004). For example, in the domestic horse (*Equus callabus*), ass (*E. asinus*), mountain zebra (*E. zebra*) and plains zebra (*E. quagga,* formerly *E. burchelli*), 5-6 alleles per species have been detected (Albright-Fraser *et al.* 1996; Brown *et al.* 2004; Luis *et al.* 2005; Janova *et al.* 2009), in contrast to the majority of species which have little to no sequence variation at this locus (e.g. Yuhki & O'Brien 1997; Takada *et al.* 1998; Chardon *et al.* 1999; Wagner *et al.* 1999). The *DRA* and *DQA* loci are known to be paralogous, encoding the α-chain of a MHC class II molecule, and have a similar function in presenting peptides derived from extracellular proteins. However, the considerable difference in levels of diversity between these genes remains unexplained. Several studies have described the variability of ELA-*DRA* and *DQA* loci of the MHC (Bailey 1994; Fraser & Bailey 1998; Brown *et al.* 2004), but there is still little understanding of the functional significance of these observed differences in ELA genes and how selection may be acting at the molecular level (*but see* Janova *et al.* 2009).

In this study, I investigated the molecular evolution of two MHC class II genes, ELA-*DRA* and *DQA*, within the genus *Equus.* This study combined previously discovered allelic data (Fraser & Bailey 1998; Brown *et al.* 2004; Janova *et al.* 2009) with new genetic data collected from natural populations of plains zebra (*E. quagga/ E. burchelli).* My objectives were to: (1) characterize inter-specific genetic variation, (2) elucidate evolutionary relationships among alleles and (3) detect molecular-level patterns of selection at these loci. I hypothesize that these genes are highly variable and under balancing selection, and that positive selection is occurring at specific functional codon sites in equids. A better understanding of the variability and evolution of ELA genes will

provide valuable background for future studies that aim to examine the genetic basis of susceptibility or resistance to pathogens in both domestic and wild equids.

Methods

Sample collection and DNA isolation

Fecal, blood and tissue samples were collected from plains zebra (*E. quagga/ E. burchelli*) in two parks of southern Africa: Etosha National Park, Namibia (n = 38) and Kruger National Park, South Africa (n = 33). For the purposes of consistency with historical ELA allele nomenclature, I hereafter refer to the species by its former scientific name, *E. burchelli*. With fecal samples, three to five pellets were collected from each individual and allowed to dry. Epithelial cells from the outermost mucosal layer were scraped from the desiccated pellets using a sterile razor blade. Tissue samples were preserved in DMSO/EDTA/Tris/salt solution and blood samples in ethylenediaminetetraacetic acid (EDTA). All samples were stored at -20˚C until DNA extraction. Sample collection was approved by the Animal Care and Use Committee (Protocol #R217-0510B) at UC Berkeley.

Whole genomic DNA was extracted from blood and tissue using Qiagen kits (Valencia, CA). Non-invasive samples, collected from feces, are subject to contamination, enzyme degradation (e.g. Linn 1981), and hydrolytic and oxidative damage that may result in lower DNA yield and increased error rates─ most commonly allele dropout (Taberlet *et al.* 1999). Thus, I used the AquaGenomics protocol (MultiTarget Pharmaceuticals, Inc.) optimized for fecal DNA extraction. A few fecal samples suffered degradation which resulted in failed PCR-amplifications. These degraded samples were re-extracted using the QIAmp fecal extraction kit (Qiagen), also designed specifically for fecal DNA extraction.

PCR-amplification and sequencing

I targeted two *MHC* loci of the Equine Lymphocyte Antigen (ELA) system, ELA-*DRA* and *DQA*, by polymerase chain reaction (PCR) (Saiki *et al.* 1988) and genotyped these loci through direct sequence-based typing. I amplified 246 bp of the *DRA* using equid-specific primers, Be3 and Be4 (Albright-Fraser *et al.* 1996), and 205 bp of the *DQA* using the primers DQA-2e and DQA-2f (Fraser & Bailey 1998). These primers targeted the functionally significant exon 2 of both genes, a region consisting of antigen binding sites (ABS) as predicted by their human lymphocyte antigen (HLA) equivalent (Reche & Reinherz 2003). PCR mixes (total reaction volume of 15µL) for both genes contained approximately 25-50ng DNA, 2uL GeneAmp 10x PCR buffer (100mM Tris-Cl, pH 8.3, 500mM KCl, 15mM $MgCl_2$, 0.01% (w/v) gelatin), 1U Ampli*Taq* Gold DNA polymerase (Applied Biosystems), 0.4 mM dNTPs, 15µg bovine serum albumin (New England BioLabs) and 0.50 µM of each primer.

Amplification of the *DRA* locus used the following "touch-down" thermocycling profile: an initial denaturation at 95ºC for 10 min; 2 cycles of 94ºC for 1 min, 60ºC for 1 min, and 70ºC for 35 s; 18 cycles of 93ºC for 45 s, 59ºC for 45 s, and 70ºC for 45 s, with the annealing temperature decreasing by 0.5ºC with each cycle; 35 cycles of 92ºC for 30 s, 50ºC for 30 s, and 70ºC for 1 min; final extension at 72ºC for 10 min to allow for

complete amplification of the targeted gene. PCR-amplification of the *DQA* locus used the following thermocycling profile: an initial denaturation at 95ºC for 6 min; 40 cycles of 94ºC for 45 s, 56ºC for 45 s, and 72ºC for 1 min; final extension at 72ºC for 5 min.

DRA amplicons were purified prior to sequencing by incubating with Exonuclease I and Shrimp Alkaline Phosphatase at 37°C for 30 minutes. Purified products were cycle-sequenced in both forward and reverse directions using the Big Dye® Terminator v.3.1 kit and run on an ABI 3730 automated sequencer (Applied Biosystems).

Identification of MHC alleles

Sequence chromatograms were edited and aligned using the software Geneious 4.7 (Drummond *et al.* 2010). Allelic phase for *DRA* heterozygous sequences was determined by computational inference with the haplotype reconstruction program PHASE v.2.1 (Stephens *et al.* 2001). This program has been found to be accurate in determining allelic phase even in extremely variable loci, such as the MHC (Bos *et al.* 2007) and, therefore, is considered to be a reliable method for allele identification. I conducted five runs, using different initial random seed values, and compared phase results across runs. A threshold posterior probability of 0.9, a value considered significantly higher than the standard (*see* Harrigan *et al.* 2008), was used to assess the accuracy of the allelic phase determination. Individuals not meeting this threshold were dropped from use in further analyses.

Given the large number of heterozygous sites in the *DQA* locus and previous evidence for multiple loci (Fraser & Bailey 1998), all PCR-amplicons were cloned and sequenced to identify novel haplotypes. PCR products were extracted and purified with the QIAquick Gel Extraction Kit, (Valencia, CA) and cloning was performed using a TOPO-TA® cloning kit with Mach 1^{TM}-T1R competent cells (Invitrogen). Amplicons were ligated into pCR®4 TOPO vectors and transformed into *E. coli* competent cells. Sixteen to twenty-three positive clones per individual were picked with a sterile toothpick and screened by sequencing (protocol described above). The high number of PCR-amplified clones was sufficient to avoid errors, such as recombinant sequences generated during PCR (Bradley & Hillis 1997). Each allele was confirmed with at least two observations, meaning that it had to be found in at least one homozygous individual or two heterozygous individuals to be included in the following analyses.

Sequence data and alignments

Novel *MHC* alleles identified in *E. burchelli* were compiled with a reference panel of Equidae sequences (GenBank, NCBI), including horse (*E. callabus*), ass (*E. asinus*), onager (*E. hemionus*), kiang (*E. kiang*), plains zebra (*E. burchelli*), mountain zebra (*E. zebra*), Grevy's zebra (*E. grevyi*) and Przewalski's horse (*E. przewalski*). A list of ELA-*DRA* and *DQA* sequences from each equid species and their respective GenBank accession numbers are listed in Table S1 and Table S2, respectively. As the ELA-*MHC* nomenclature is currently in revision, names for previously discovered alleles follow designations given in Janova *et al*. (2009) and novel sequences discovered here were named based on the recommendations outlined by the MHC allele nomenclature committee (Ellis *et al.* 2006). The new nomenclature is expected to be established soon on the IPD-MHC Database (www.ebi.ac.uk/ipd/mhc). Identical alleles shared between species were given species-specific numbering. Reference and novel nucleotide, and corresponding amino acid sequences were aligned using the Geneious 4.7 sequence

alignment tool and editor (Drummond *et al.* 2010) and alignment graphics created using BioEdit v5.0.9(Hall 1999).

Statistical analyses of diversity and evolution

Standard descriptive diversity indices for each locus within the genus Equidae were calculated using MEGA4 (Tamura *et al.* 2007). These indices included the number of alleles (*A*), variable nucleotide positions (VNP), parsimony informative positions (PIP), transition/transversion bias ratio (*R*), Kimura 2-parameter gamma (K2P+Γ) evolutionary distance (*d*) and Poisson-corrected amino acid distance. The K2P+Γ model accounts for multiple hits, differences in transitional and transversional substitution rates and variation in substitution rates among sites following a gamma-shaped distribution. Estimates of the gamma shape parameter (α) were determined in PAUP*v4.0b0 (Swofford 2002) to be α = 0.9872 for the *DRA* data and α = 0.4181 for the *DQA* data. Standard error of distance estimates were obtained by using a bootstrap procedure with 10,000 pseudoreplicates.

Four different methods, implemented in RDP v.3.44 beta package (Martin *et al.* 2010), were used to test for recombination and detect potential recombinant events: (1) RDP, (2) GENECONV, (3) Maximum Chi, and (4) BootScan. The highest acceptable *p*-value for all methods was set at a conservative value of 0.10, with a Bonferroni correction for multiple comparisons and a window size of 30 variable nucleotides for all approaches except BootScan. For analyses in BootScan, 1,000 bootstrap replicates were conducted under the Kimura model (transition/ transversion ratio = 1.341), with a window size of 100 bp, step size of 20 nucleotides and cut-off value of 0.70.

Selection, averaged across the gene, was estimated using MEGA4 (Tamura *et al.* 2007) in terms of the relative rates of nonsynonymous (d_N) and synonymous (d_S) base pair substitutions, according to Nei and Gojobori (1986) with the Jukes and Cantor correction for multiple hits (Jukes & Cantor 1969). *Z*-tests of selection were performed over all sites, and separately at ABS and non-ABS, under the null hypothesis of neutrality ($d_N = d_S$) and the alternative hypotheses of non-neutrality ($d_N \neq d_S$), positive selection ($d_N > d_S$), and purifying selection ($d_N < d_S$).

Site-specific selection analyses

As selection will realistically act on only a small subset of amino acids in a protein, averaging substitution rates over entire gene regions is considered to be a conservative indicator of positive selection (Yang & Bielawski 2000). Therefore, I used a more powerful maximum-likelihood based method, implemented in the CodeML subroutine of the software PAML (Yang 2007) which allows the rates of $\omega = d_N/d_S$ to vary among codons (Yang & Bielawski 2000; Bielawski & Yang 2003). This method has been suggested to be more sensitive than other methods for detection of molecular evidence of selection (Anisimova 2003). The models employed here, called 'random-sites' models, do not require *a priori* information on the functional significance of each site and estimate the nonsynonymous to synonymous rate ratio (ω) to indicate selective pressure at the protein level ($\omega < 1$: purifying selection, $\omega = 1$: neutral evolution, $\omega > 1$: positive selection). In this analysis, I used the Equidae alignments to assess heterogeneity in ω across the two MHC genes (*DRA* and *DQA*) and to identify codons under positive selection. I fit the alignment to the following codon 'random-sites' models, in PAML: M0 (one ratio: best average ω across all sites), M1a (nearly neutral: estimates the

proportion of sites that best fit $\omega = 0$ versus those best fit by $\omega = 1$), M2a (positive selection: adds a third set of sites to M1a that have $\omega > 1$ and estimates the best fit for this added ω value and associated proportion of sites), M3 (discrete: fits proportions and ω values assuming three classes of sites labeled 0, 1, and 2 such that $\omega_0 < \omega_1 \leq \omega_2$), M7 (beta: ω is beta-distributed on [0, 1]) and M8 (beta and omega: a proportion of sites are beta-distributed on [0, 1] and the remaining proportion have an average $\omega_2 > 1$ (Yang *et al.* 2000)). M0 is the only model that does not allow for variation in ω across codon sites. Whereas M1a and M7 allow only for neutral evolution and purifying selection at some proportion of sites, M2a, M3, and M8 also allow for the possibility of positive selection at a proportion of sites.

Likelihood ratio tests (LRT) were used to compare nested models based on their log-likelihood (Nielsen & Yang 1998). I compared M0 and M3 to test for the significance of heterogeneity in ω across sites, whereas M1a was compared with M2a, and M7 with M8 to test for positive selection. Significant adaptive evolution was inferred if twice the difference in log-likelihood values was greater than the chi-square critical value for the given degrees of freedom. I used the Bayes empirical Bayes (BEB) approach (Yang *et al.* 2005) to estimate mean ω and standard errors across codon positions. Specific sites under positive selection were indicated by estimates of $\omega > 1$ and posterior probabilities > 0.95. This approach accounts for sampling errors in the maximum likelihood estimates of the parameters and has a low false positive rate. Tree files used in PAML analyses were generated using a maximum likelihood approach in PhyML (Guindon & Gascuel 2003), under the Kimura 3-parameter and the Kimura 2-paramter model of nucleotide substitution for the *DRA* and *DQA* locus, respectively. Models of nucleotide substitution and the distribution of rate variation across nucleotide sites (gamma) were estimated in PAUP*v4.0b0 (Swofford 2002).

Phylogenetic reconstructions

Phylogenetic relationships among Equidae *DRA* and *DQA* sequences were reconstructed using a Bayesian approach implemented in MrBayes 3.1 (Ronquist & Huelsenbeck 2003). The data set was partitioned and the best-fit models were determined for each codon position using the Akaike Information Criterion (AIC) in MODELTEST v.3.7 (Posada & Crandall 1998). Bayesian inference involved running six Metropolis-coupled MCMC chains (1 cold and 5 heated) simultaneously at *n* incremental temperature of 0.1, and chains were run for seven and sixteen million generations for the *DRA* and *DQA* data, respectively. Trees were sampled every 100 generations and the first 25% of trees found were discarded, leaving the remaining trees to be used for estimating the consensus tree. Two independent analyses were conducted and results were compared to check for convergence by confirming that the average deviation of split frequencies approached 0 (with values less than 0.01). I also checked that the potential scale reduction factor (PSRF) approached 1 and that chains mixed sufficiently (with chain mixing values greater than 0.2 between chain pairs). Finally, I used the program Tracer v1.4 (Rambaut & Drummond 2007) to ensure whether sampling from the posterior distribution of each parameter was sufficient and had reached a large enough effective sample size (ESS > 200) for accurate parameter estimation. Posterior probabilities, representing the probability that a specific node is observed, were recorded. This analysis

was run on both non-partitioned and partitioned data, and the optimal model was determined using Bayes Factors.

*DRA s*equences from *Bos taurus* (DQ821713)*, Ovis aries* (Z11600*)* and *Sus scrofa* (AY754888*)* obtained from GenBank (NCBI) were used as outgroups. For *DQA* trees, available sequences from *B. taurus* (AB548942), *O. aries* (M33304) and *S. scrofa* (EU195146) were used as outgroups.

Results

Alleles amplified from the *DRA*, exon 2, in *E. burchelli* represented a single locus. Overall, I found 9 unique *DRA* alleles with haplotype phase certainties greater than the threshold probability value of 90% and which were observed at least twice in the sample. Of the alleles observed, five were novel sequences (*DRA*07-*11)* never seen before in plains zebra [GenBank: HQ637392- HQ637396]. Two of these newly discovered alleles have previously been found in other equid species (*Eqbu-DRA*07* is identical to *Eqas-DRA*01* of *E. asinus*; *Eqbu-DRA*08* identical to *Eqca-DRA*04* of *E. callabus*).

In the *E. burchelli* sample, 21 unique *DQA* alleles were found through cloning which met the requirements for this study. I found 13 novel alleles in *E. burchelli, Eqbu-DQA*09-*21* [GenBank: HQ637397- HQ637409]. One of these, *Eqbu-DQA*09*, is identical to the *E. callabus* allele, *Eqca-DQA*07*. Cloning of the *DQA* revealed between 1- 4 different alleles in each individual, indicating the presence of at least two *DQA* homologous loci.

Inter-and intra-specific analyses of diversity

Nucleotide alignments of all *DQA* sequences from *Equus* revealed considerable sequence diversity at this locus within the genus and at the species level (Figure S1). This observation is consistent with the extreme level of polymorphism typically found at MHC genes [6]. In contrast, *DRA* alignments showed notably lower levels of nucleotide variation (Figure S2). However, it should be noted that the nucleotide and amino acid diversity observed at the *DRA* in Equidae is unusually high relative to what has been reported at this locus in other taxa (Table 1).

Both within *E. burchelli* and among Equidae, genetic diversity (including number of variable sites, number of parsimony informative sites, number of alleles, nucleotide diversity) was greater at the *DQA* than *DRA*. (Table 2 and 3). Mean evolutionary divergence was low (1.3%) across all *DRA* sequences (Table 2), ranging from 0- 3.5% in all pairwise sequence comparisons. In contrast, mean divergence was higher at the *DQA* (13.7%) and ranged from 0- 52.1% between sequence pairs. Interestingly, amino acid distances were greater than evolutionary distances between pairs of nucleotide sequences at both loci. Within other *Equus* species, mean evolutionary distances showed a similar pattern with the exception of *E. asinus* and *E. hemionus*, where average sequence divergences at the *DQA* locus were low (1.5% and 2.7%, respectively), however sample sizes from both species were also low (Table 3).

Global selection analyses

The *DRA* and *DQA* nucleotide sequence encoded an 81 and 67 amino acid protein sequence, respectively (Figures 1 and 2). Protein sequence alignments, including reference Equidae data, revealed 8 synonymous and 7 nonsynonymous mutations at the *DRA* locus. In contrast, the *DQA* exhibited 60 synonymous and 37 nonsynonymous mutations. *Eqbu-DQA*21* had a stop codon at position 64 (Figure 2) and was excluded from all other analyses with the exception of phylogenetic reconstructions. Along with the cloning results, this observation implied the presense of a duplicate non-functional *DQA* locus.

Analyses of the d_N/d_S ratio averaged across the whole coding region suggested that purifying selection is occurring at the *DRA* ($d_N/d_S = 0.32$) and no selection, or neutral evolution ($d_N/d_S = 0.99$), at the *DQA* (Table 2). By species, evidence for positive selection was only found at the *DQA* within *E. kiang* ($d_N/d_S= 2.36$; Table 3). *Z*-tests performed across all codon sites were not statistically significant ($p > 0.05$), and therefore I could not reject (at the 5% level) the null hypothesis of neutral evolution at both MHC loci (Table 4). In summary, estimates of d_N/d_S suggested it is unlikely that positive selection is acting at the level of the entire gene (with $d_N/d_S \leq 1$).

Site-specific selection analyses

It is unlikely for selection to act uniformly across a gene over evolutionary time, but more probable for it to occur at specific sites based on their functional role. For the *DRA*, Z-tests performed on non-ABS separately were significant ($p = 0.049$) providing weak evidence for purifying selection at these sites, whereas I could not reject the null hypothesis of neutral evolution at the ABS (Table 4). At the *DQA*, *Z*-tests by site type also could not reject the null hypothesis of neutrality ($p > 0.05$). However, for both loci, results from the selection analyses in PAML revealed that the model allowing for variable evolutionary rates across codon sites (M3) provided a better fit to the data than the model of one evolutionary rate across sites (M0). Also, models including positive selection (M2a and M8) had higher log-likelihoods that those excluding positive selection (M1a and M7) (Table 5).

At the *DRA*, both M2a and M8 had equivalent likelihoods and suggested that approximately 10% of sites were possibly under positive selection ($\omega = 3.40$) with the remaining sites under purifying selection ($\omega = 0.04$) (Table 5). Using a LRT, the model of one evolutionary rate across sites (M0) was rejected ($p = 0.006$) for the alternative model predicting variable rates of evolution (M3) across *DRA* codons. However, the models of neutral evolution (M1a, M7) could not be rejected ($p = 0.188$, $p = 0.204$). Posterior means of ω estimated across *DRA* codons under positive selection models predicted four sites (positions 14, 19, 47, 49) that may be under selection ($\omega > 1$), two of which are also putative ABS based on the HLA equivalents (Reche & Reinherz 2003). However, as posterior probabilities for these site predictions were less than 95% and positive selection models (M2a and M8) by which these sites were identified were not significant, the hypothesis that selection is occurring at these specific *DRA* codons requires further investigation.

At the *DQA*, the discrete model (of 3 discrete evolutionary rate classes: M3) had the highest log-likelihood and estimated that approximately 44% of codon sites had ω values greater than one (36% with $\omega = 1.68$; 8% with $\omega = 6.80$) with the remaining 56% of sites

being assigned ω values close to 0 ($\omega = 0.08$) (Table 5). Likelihood ratio tests revealed significant variation in selection across codon sites and positive selection occurring at specific sites ($p< 0.001$) (Table 5). Posterior means of ω across *DQA* codon sites, estimated by models M2a and M8, predicted that 5 codons (positions 2, 43, 53, 57, 67) were under significant positive selection. All of these codons are also known as putative ABS (Figure 3). Furthermore, two *DQA* codons (positions 52, 64) were also predicted to be under selection, although with non-significant posterior probabilities (< 95%).

Recombination analyses

There was no evidence for recombination occurring at either MHC locus, even when using a very conservative cutoff for the highest acceptable *p*-value ($p = 0.10$) and small window sizes. Despite these measures, which are known to increase the potential for detecting false positive recombinant events (Martin *et al.* 2010), no recombination was detected. This supports the conclusion that recombination does not play a major role in the generation of diversity at these loci.

Phylogenetic reconstructions and inter-specific allele sharing

Bayes factors suggested that the most probable evolutionary model for both loci required partitioning of the data by codon position. For the *DRA* 243bp sequence, model selection by codon position (pos) indicated the following AIC-selected nucleotide substitution models: pos 1 = Felsenstein 81 model (Felsenstein 1981) with unequal base frequencies (F81uf), pos 2 = Tamura-Nei model (Tamura & Nei 1993) with invariable rates (TrN+I), pos 3 = transversional model with invariable rates (TVM+I). Whereas, for the *DQA*, the best fit models were: pos 1 = transversional model with gamma-distributed rates (TVM+Γ), pos 2 = symmetrical model (Zharkikh 1994) with gamma-distributed rates (SYM+Γ), pos 3 = Tamura-Nei model (Tamura & Nei 1993) with equal base frequencies (TrNef). Therefore, for both loci, phylogenetic analyses in MrBayes were conducted using the general time reversible (GTR) model (Tavaré 1986) as this model encompasses all aforementioned models. However, equal rates versus gamma-distributed rates across sites were specified for the *DRA* versus *DQA*, respectively.

Bayesian phylogenetic analyses revealed widespread sharing of MHC lineages across equid species (Figures 4 and 5), with results from two trials resulting in nearly identical trees. For both loci, alleles were found distributed throughout the evolutionary tree and not clustered by species, such that alleles from different species appear to be more closely related than alleles from the same species. Also, there were many unresolved nodes, with posterior probabilities < 95%, throughout the tree. The *DRA* tree had only one well supported clade including all equid *DRA* alleles. In contrast, the *DQA* tree exhibited multiple well supported clades (posterior probability > 95%). There was one major clade which formed two distinct clusters, encompassing the majority of equid *DQA* alleles, but also a second smaller, more divergent clade comprised of 6 alleles. This smaller clade included the allele that contains a stop codon, *Eqbu-DQA*21*. Alleles *Eqbu- DQA*18*, *Eqbu-DQA*08* and *Eqca-DQA*13* fell out basal to both clades.

I observed a large number of identical alleles across species (Table 3). Overall, there were 33 and 55 alleles in *DRA* and *DQA*, respectively, when accounting for all unique alleles in each *Equus* species (i.e. allowing for identical alleles across species). Identical allele sharing was more prevalent at the *DRA* locus, with 7 of the 22 unique alleles found

in multiple species, whereas a lower proportion of the unique haplotypes (5 out of 48) were shared by two or more species at the *DQA* locus.

Discussion

The characterization of diversity and selection patterns within MHC genes is imperative for understanding their adaptive significance in host immune function. This study found elevated levels of polymorphism and compelling evidence for selection acting on the class II MHC genes, *DRA* and *DQA,* within the genus *Equus* with the contribution of many novel alleles identified in *E. burchelli*. In particular, the average pair-wise amino acid distance among alleles was observed to be greater than nucleotide-based distances in both loci, reflecting an excess of nonsynonymous mutations relative to synonymous mutations. Although global estimates of d_N/d_S averaged across all codon sites contradict the hypothesis of positive selection at these loci, codon-based evolution models that allowed for heterogeneous selection pressure across codon sites best-fit the data. Furthermore, codon models incorporating positive selection were also significant at the *DQA*. Most notably, site-specific selection analyses at this locus suggested that positive selection is occurring at particular codons associated with foreign antigen binding.

Selection at antigen binding sites

Despite the observation of high levels of functional diversity, whole gene-level selection analyses based on the nonsynonymous/ synonymous substitution rate ratios (d_N/d_S) revealed no evidence for positive selection at either locus in *Equus*. However, it is well known that for many functional proteins d_S is often greater than d_N due to strong functional and structural constraints (i.e. purifying selection). Consequently, selection detection methods that average over entire coding regions can be misleading when selective pressures differ substantially across codons; They are unlikely to find elevated nonsynonymous mutation rates and, therefore, have low power to detect signatures of positive selection (e.g. Akashi 1999; Crandall *et al.* 1999). The codon models implemented in this study, however, allowed for selection to vary across codon sites and did, in fact, suggest that a large proportion of sites were conserved, particularly at the *DRA*. More importantly, as even small, single amino acid changes can have a significant impact on gene function these models proved to be valuable for detecting specific targets of selection.

The primary function of classical MHC molecules is to initiate host immune response through the presentation of foreign and self-peptides to T-cells. Studies have shown incredible diversity and elevated nonsynonymous mutations at the ABS of these genes, which is believed to increase the host's ability to recognize a diverse range of pathogens (Hughes & Nei 1988; Hughes & Nei 1989). This underlies the hypothesis that pathogen-driven selection is a primary mechanism sustaining extreme diversity at the MHC (Doherty & Zinkernagel 1975; Hedrick & Kim 1998). In agreement with this, I found that all five *DQA* codons under significant positive selection were also predicted to be ABS (Figure 3). Of the two sites where weaker statistical support for selection was found, only one of these (positions 52) was not a putative ABS. However, this codon was noted to be proximate to an ABS and may play a potential associative role in peptide recognition.

This finding is significant as it not only supports the hypothesized pathogen-driven mechanism driving the diversity observed at the *DQA*, but also identifies candidate amino acid residues that may play a significant role in equid immune response.

Effects of recombination

Although the maximum likelihood based approach used in this study has proven to be powerful in testing for site heterogeneity in selection and in identifying critical amino acids under positive selection (Anisimova *et al.* 2001; Anisimova *et al.* 2002), the presence of recombination can violate the assumptions of the codon-models. I expect that, even if recombination has occurred during the evolution of these genes, the effects on the outcome of the results would be minimal. Anisimova *et al*. (2003) tested the effect of recombination through simulations and concluded that the likelihood-ratio test (LRT) was robust to the presence of low levels of recombination in a dataset. At higher levels of recombination, however, false positive detection rate could be extremely high (up to 90%). Recombination can be difficult to detect, but I found no evidence for its occurrence when using four different approaches. Moreover, M7 and M8 in CodeML, have been shown to be relatively robust to the influence of recombination on selection estimates (Anisimova *et al.* 2003). As the results from LRTs of all three sets of nested models on the *DQA* were highly significant, including M7 versus M8, I conclude that the conclusions hold up even under the low likelihood of undetected recombination.

Trans-species polymorphisms and balancing selection

Balancing selection is expected to preserve high levels of polymorphisms at MHC loci by retaining alleles during species diversification events (Klein 1980; Klein 1987). The lack of allele clustering by species, in reconstructions of *DQA* and *DRA* phylogenies, suggests that MHC allele divergence pre-dates that of species divergence in Equidae. This pattern contrasts that previously found in equid phylogenies based on neutral genetic markers, including microsatellites (Kruger *et al.* 2005) and mitochondrial DNA (George & Ryder 1986), as well as non-neutral globin gene trees (Oakenfull & Clegg 1998), all of which have shown distinct allele segregation by taxon. The discordance between MHC gene phylogenies and other gene phylogenies has similarly been seen among other vertebrate taxa (e.g. Kundu & Faulkes 2007) and has been attributed to balancing selecting acting on these loci due to their role in foreign peptide recognition. Trans-species polymorphisms were well supported in the Equidae *DQA* phylogeny, providing evidence for balancing selection acting on this locus. However, the *DRA* data revealed only one well supported clade (posterior probability > 95%) and, thus, caution must be used in its interpretation. Specifically, the limited availability of sequence variation at the *DRA* largely affected the ability to predict the phylogenetic relationships among alleles and, thus, further examination of diversity in flanking regions of this locus would be useful for clarifying the mode of evolution occurring at this locus. However, the observations of extensive allele sharing among species, in conjunction with unique levels of *DRA* amino acid diversity in *Equus* relative to other taxa (see further discussion below), is compatible with the hypothesis that selection is acting to promote or maintain diversity at this locus in equids.

MHC gene evolution and evidence for *DQA* duplication

Cloning results suggested at least two *DQA* loci in *E. burchelli*, corroborating a previous study in the domestic horse (Fraser & Bailey 1998). Fraser and Bailey (1998) discovered that the horse allele, *Eqca-DQA*13*, is derived from a *DQA* homologue localized to chromosome 5 separate from the primary MHC cluster on chromosome 20. This represented the first time MHC genes have been found distributed on more than one chromosome (Bailey *et al.* 2000). However, little is known about whether this locus is polymorphic. In the *DQA* phylogeny, the plains zebra allele, *Eqbu-DQA*08,* clustered with this putative duplicate allele basal to the primary clades and, therefore, could be a variant of the duplicate locus. Further study is necessary to determine the functionality and expression of this second *DQA* locus.

Bayesian phylogenies showed at least two *DQA* allele clades (Figure 5); one of which encompasses the majority of all equid *DQA* alleles known, to date. The second smaller clade is more divergent and includes the putative 'pseudogene' allele, *Eqbu-DQA*21*. This allele may be the result of a deleterious mutation that arose relatively recently, as the other alleles in this cluster encode potentially functional alleles (i.e. without stop codons). It is possible the alleles of this clade are derived from a paralogous locus which is gradually becoming dysfunctional through an accumulation of deleterious mutations, as would be expected under the 'birth and death' model which has been a hypothesized mode of evolution for MHC gene families (Nei & Rooney 2005). This model suggests that new genes are created by gene duplication and either are maintained over long periods of time or become non-functional through mutations. However, it is alternatively possible that the *DQA*21* allele has acquired a new, unknown function as the frame-shift mutation present in the allele generated a stop codon present at the very end of the gene, thus only truncating the protein by four amino acid residues (*see* Figure S1 and Figure 2). In addition, *Eqbu-DQA*18* was found to be highly divergent from the other *DQA* alleles in Equidae and could also potentially be an allele derived from a *DQA* homologue.

Unique *DRA* diversity in Equidae

Inter-specific analyses of diversity and divergence in MHC alleles revealed that the *DQA* is considerably more polymorphic than the *DRA* in Equidae, with elevated nonsynonymous substitution rates. This finding is concordant with previous studies in other vertebrate species on *DQA* orthologs (e.g. Bondinas *et al.* 2007; O'Connor *et al.* 2007; Chen *et al.* 2010). However, the nucleotide and functional diversity in the *DRA* locus was shown to be unusually high relative to what has been observed in other taxonomic groups (Table 1), supporting the results of previous equid MHC studies (Bailey 1994; Albright-Fraser *et al.* 1996; Brown *et al.* 2004; Janova *et al.* 2009). This observation is particularly compelling because little to no variation in the *DRA* locus has been found in most vertebrate species, for example in humans (Chu *et al.* 1994), dogs (Wagner *et al.* 1999), cats (Yuhki & O'Brien 1997), goats (Takada *et al.* 1998) and pigs (Chardon *et al.* 1999). Chu *et al.* (1994) found that although very low levels of *DRA* polymorphisms exist in mice, these molecules remain involved with peptide binding and suggested that the *DRA* is under strong functional constraints, such that any mutations would be deleterious to peptide-presenting function. Similarly, it is possible that the reduced *DRA* diversity observed in other taxa may be the result of multiple selective sweeps occurring independently across vertebrate lineages. Alternatively, but not

exclusive of this hypothesis, functional constraints that are present in other taxa may have become relaxed in equids. The significance of this unique level of diversity within the Equidae *DRA* remains unclear, though I hypothesize that this locus plays a vital role in response to a unique suite of pathogens or parasites specific to the genus. MHC diversity has also been suggested to be associated with mate recognition and preferences (or inbreeding avoidance) in some species (Jordan & Bruford 1998; Reusch *et al.* 2001). Therefore, further research is necessary to address the potential role sexual selection and parasite-mediated selection could play in the patterns of diversity at the *DRA*.

Conclusions

Much of the research on the Equidae MHC, to date, has been conducted using samples from captive or domestic individuals (e.g. Hedrick *et al.* 1999; Brown *et al.* 2004; Janova *et al.* 2009). Here, focused sampling from natural populations of plains zebra substantially increased the number of known MHC alleles, nearly doubling and tripling that which has previously been identified in this species at the *DRA* and *DQA*, respectively. Wild equid populations are subject to strong selective pressure by parasites and pathogens (e.g. nematode infections and anthrax in Etosha National Park, Namibia), and therefore further study on these populations would substantially advance current knowledge of immune gene evolution and its role in host fitness under natural conditions. This study also highlights the need for more extensive sampling from wild vertebrates in order to capture the full extent of variation at MHC genes. Elucidating patterns of selective pressure across functional immune genes can be especially informative for identifying candidate disease genes and significant protein residues. However, future research linking these results to gene function and ecology is necessary to better understand the mechanisms underlying adaptation in nature.

Acknowledgments

I would like to thank the Getz Lab, Bowie Lab and two anonymous reviewers for valuable comments on the manuscript. I would also like to specifically acknowledge W.M. Getz, R. Bowie, K. Goodman, M. Hadjistylli and E. Rubidge for discussions concerning the genetic methods and analyses; O. Putzeys, H. Ganz, and members of the Getz Lab for help collecting samples in the field; E. DeFranco and K. Hall for assistance in the laboratory. Finally, I am grateful to South African National Parks, the Namibian Ministry of Environment and Tourism and the Etosha Ecological Institute for supporting the sample collection phase of this project. The funding for this research was provided by the National Institute of Health Ecology and Evolution of Infectious Disease (NIH-EEID) GM083863 to WMG and the Carolyn Meek Memorial Scholarship and National Science Foundation Doctoral Dissertation Improvement Grant (NSF-DDIG) MCINS-20091291 to PLK.

Tables

Table 1. Diversity of the ELA-*DRA*, exon 2, by taxon

Taxonomic group	MHC symbol	No. of nucleotide sequences	No. of protein sequences
Bovine	BoLA	1	1
Canine	DLA	1	1
Human	HLA	1	1
Non-Human Primate	NHP	13	2
Ovine	OLA	3	3
Swine	SLA	4	3
Equine*	ELA	22	10

Information extracted from the international ImMunoGeneTics (IMGT) information system® (http://www.imgt.org) and the Immuno Polymorphism Database - MHC (IPD-MHC) (http://www.ebi.ac.uk/ipd/mhc/).
*Equine data was compiled in this study.

Table 2. Indices of diversity and selection at the ELA-*DRA* and *DQA*

	Length (bp)	*N*	*A*	PIP/VNP	R	K2P distance (%)	AA distance (%)	d_N	d_S	d_N/d_S
DRA	243	33	22	9/15	4.75	1.3(0.4)	1.7 (0.7)	0.008 (0.003)	0.025 (0.011)	0.32
DQA	201	55	48*	70/96	1.34	13.7(2.2)	21.6 (4.4)	0.105 (0.019)	0.106 (0.021)	0.99

Length = number of base pairs (bp); *N* = number of alleles when considering identical alleles across taxa separately; *A* = number of alleles across taxa; PIP = parsimony informative positions; VNP = variable nucleotide positions; *R* = transition/ transversion bias; K2P distance = average mean evolutionary distance determined using the Kimura 2-parameter gamma model (K2P+Γ); AA = average mean Poisson-corrected amino acid distance; d_S = synonymous base pair substitution rate; d_N = non-synonymous base pair substitution rate; Standard errors of estimates are shown in parentheses. *Includes the allele *Eqbu-DQA*21,* with a stop codon, while all other estimates exclude it.

Table 3. ELA-*DRA* and *DQA* diversity and selection within *Equus spp.*

Species	*DRA*			*DQA*		
	d (%)	d_N / d_S	*Shared/ Total*	*d (%)*	d_N / d_S	*Shared/ Total*
E. asinus	1.2 (0.4)	0.21	3/6	1.5 (0.9)	0.72	0/2
E. burchelli	1.2 (0.4)	0.23	6/11	16.9 (2.7)	0.78	4/20
E. callabus	1.4 (0.5)	0.67	1/5	13.3 (2.2)	1.10	3/21
E. grevyi	1.3 (0.7)	0.00	1/2	12.1 (3.4)	0.88	1/2
E. hemionus	1.3 (0.7)	0.81	1/2	2.7 (1.2)	0.53	0/2
E. kiang	1.3 (0.7)	0.81	2/2	8.6 (2.2)	2.36	0/3
E. przewalski	n/a	n/a	n/a	n/a	n/a	1/1
E. zebra	0.8 (0.4)	0.40	4/5	12.4 (2.6)	1.27	3/4

d = Mean evolutionary K2P distance; d_N/d_S = nonsynonymous to synonymous mutation rate ratio; Shared/ Total = number of shared out of total alleles found within species of the genus *Equus*.

Table 4. Selection tests over all sites, antigen binding sites (ABS) and non-antigen binding sites (non-ABS)

Locus		Sites		
		All	*ABS*	*Non-ABS*
DRA	*N*	81	20	61
	d_N/d_S	0.302	n/a*	0.202
	Z; $d_N \neq d_S$	0.143	0.121	0.096
	Z; $d_N > d_S$	1	0.058	1
	Z; $d_N < d_S$	0.074	1	0.049
DQA	*N*	67	18	49
	d_N/d_S	0.990	1.013	0.915
	Z; $d_N \neq d_S$	0.975	0.978	0.715
	Z; $d_N > d_S$	1	0.489	1
	Z; $d_N < d_S$	0.488	1	0.359

N = number of codons; d_N/d_S = synonymous to non-synonymous rate ratio; *Z* test *p*-values for rejecting the null hypothesis of neutrality ($d_N = d_S$) for the alternative hypotheses of non-neutrality ($d_N \neq d_S$), positive selection ($d_N > d_S$), and purifying selection ($d_N < d_S$). ***** There were no synonymous mutations, therefore d_N/d_S is undefined.

Table 5. Parameter estimates, log-likelihood values and predicted sites under selection from codon models

Locus	Model code	*P*	ℓ	Parameter estimates	Sites under positive selection	2Δℓ (*p*-value)
DRA	M0 (one ratio)	1	-481.93	ω= 0.353	None	
	M3 (discrete)	5	-474.76	ω_0= 0.044, p_0= 0.589 ω_1= 0.044, p_1= 0.315 **ω_2= 3.40**, p_2= 0.096	Not analysed	14.34 (p= 0.006)
	M1a (nearly neutral)	1	-476.43	ω_0= 0, p_0= 0.787 ω_1= 1, p_1= 0.213	Not allowed	
	M2a (positive selection)	3	-474.76	ω_0= 0.044, p_0= 0.904 ω_1= 1, p_1= 0 **ω_2= 3.40**, p_2= 0.096	14,19, 47, 49	3.34 (p=0.188)
	M7 (beta)	2	-476.35	p= 0.005, q= 0.020	Not allowed	
	M8 (beta and omega)	4	-474.76	p_0= 0.904 , p_1= 0.096, p= 4.61, q= 99.0, **ω= 3.40**	14,19, 47, 49	3.18 (p= 0.204)
DQA	M0 (one ratio)	1	-1612.03	ω= 0.984	None	
	M3 (discrete)	5	-1513.2	ω_0= 0.078, p_0= 0.556 **ω_1= 1.68**, p_1= 0.364 **ω_2= 6.80**, p_2= 0.080	Not analysed	197.68 (p< 0.001)
	M1a (nearly neutral)	1	-1545.23	ω_0= 0.043, p_0= 0.556 ω_1= 1, p_1= 0.444	Not allowed	
	M2a (positive selection)	3	-1516.86	ω_0= 0.047, p_0= 0.521 ω_1= 1, p_1= 0.389 **ω_2= 4.91**, p_2= 0.090	**2***, **43***, **53***, **57***, **67***	66.8 (p< 0.001
	M7 (beta)	2	-1548.22	p= 0.104, q= 0.119	Not allowed	
	M8 (beta and omega)	4	-1518.57	p_0= 0.909 , p_1= 0.091, p= 0.02, q= 0.02, **ω= 5.15**	**2***, **43***, 52, **53*, 57***, 64, **67***	29.65 (p< 0.001)

P = number of free parameters in the ω distribution; ℓ = log-likelihood; Model parameter estimates include the nonsynonymous to synonymous rate ratio (ω) and proportion of sites (p) under each ω site class. Estimates for ω that are evidence for positive selection are bolded. Sites under selection were predicted using the Bayes Empirical Bayes (BEB) approach: sites inferred to be under positive selection with posterior probabilities >95% are in bold and sites with posterior probabilities of > 99% are indicated by an asterisk (*).

Figures

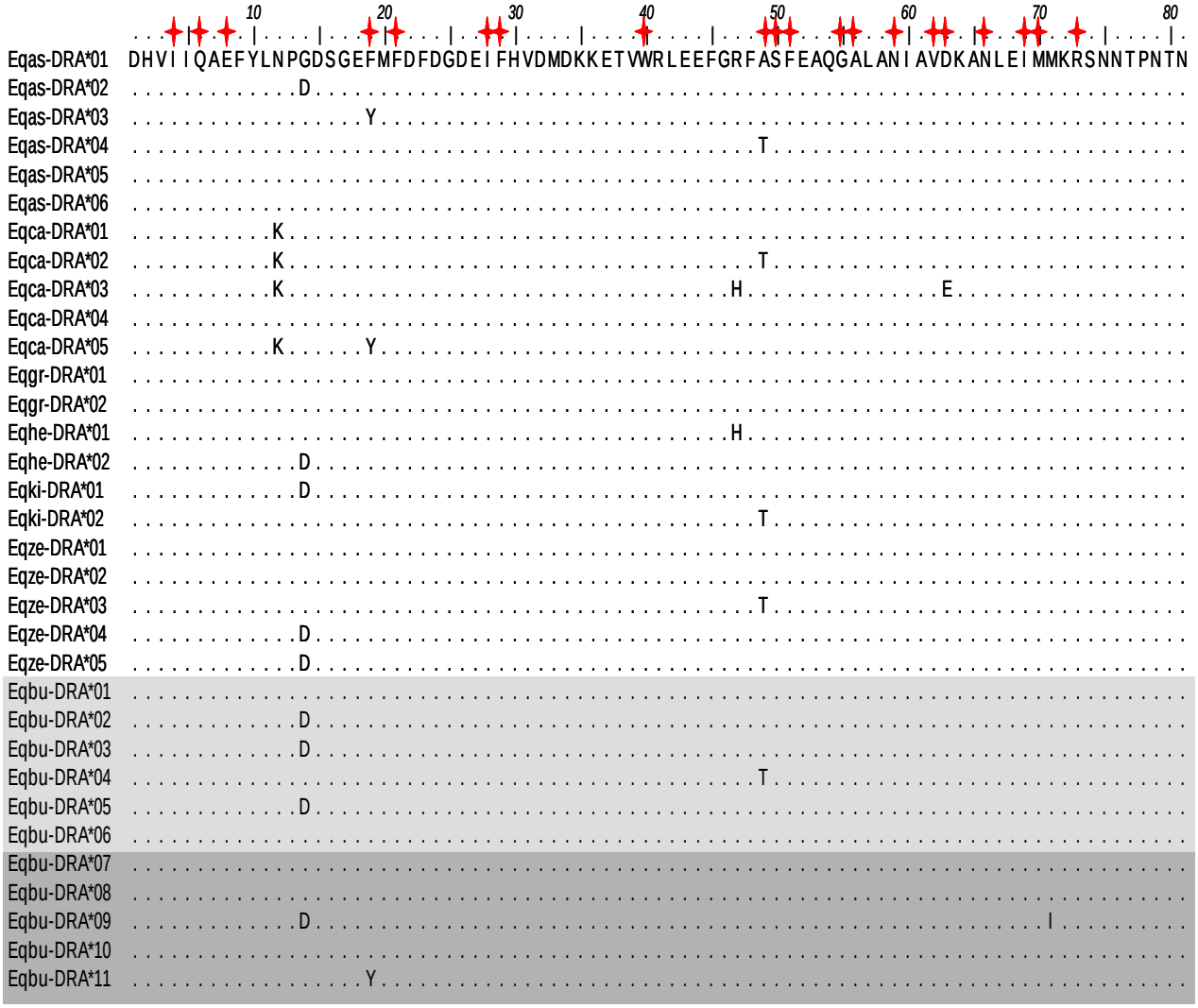

Figure 1. Predicted amino acid alignment of the ELA-*DRA* locus
Dots indicate sequence identity to first sequence in alignment, *Eqas-DRA*01*. *E. burchelli* alleles are shown in gray, with light gray highlighting alleles previously known and dark gray highlighting new alleles discovered in this study. Red stars above amino acids indicate putative antigen binding sites, based on the human HLA equivalents (Reche & Reinherz 2003).

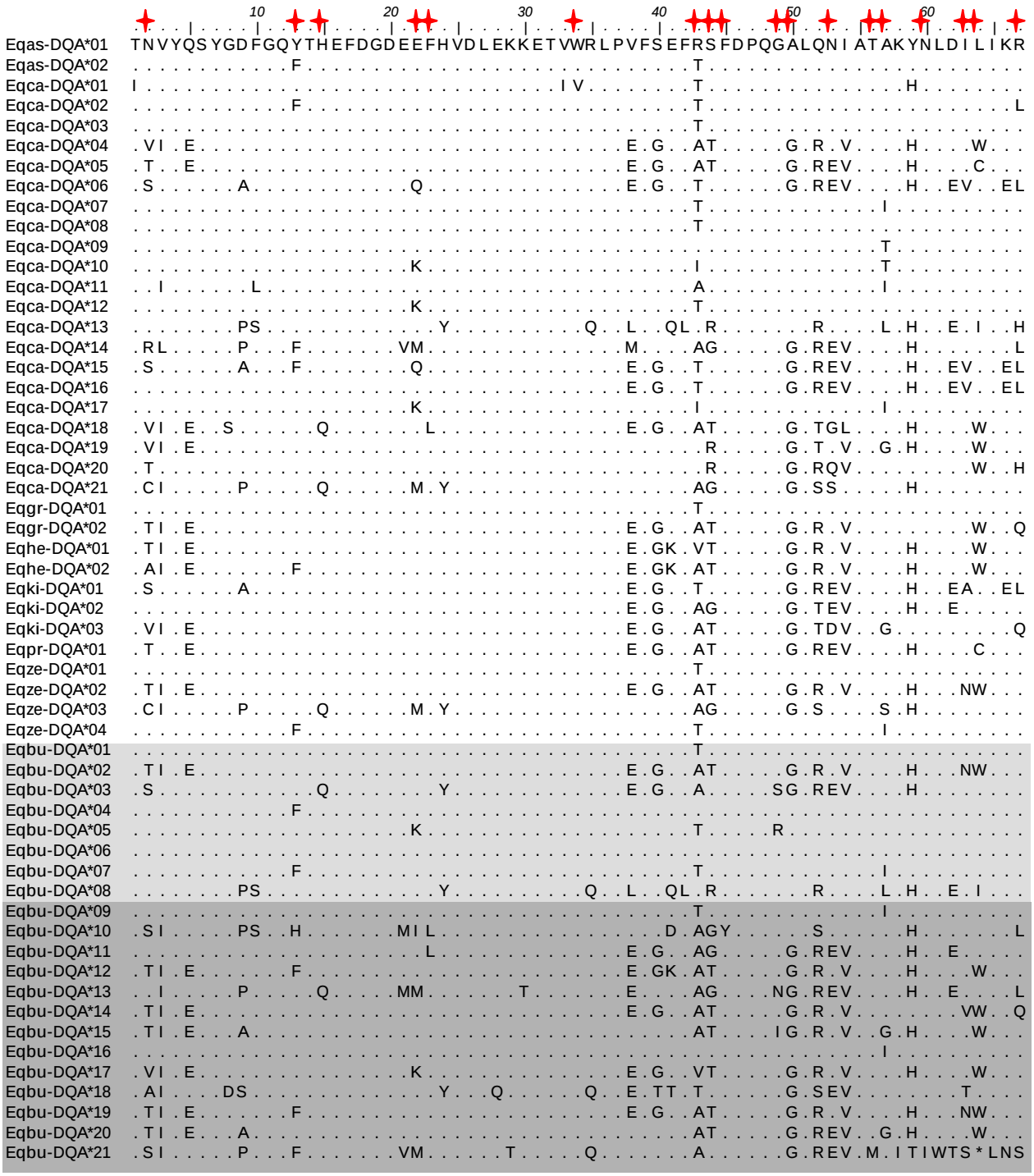

Figure 2. Predicted amino acid alignment of the ELA-*DQA* locus

Dots indicate sequence identity to first sequence in alignment, *Eqas-DQA*01*. *E. burchelli* alleles are shown in gray, with light gray highlighting alleles previously known and dark gray highlighting new alleles discovered in this study. Red stars above amino acids indicate putative antigen binding sites, based on human HLA equivalents (Reche & Reinherz 2003). The asterisk (∗) represents a stop codon.

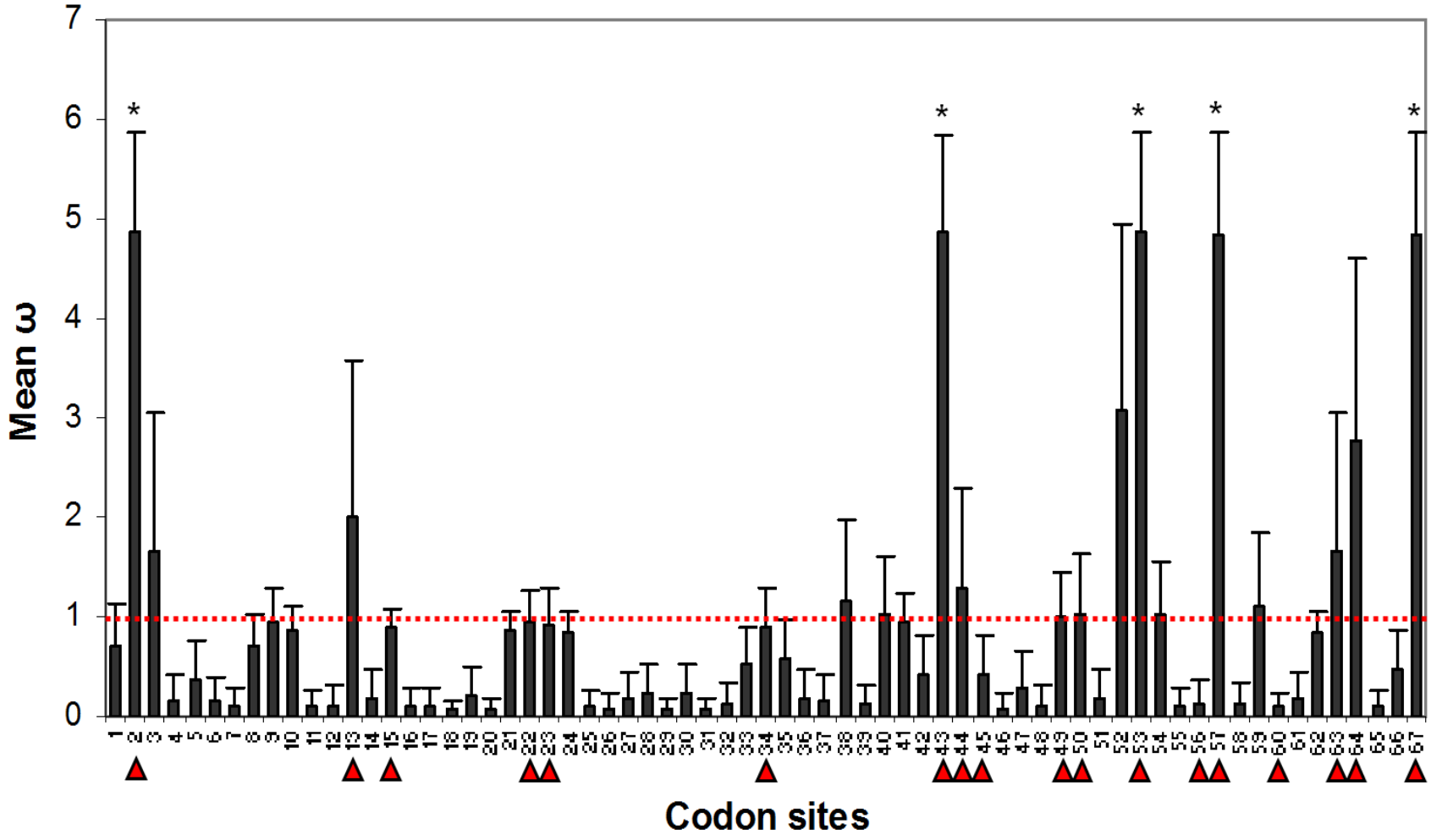

Figure 3. Posterior means of *ω* across *DQA*, exon 2, codon sites

Posterior means of ω calculated over 11 site classes under the random-sites, codon-based model M8 (beta and omega) and Bayes empirical Bayes (BEB) approach as implemented in PAML (Yang 2007). Error bars indicate S.E. of the mean. The asterisk (∗) indicates significant positive selection with a posterior probability > 95%. The dashed red line shows where $\omega = 1$. The red triangle (▲) notates predicted antigen binding sites based on HLA equivalents (Reche & Reinherz 2003).

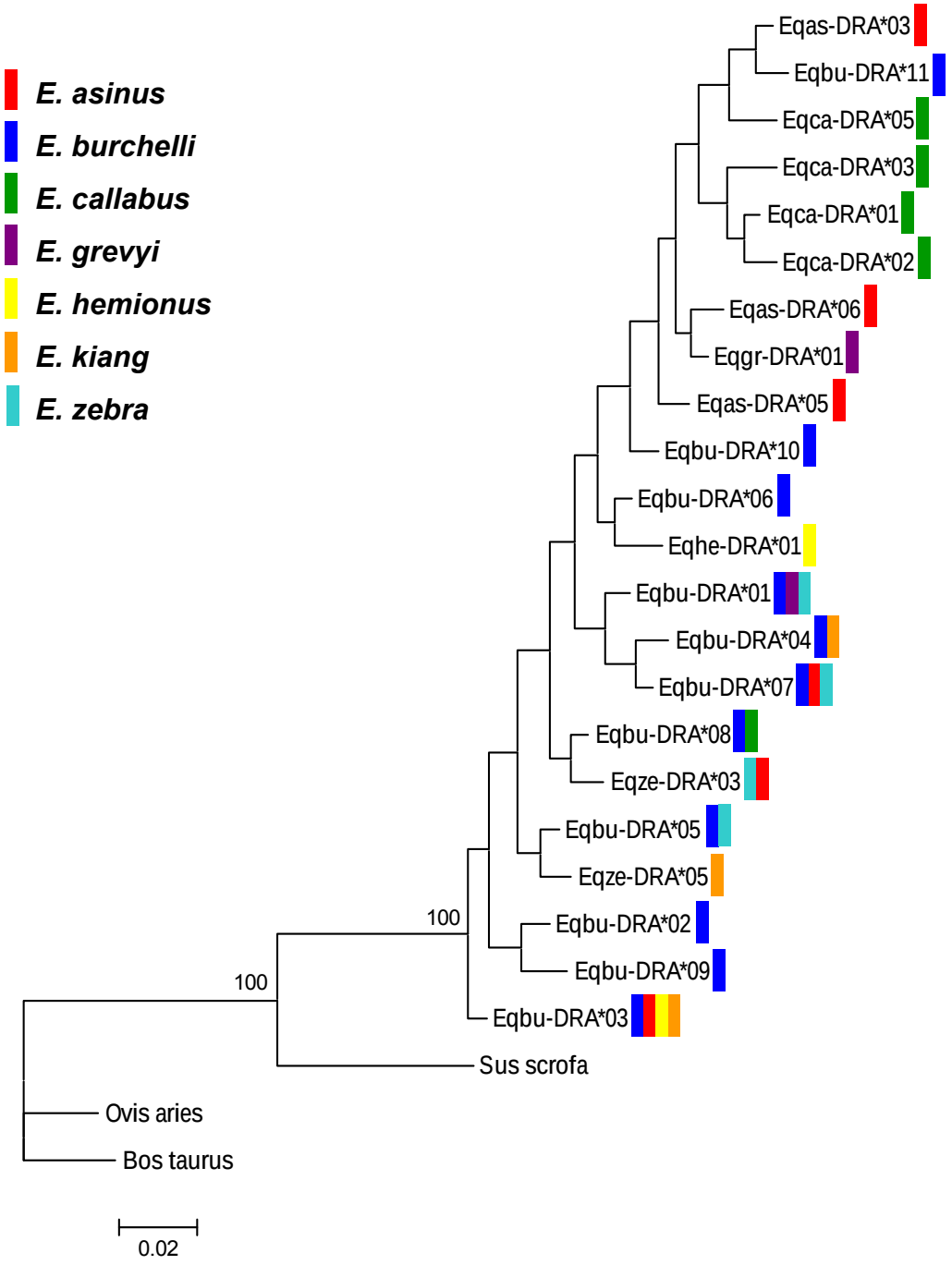

Figure 4. Bayesian reconstruction of unique *DRA* alleles in Equidae
Sequence data (243 bp) was partitioned by codon position and a GTR nucleotide substitution model was used, with equal rates across sites. Posterior probabilities > 50% are reported at the nodes. Identical alleles across multiple species are indicated by the appropriate colored bars (*see* legend) and names were omitted from the tree: ***Eqbu-DRA*01*** = *Eqgr-DRA*02 = Eqze-DRA*02;* ***Eqbu-DRA*04*** = *Eqki-DRA*02;* ***Eqbu-DRA*07*** = *Eqas-DRA*01 = Eqze-DRA*01;* ***Eqbu-DRA*08*** = *Eqca-DRA*04;* ***Eqze-DRA*03*** = *Eqas-DRA*04;* ***Eqbu-DRA*05*** = *Eqze-DRA*04;* ***Eqbu-DRA*03*** = *Eqas-DRA*02 = Eqhe-DRA*02 = Eqki-DRA*01*. Sequences from *Bos taurus* (DQ821713), *Ovis aries* (Z11600) and *Sus scrofa* (AY754888) were used as outgroups.

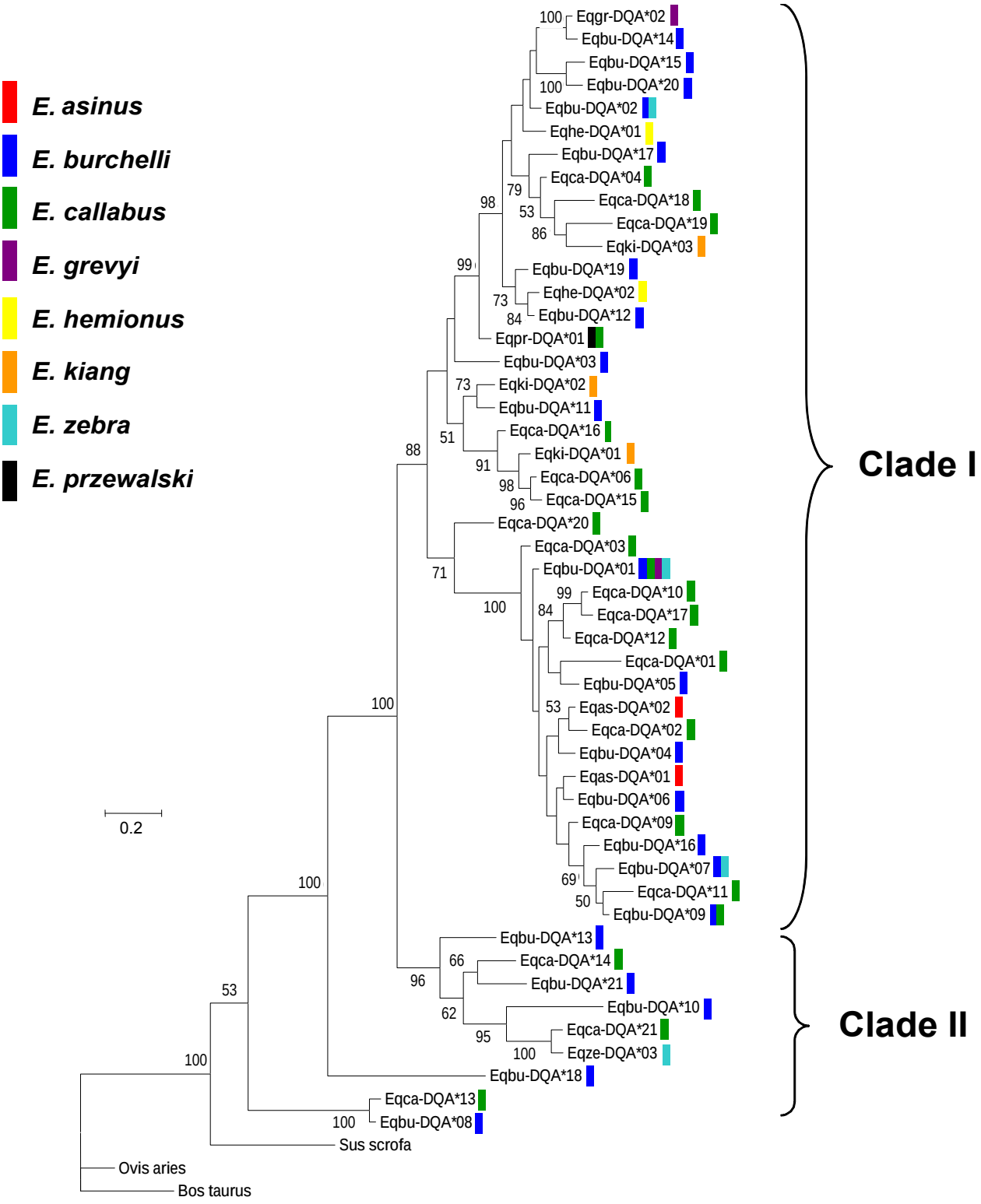

Figure 5. Bayesian reconstruction of unique *DQA* alleles in Equidae

Sequence (205 bp) partitioned by codon, using a GTR substitution model with Γ-distributed rates. Posterior probabilities > 50% are reported at nodes. Identical alleles across species indicated by colored bars (*see* legend) and names omitted from the tree: ***Eqbu-DQA*02*** = *Eqze-DQA*02;* ***Eqpr-DQA*01*** = *Eqca-DQA*05;* ***Eqbu-DQA*01*** = *Eqca-DQA*08 = Eqgr-DQA*01 = Eqze-DQA*01;* ***Eqbu-DQA*07*** = *Eqze-DQA*04;* ***Eqbu-DQA*09*** = *Eqca-DQA*07. Eqbu-DQA*21* has a stop codon, but was included in this analysis. Sequences from *B. taurus* (AB548942), *O. aries* (M33304) and *S. scrofa* (EU195146) were used as outgroups.

Supplementary materials

Table S1. ELA-*DRA* allele sequences. Species, nomenclature and GenBank (NCBI, NIH) accession numbers listed for each allele.

Locus (ELA-)	Species	Nomenclature	Accession Number
DRA	*E. asinus*	*Eqas-DRA*0101*	L47171
DRA	*E. asinus*	*Eqas-DRA*0201*	AF541938
DRA	*E. asinus*	*Eqas-DRA*0301*	AJ575296
DRA	*E. asinus*	*Eqas-DRA*0401*	AJ575297
DRA	*E. asinus*	*Eqas-DRA*0501*	AJ575298
DRA	*E. asinus*	*Eqas-DRA*0601*	FJ487912
DRA	*E. burchelli*	*Eqbu-DRA*0101*	AJ575299
DRA	*E. burchelli*	*Eqbu-DRA*0201*	EU930120
DRA	*E. burchelli*	*Eqbu-DRA*0301*	EU930126
DRA	*E. burchelli*	*Eqbu-DRA*0401*	EU930121
DRA	*E. burchelli*	*Eqbu-DRA*0501*	EU930118
DRA	*E. burchelli*	*Eqbu-DRA*0601*	EU930122
DRA	*E. callabus*	*Eqca-DRA*0101*	L47174
DRA	*E. callabus*	*Eqca-DRA*0201*	M60100
DRA	*E. callabus*	*Eqca-DRA*0301*	L47172
DRA	*E. callabus*	*Eqca-DRA*0401*	AJ575295
DRA	*E. callabus*	*Eqca-DRA*0501*	FJ716134
DRA	*E. grevyi*	*Eqgr-DRA*0101*	EU930116
DRA	*E. grevyi*	*Eqgr-DRA*0201*	EU930125
DRA	*E. hemionus*	*Eqhe-DRA*0101*	L47173
DRA	*E. hemionus*	*Eqhe-DRA*0201*	EU930128
DRA	*E. kiang*	*Eqki-DRA*0101*	EU930127
DRA	*E. kiang*	*Eqki-DRA*0201*	FJ657514
DRA	*E. zebra*	*Eqze-DRA*0101*	EU930123
DRA	*E. zebra*	*Eqze-DRA*0201*	EU930124
DRA	*E. zebra*	*Eqze-DRA*0301*	EU930129
DRA	*E. zebra*	*Eqze-DRA*0401*	EU930117
DRA	*E. zebra*	*Eqze-DRA*0501*	EU930119

Table S2. ELA-*DQA* allele sequences. Species, nomenclature and GenBank (NCBI, NIH) accession numbers listed for each *DQA* allele.

Locus (ELA-)	Species	Nomenclature	Accession Number
DQA	*E. asinus*	*Eqas-DQA*0101*	U92521
DQA	*E. asinus*	*Eqas-DQA*0201*	U92522
DQA	*E. burchelli*	*Eqbu-DRA*0101*	EU935837
DQA	*E. burchelli*	*Eqbu-DRA*0201*	EU935829
DQA	*E. burchelli*	*Eqbu-DRA*0301*	EU935833
DQA	*E. burchelli*	*Eqbu-DRA*0401*	EU935834
DQA	*E. burchelli*	*Eqbu-DRA*0501*	EU935835
DQA	*E. burchelli*	*Eqbu-DRA*0601*	EU935836
DQA	*E. burchelli*	*Eqbu-DRA*0701*	EU935832
DQA	*E. burchelli*	*Eqbu-DRA*0801*	EU930130
DQA	*E. callabus*	*Eqca-DRA*0101*	L33909
DQA	*E. callabus*	*Eqca-DRA*0201*	U92505
DQA	*E. callabus*	*Eqca-DRA*0301*	U92506
DQA	*E. callabus*	*Eqca-DRA*0401*	U92507
DQA	*E. callabus*	*Eqca-DRA*0501*	U92508
DQA	*E. callabus*	*Eqca-DRA*0601*	U92510
DQA	*E. callabus*	*Eqca-DRA*0701*	U92511
DQA	*E. callabus*	*Eqca-DRA*0801*	U92512
DQA	*E. callabus*	*Eqca-DRA*0901*	U92513
DQA	*E. callabus*	*Eqca-DRA*1001*	U92514
DQA	*E. callabus*	*Eqca-DRA*1101*	U92515
DQA	*E. callabus*	*Eqca-DRA*1201*	U92516
DQA	*E. callabus*	*Eqca-DRA*1301*	U92517
DQA	*E. callabus*	*Eqca-DRA*1401*	U92518
DQA	*E. callabus*	*Eqca-DRA*1501*	U92519
DQA	*E. callabus*	*Eqca-DRA*1601*	AF115324
DQA	*E. callabus*	*Eqca-DRA*1701*	AF115325
DQA	*E. callabus*	*Eqca-DRA*1801*	AF115326
DQA	*E. callabus*	*Eqca-DRA*1901*	AF115327
DQA	*E. callabus*	*Eqca-DRA*2001*	AF115328
DQA	*E. callabus*	*Eqca-DRA*2101*	AF115329
DQA	*E. grevyi*	*Eqgr-DQA*0101*	EU930136
DQA	*E. grevyi*	*Eqgr-DQA*0201*	EU930131
DQA	*E. hemionus*	*Eqhe-DQA*0101*	U92520
DQA	*E. hemionus*	*Eqhe-DQA*0201*	EU930135
DQA	*E. kiang*	*Eqki-DQA*0101*	EU930132
DQA	*E. kiang*	*Eqki-DQA*0201*	EU930133
DQA	*E. kiang*	*Eqki-DQA*0301*	EU930134
DQA	*E. przewalski*	*Eqpr-DQA*0101*	U92509
DQA	*E. zebra*	*Eqze-DQA*0101*	EU935838
DQA	*E. zebra*	*Eqze-DQA*0201*	EU935828
DQA	*E. zebra*	*Eqze-DQA*0301*	EU935830
DQA	*E. zebra*	*Eqze-DQA*0401*	EU935831

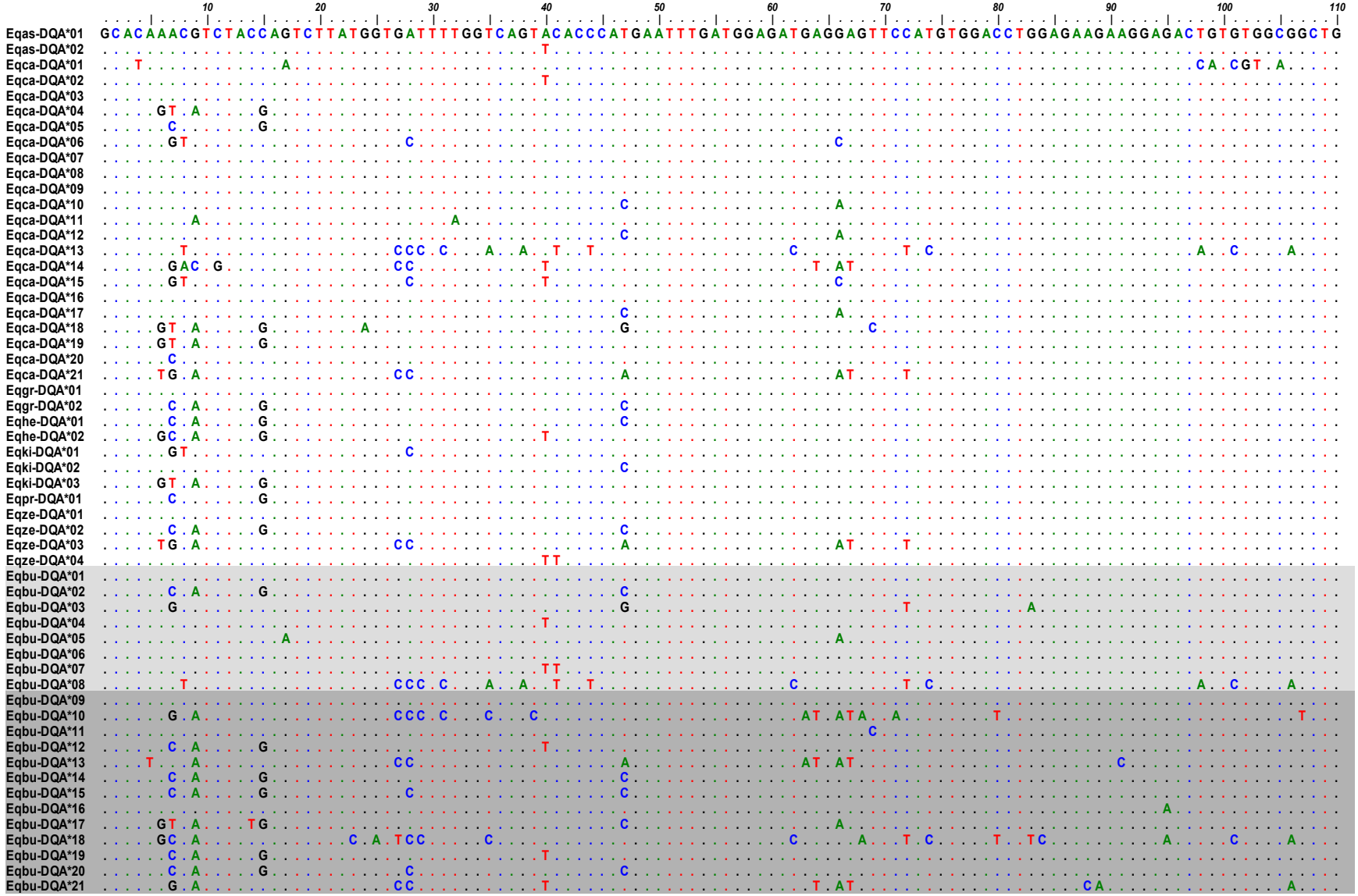

Figure S1a. Nucleotide alignment of ELA-*DQA* alleles identified in Equidae (nucleotide positions #1-110)

Dots indicate identity to first sequence in alignment, *Eqas-DQA*01*. The thirteen novel *E. burchelli* alleles that were identified in this study (*Eqbu-DQA*09 - *21*) are highlighted in dark gray. Previously reported *E. burchelli* alleles are shown in light gray.

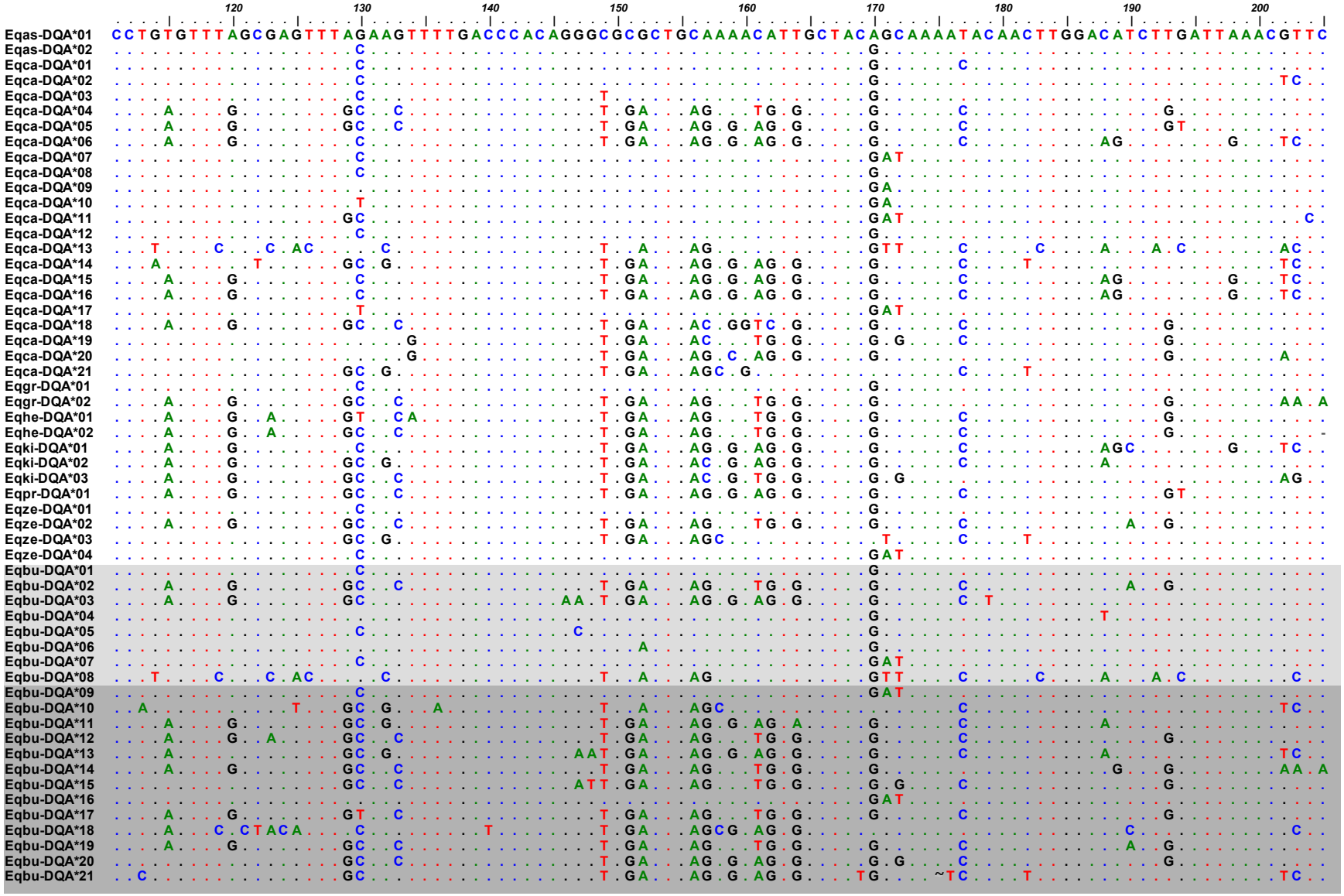

Figure S1b. Nucleotide alignment of ELA-*DQA* alleles identified in Equidae (nucleotide positions #111-205)

Dots indicate identity to first sequence in alignment, *Eqas-DQA*01*. The thirteen novel *E. burchelli* alleles that were identified in this study (*Eqbu-DQA*09 - *21*) are highlighted in dark gray. Previously reported *E. burchelli* alleles are shown in light gray. *Eqbu-DQA*21* has a frame-shift mutation (~) at position 176.

Figure S2a. Nucleotide alignment of ELA-*DRA* alleles identified in Equidae (nucleotide positions #1-130).
Dots indicate identity to first sequence in alignment, *Eqas-DRA*01*. *E. burchelli* alleles are shown in gray. Five novel *E. burchelli* alleles identified in this study (*Eqbu-DRA*07 -*11)* are highlighted in dark gray, whereas alleles discovered in previous studies are in light gray.

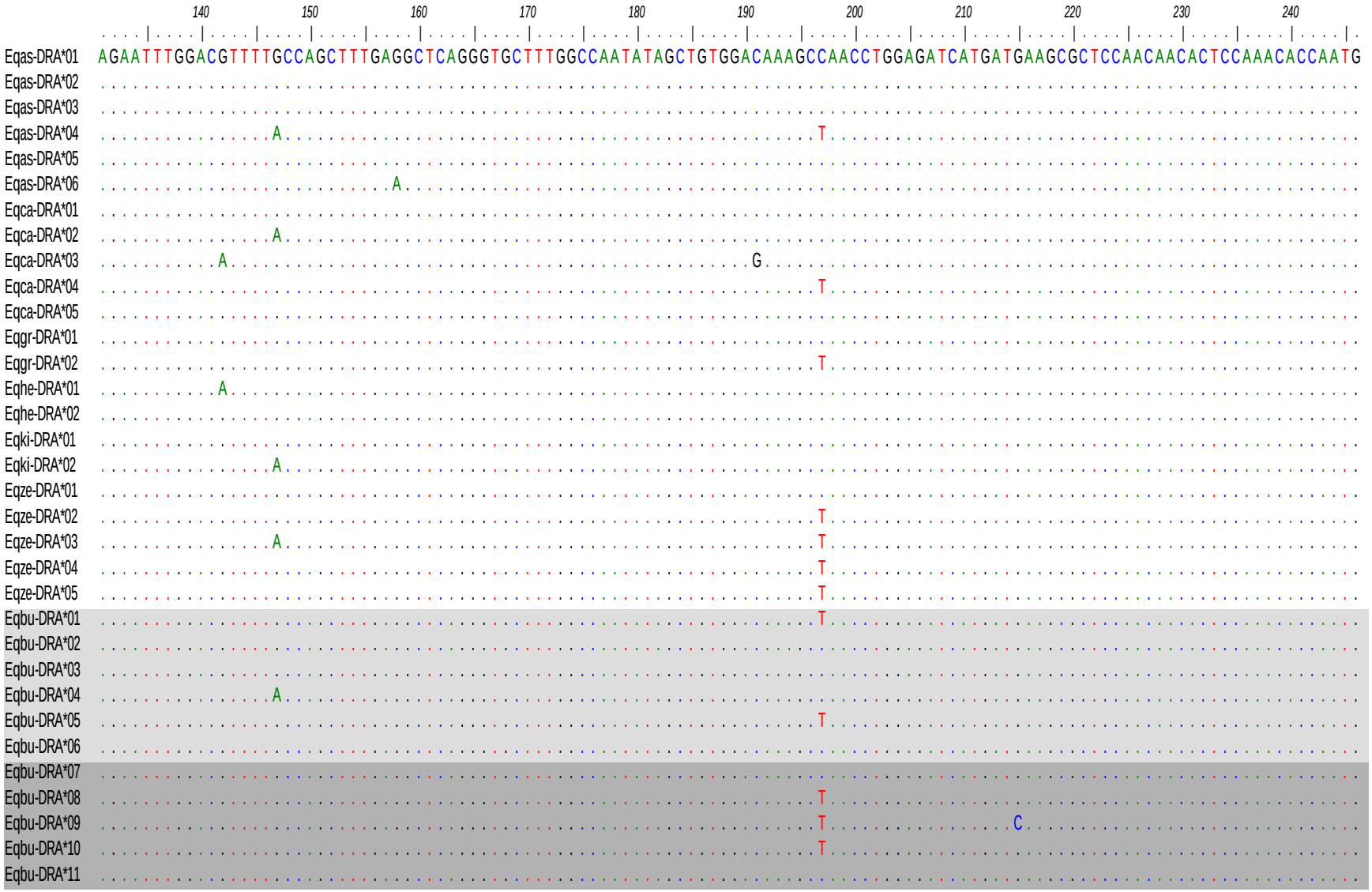

Figure S2b. Nucleotide alignment of ELA-*DRA* alleles identified in Equidae (nucleotide positions #131-246). Dots indicate identity to first sequence in alignment, *Eqas-DRA*01*. *E. burchelli* alleles are shown in gray. The five novel *E. burchelli* alleles identified in this study (*Eqbu-DRA*07 -*11)* are highlighted in dark gray, whereas alleles discovered in previous studies are highlighted in light gray.

Chapter 3. Relative influence of demography and selection on immune gene variation in plains zebra (*Equus quagga*) populations of southern Africa

> "*Living in an entirely different physical as well as biotic environment, such a population would have unique opportunities to enter new niches and to select novel adaptive pathways"*
>
> ─ Ernst Mayr

Abstract

Demography, migration and natural selection are predominant evolutionary forces shaping the distribution of genetic variation among natural populations. Many studies have used neutral genetic markers to make inferences about population history. However, there is a growing recognition that investigating functional coding loci, which directly reflect fitness, is critical to our understanding of species' ecology and evolution. Immune genes, such those of the Major Histocompatibility Complex (MHC), are significant for their role in pathogen recognition and provide an excellent system for studying selection and adaptation. I contrasted diversity patterns of neutral data (13 microsatellite loci and a nuclear intron) with MHC loci, *ELA-DRA* and *DQA*, in two plain zebra (*Equus quagga*) populations of southern Africa: Etosha National Park, Namibia, and Kruger National Park, South Africa. Results from neutrality tests, along with observations of elevated diversity and low differentiation across populations, supported previous genus-level evidence for balancing selection at these loci. Despite being low, MHC population divergence was significant and may be attributed to drift effects typical of geographically separated populations experiencing little to no gene flow. At the *DRA* locus, zebra in Etosha exhibited geographic differentiation concordant with microsatellites and reduced levels of diversity due to highly skewed allele frequencies that could not be explained by demography, thus suggestive of spatially heterogeneous selection and local adaptation. This study highlights the complexity in which selection acts on gene diversity, suggesting a role for both ongoing balancing selection and heterogeneity in selective pressures over populations. It also underscores the need for future studies on the ecological mechanisms (e.g. pathogen diversity) that drive these patterns in adaptive variation among natural populations.

Introduction

Elucidating the mechanisms that shape genetic variation across populations and loci requires teasing apart the impacts of various micro-evolutionary forces that may be at play. The processes of gene flow and genetic drift are known to influence gene diversity, and equivalently affect all loci in an organism's genome (Lewontin & Krakauer 1973; Nielsen 2005). As migration and demographic factors (e.g. changes in population size or structure) can in turn affect these processes, the mutational patterns and allelic distributions of effectively neutral genes should reflect population history. In contrast, selection is expected to act differentially across loci depending on the adaptive significance of a gene region and the magnitude of the selective force. Therefore, contrasting patterns of diversity among populations at neutral and functional loci can be especially informative for illuminating the relative effects of selection versus demography on gene variation and population dynamics.

Examination of neutral genetic markers, such as microsatellites, single nucleotide polymorphisms (SNPs) or mitochondrial DNA (mtDNA), has proven to be valuable for studying the evolution and demographic history of a population or species (Bruford & Wayne 1993; Avise 2004; Lowe *et al.* 2004). Neutral data has been informative for studying relatedness, mating behavior, dispersal patterns, changes in population size, population genetic structure and speciation (see Blouin *et al.* 1996; Paetkau *et al.* 1997; Balloux & Lugon-Moulin 2002; Brumfield *et al.* 2003; Fabiani *et al.* 2003). However, it is adaptive variation that reflects natural selection, fitness and the potential to adapt to changing environments, and is of critical importance to the study of evolution, ecology and conservation (van Tienderen *et al.* 2002; Sommer 2005). In the past decade, there has been a surge in studies combining analyses of neutral and functional genes, such as those involved in immune response, to elucidate how selection shapes gene diversity among natural populations (e.g. Landry & Bernatchez 2001; Miller *et al.* 2001; Cutrera & Lacey 2006; Bos *et al.* 2008; Mona *et al.* 2008; Loiseau *et al.* 2009; Miller *et al.* 2010).

The Major Histocompatibility Complex (MHC) is a gene family that plays a critical role in the immune system of vertebrates. Class II MHC genes encode cell-surface glycoproteins, responsible for directly binding foreign antigens from extracellular pathogens (e.g. eukaryotic parasites and bacteria) and presenting them to T-lymphocytes to elicit an immune response. These genes typically exhibit elevated levels of polymorphism and highly divergent MHC variants have been shown to persist over long time periods, even through the course of speciation events (Klein 1986). This maintenance of MHC polymorphism across species, also known as 'trans-species polymorphisms', can largely be explained by the action of balancing selection (Takahata 1990). Variation in the MHC has been shown to be particularly concentrated within the peptide binding region (PBR) (Hughes & Nei 1989) and suggests that pathogen recognition is a selective driving force, such that increased allelic diversity allows for recognition of a broader spectrum of pathogens (Hedrick & Kim 1998). Beyond implications for disease resistance, MHC variants have also been shown to influence other biological traits such as mate preference, kin recognition and maternal-fetal interactions (reviewed in Edwards & Hedrick 1998; Bernatchez & Landry 2003; Piertney & Oliver 2006). As a well-studied gene region, the MHC remains an important model

with which to test hypotheses regarding the influence of pathogen-mediated selection on host diversity and, ultimately, on fitness and evolutionary potential.

Much debate still surrounds the mechanism by which parasite-driven balancing selection maintains diversity in MHC genes (Sommer 2005). Overdominant selection, or heterozygote advantage, makes the basic assumption that heterozygous individuals should recognize a broader range of pathogens (Doherty & Zinkernagel 1975). Whereas, frequency-dependent selection suggests that the advantage of a particular allele varies with its frequency, with common alleles selected against and rare alleles selected for (Takahata & Nei 1990). In reality, however, natural populations are exposed to fluctuating environmental conditions and, subsequently, host-pathogen interactions are expected to vary over time and space (e.g. Miller *et al.* 2001). This model, presenting allele fitness as spatiotemporally heterogeneous, is referred to as 'diversifying selection in space and time' (Hedrick 2002).

Recent studies have provided evidence for balancing selection acting on MHC genes over the history of the genus *Equus* (Janova *et al.* 2009; Kamath & Getz 2011). The MHC in equids, also called the Equine Lymphocyte Antigen (ELA), has been found to be unique in its organization and diversity. Most notably, the ELA- *DRA* gene possesses an extreme level of diversity relative to other mammals (Albright-Fraser *et al.* 1996; Brown *et al.* 2004). The ELA-*DQA* locus is also exceptionally variable in equids and has a duplicate functional copy present on an entirely different chromosome, an observation that has not been recorded in any other vertebrate taxa (Fraser & Bailey 1998). While there have been some studies in equid MHC evolution (Janova *et al.* 2009; Kamath & Getz 2011), none have characterized selection among wild populations.

Plains zebra (*Equus quagga*) of Etosha National Park, Namibia, are the main host species of anthrax, a deadly disease caused by the bacterial pathogen, *Bacillus anthracis*, which occurs in severe and consistent annual outbreaks (Turnbull *et al.* 1989; Lindeque & Turnbull 1994). In addition, zebra in Etosha are appreciably infected by gastrointestinal (GI) parasites, found in the population at nearly 100% prevalence (Turner & Getz 2010). In zebra of Kruger National Park, South Africa, anthrax outbreaks are sporadic, occurring on an approximate decadal cycle (de Vos 1990) and although GI nematode prevalence in plains zebra is similarly high, there is a greater nematode species richness in Kruger when contrasted to Etosha (Matthee *et al.* 2004). Given the significance of MHC class II genes in pathogen recognition, I predict that they are likely affected by pathogen-mediated selection that varies among natural zebra populations.

In this study, I examined the distribution of genetic variation in the PBR of two class II MHC genes, *DRA* and *DQA*, in plains zebra populations to shed light on how selection is operating at these genes. A baseline for demography, including genetic structure and inferences regarding changes in population size, was established through analyses that combine the use of both neutral microsatellite and nuclear intron data. I employed tests for selection and demography on both functional and neutral data to reveal differences in selective pressures occurring across loci and zebra populations. Such tests may provide valuable insight for inferring the importance of immune genes in local adaptation and for pinpointing candidate alleles that may play a particular ecological role, such as in pathogen resistance.

Methods

Sampling

Blood, tissue and fecal samples were collected from plains zebra (*E. quagga*) from two populations in southern Africa: Etosha National Park, Namibia (n = 84) and Kruger National Park, South Africa (n = 89). Zebra were sampled between June 2007 and November 2008, so these analyses are unlikely to be influenced by temporal variation. Sample preservation and genomic DNA extraction protocols are outlined in Kamath & Getz, 2011. The use of DNA extracted from fecal material occasionally resulted in failed PCR-amplifications, which was attributed to enzyme degradation, hydrolytic and oxidative damage. Therefore, in several cases only a subset of the samples collected were used in downstream genetic analyses (Table S1).

MHC sequencing and cloning

The exon 2 coding regions of the MHC class II loci, ELA-*DRA* and ELA-*DQA*, were amplified from genomic DNA using primers, polymerase chain reaction (PCR) cycling conditions, and sequencing protocols described in Kamath & Getz, 2011. This resulted in 246 bp and 205 bp amplicons encompassing the functional PBR of the *DRA* and *DQA*, respectively. I sequenced loci in both forward and reverse directions to confirm heterozygous base positions. Sequence chromatograms were edited and aligned manually using Geneious 5.0 (Drummond *et al.* 2010). Allelic phase for heterozygous bases was determined with the haplotype reconstruction program, PHASE v2.1 (Stephens *et al.* 2001), known to be accurate in determining phase in even extremely variable MHC loci (Bos *et al.* 2007). I ran two runs, with different initial random seed values, and compared results across runs. I used a threshold posterior probability of 0.80, a value higher than the standard (Harrigan *et al.* 2008), to conclude that phase determination was accurate and only individuals meeting this threshold were used in further analysis.

The *DQA* locus is known to have two duplicate copies with up to four alleles per individual (Fraser & Bailey 1998; Kamath & Getz 2011). Therefore, for heterozygous individuals, molecular cloning was carried out using a TOPO-TA® cloning kit with Mach 1™ T1R competent cells (Invitrogen) to identify alleles. Spurious sequences can form during PCR and cloning due to polymerase errors and the formation of heteroduplexes or chimeras, particularly when amplifying multiple sequences simultaneously (Jansen & Ledley 1990). As this may be especially problematic during cloning, I minimized potential error by using a high fidelity polymerase (Platinum® Taq DNA Polymerase High Fidelity, Invitrogen) in the initial PCR. PCR products were run on a 1% agarose gel and bands of the correct length were excised and purified using QIAquick Gel Extraction Kits (Qiagen), ligated into pCR®4 TOPO vectors and transformed into *E. coli* competent cells. Following an overnight incubation at 37°C, sixteen to twenty-three positive clones were picked for each individual sample and clones were directly sequenced using the primers used in the original PCR. Allele sequences were confirmed with at least two observations (i.e. amplified in at least two independent PCRs or cloning reactions from the same individual or seen in two different individuals). Sequences not meeting these criteria were considered to be erroneous and not considered in subsequent analyses.

β-Fibrinogen intron 7 amplification

I amplified 668 bp of the *Fibrinogen* beta chain gene, intron 7, (*β-Fibr*) with primers that were developed using the publicly-available horse genome and Primer3 (Rozen & Skaletsky 2000): *β-Fibr10* (5'-CAGTAGTATCTGCCGTTTGG-3') and *β-Fibr11* (5'-GAGGGCGACAAATACCAAC- 3'). Description of the PCR protocol used can be found in the Supplementary materials (PCR protocols). Amplified PCR products were cleaned using Exo-SAP-IT (USB Corporation) and sequenced in both directions on an ABI 3730 sequencer (Applied Biosystems) under standard cycling conditions. Due to the presence of a length mutation in the *β-Fibr* intron, I used the program OLFinder to help resolve insertion/deletions (indels) for individuals with evidence of their presence (Dixon 2010). Sequences were edited and alleles determined and confirmed as described above for MHC loci.

Microsatellite genotyping

I genotyped 15 microsatellite loci, previously isolated from the domestic horse (*Equus callabus*): Aht21 (Swinburne *et al.* 1997), Asb23 (Irvin *et al.* 1998), Cor014 (Hopman *et al.* 1999), Hmb1 (Binns *et al.* 1995), Hms7 (Guerin *et al.* 1994), Htg7, Htg9, Htg14, Htg15 (Marklund *et al.* 1994), Lex20, Lex33 (Coogle *et al.* 1996), Lex52 (Coogle & Bailey 1997), Ucdeq505 (Eggleston-Stott *et al.* 1997), Um011 (Meyer *et al.* 1997), Vhl47 (van Haeringen *et al.* 1998). Forward primers were modified at the 5'-end by the addition of a fluorescent label: HEX, 6-FAM (Invitrogen), NED or PET (Applied Biosystems). Details of the PCR protocols used can be found in the Supplementary materials (PCR Protocols & Table S2). Allele fragments were scored for size against the LIZ-500 size standard through electrophoresis using an ABI3730 DNA Analyser, followed by visualization with GeneMapper v.4.0 (Applied Biosystems) software.

Noninvasive genetic sampling is known to result in high genotyping error rates and allelic dropout. A portion of the samples used in this study were derived from feces, and thus subject to these problems. To address this issue, I used a comparative genotyping protocol (Frantz *et al.* 2003) (Figure S1) which is a modification of the commonly implemented multi-tubes genotyping approach (Taberlet *et al.* 1996). This protocol has been shown to efficiently reduce error rates as it minimizes the number of PCRs necessary to arrive at a consensus genotype (see Hansen *et al.* 2008). Furthermore, I quantified genotyping error using paired blood and fecal samples ($n = 42$) obtained from individual zebra in Etosha. Using both blood and feces obtained from an individual, two PCRs per sample were performed and total error rate was quantified. Further breakdown of factors contributing to this error included the error rate due to false alleles, allelic dropout and multiple alleles.

To confirm that microsatellite loci follow the assumptions of neutrality, Fisher's exact tests for Hardy-Weinberg (H-W) equilibrium (testing the alternative hypotheses of heterozygote excess or deficiency) and genotypic linkage disequilibrium (LD) between pairs of loci were conducted in GENEPOP 4.0 (Raymond & Rousset 1995). Significance of exact tests was determined with a Markov chain (MC) algorithm (Guo & Thompson 1992) using default parameters, and corrected for multiple comparisons through a sequential Bonferroni procedure (Holm 1979). Null allele frequency (NAF) (frequency of non-amplified alleles resulting in an apparent homozygote) was estimated per locus (10,000 bootstrap pseudoreplicates) using the expectation maximum algorithm (EMA)

implemented in FreeNA (Chapuis & Estoup 2007). Two microsatellite loci, Asb23 and Lex33, had low amplification success in samples from Kruger, with 63% and 71% missing data, respectively. Therefore, genetic analyses were conducted with these loci excluded.

Intra-population genetic diversity

To assess general patterns of intra-population diversity, I calculated the average number of alleles (A), expected heterozygosity (H_E) and the fixation index (F_{IS}) at all loci using Genalex (Peakall & Smouse 2006). Allelic richness, being particularly sensitive to differences in sample size, was also corrected (A_{CORR}) through a rarefaction approach implemented in the program HP-RARE 1.1 (Kalinowski 2005). Allelic variation in zebra populations was also quantified at sequence-based loci (*β-Fibr*, *DRA*, *DQA*) in terms of number of segregating sites (S), haplotype diversity (H_D; Nei 1987), nucleotide diversity (π; Nei & Tajima 1981) and the mean number of pair-wise nucleotide differences (k; Tajima 1983) in DNAsp (Rozas *et al.* 2003). Empirical distributions were generated to determine sampling variance and standard deviations of parameter estimates.

The *DQA* sequences found in this study are derived from at least two unresolved loci and are referred to as 'alleles' although I were not able to assign specific alleles to a locus. Therefore, I estimated allele frequencies as the number of individuals carrying a particular allele out of the total number of alleles. Sequenced homozygotes were assumed to have two copies of the observed allele, but for cloned heterozygotes each allele observed was only counted once. I recognize that this method for determining allele frequency will overestimate the frequency of rare alleles and underestimate the frequency of common alleles (see Ekblom *et al.* 2007; Miller *et al.* 2010). Therefore, I alternatively assessed intra- and inter-population variability at the *DQA* locus using measures independent of allele frequency: mean number of alleles per individual, total number of alleles per population and average percent difference (APD). APD is based on the average percentage of sequences that differ among all possible individual pair-wise comparisons, calculated as outlined in Yuhki and O'Brien (1990) and can be used to estimate within-population genetic variation from multi-locus data. As the *DQA* diversity data was not normally distributed, I used a Kruskal-Wallis rank randomization test to test the null hypothesis of group-mean equality. Differences in variance between population samples were addressed by a Levene test based on absolute residuals of each diversity observation to the respective population diversity mean. A Wilcoxon sign-rank test (Z test) was performed to account for unequal variance between population samples in group-mean comparisons.

I further tested for significant differences in the distribution of the number of *DQA* alleles per individual across populations, through goodness-of-fit contingency analyses, with significance assessed through calculation of the Chi Square statistic (χ^2). Statistical analyses were performed using JMP 4.0 (SAS Institute Inc.) software.

Population differentiation

Gene flow and drift are expected to affect neutral loci to the same degree, whereas selection acts on specific loci (Lewontin & Krakauer 1973) and discordance in population structure from that observed at neutral loci may provide evidence for selection (Lynch *et al.* 1999). Therefore, I contrasted the partitioning of genetic variation at functional MHC

loci to that of neutral loci to control for the confounding effects of migration and population size, and shed light on the spatial heterogeneity of selection pressure across populations.

Population genetic structure was assessed in Arlequin v3.1 (Excoffier *et al.* 2005) by computing conventional *F*-statistics (F_{ST}; Wright 1951). For sequence data, the F_{ST} estimator Φ_{ST} (Weir & Cockerham 1984) was calculated from a Kimura-2 parameter distance matrix and significant deviations from 0 determined by 1,000 permutations. For microsatellite loci, I also calculated *R*-statistics (R_{ST}; Rousset 1996, Goodman 1997) which assume a stepwise mutation model and is based on the sum of squared size differences among alleles (Slatkin 1995). Ninety-five percent confidence intervals were generated for microsatellite F_{ST} estimates by bootstrapping (10,000 replicates) in FSTAT v2.9 (Goudet 1995) and MHC values falling outside of this interval were considered to be significantly different (see Landry & Bernatchez 2001). MHC estimates were also directly compared to F_{ST} values derived from the neutral intron, *β-Fibr*. Population and allele divergence was further assessed by global exact tests of population differentiation (Markov chain steps = 100,000, dememorization steps = 10,000) in Arlequin v3.1 (Excoffier *et al.* 2005).

At the *DQA* locus, I calculated a measure of differentiation F′ (analogous to F_{ST}), to account for the bias in allele frequencies (discussed previously). This measure is derived from the similarity index (Lynch 1990), and because it uses percent similarity among pair-wise sequences of individuals to estimate population sub-division, it is analogous to APD. Standard errors of F′ were estimated by applying a Taylor expansion approximation (see Lynch & Crease 1990).

Sequence analyses of selection versus demography

I conducted hypothesis testing to check for significant departures from neutrality at MHC loci using approaches that reflect both recent and historical processes and/ or events, implemented in Arlequin 3.5 (Excoffier *et al.* 2005) and DNAsp 4.5 (Rozas *et al.* 2003). Significance was tested with 10,000 simulations. First, I used Slatkin's Markov-Chain Monte Carlo (MCMC) implementation (Slatkin 1994) of the Ewens-Watterson (E-W) Test (Ewens 1972; Watterson 1978) to test for recent selection or demographic events (i.e. rapid population expansion or bottlenecks) affecting mutational patterns. The E-W test compared the observed homozygosity (F_{obs}) with the expected homozygosity (F_{exp}), based on a random sample of the same size consisting of the same number of alleles, simulated under the assumption of neutrality (10,000 replicates). Excess heterozygosity, as expected under balancing selection or a recent population bottleneck, would be indicated by significantly lower F_{obs} than expected. Therefore, this test is particularly useful for detecting cases of overdominance (i.e. heterozygote advantage). As E-W tests are based on allele frequency distributions, they are generally used to reflect relatively recent (i.e. over the history of a population) processes or events (see Garrigan & Hedrick 2003).

Tajima's *D* (*D*; Tajima 1989) and Fu's F_S (F_S; Fu 1997) were calculated for each population to test for departures from the null hypothesis of neutral evolution using Arlequin v3.1 (Excoffier *et al.* 2005). Tajima's *D* and Fu's F_S are analogous to the E-W test, but because they are based on sequence data (infinite site model) rather than allelic data (infinite allele model) they reflect historical selective pressures or demographic

changes, in addition to recent events. Fu's *F* measures the probability of observing a certain number of alleles given the average pair-wise sequence divergence (θ), whereas Tajima's statistic contrasts θ to that which would be expected under neutrality given the number of segregating sites (*S*). Negative values for both statistics imply population expansion or purifying selection, and positive values suggest either a population bottleneck or balancing selection.

These neutrality tests depend on the assumption that population size has been constant through time (Nei 1987). Therefore, in order to tease apart the signatures of demography and selection, I contrasted neutrality test results to those performed on the intron (*β-Fibr*). In addition, allele frequency distributions at neutral and functional loci were statistically compared across populations with contingency analyses in JMP 4.0 (SAS Institute Inc.).

Molecular detection of selection

For molecular-level evidence of selection, I estimated rates of non-synonymous to synonymous mutations ($\omega = d_N/d_S$) at coding genes (*DRA, DQA*) using maximum likelihood models of codon-substitution in the CODEML subroutine of PAML (Yang 2007). I performed maximum likelihood ratio tests (LRTs) to test for significant positive selection by comparing the likelihoods of models of neutral evolution (M1a, M7) to those incorporating positive selection (M2a, M8) as described in Kamath & Getz, 2011. I also compared models assuming one evolutionary rate across codon sites (M0) to those allowing for heterogeneous rates (M3).

Demographic inference

I constructed mismatch distributions for sequence-based loci in Arlequin 3.5 (Excoffier *et al.* 2005), plotting frequency distributions of pair-wise differences among alleles, to test the null hypothesis of recent increase in population size. This analysis compares the observed frequency distribution of pair-wise differences to the expected uni-modal distribution of a population that has undergone a sudden expansion, generated through coalescent simulations (Hudson *et al.* 1992; Rogers & Harpending 1992). Multi-modality and deviations from the expected distribution may be indicative of a population that has reached stationary demographic equilibrium and right-shifted peaks may represent older expansion events. Goodness-of-fit tests for population expansion were conducted by calculating the sum of squared deviation (SSD) and raggedness index (RI) (Harpending *et al.* 1998), and significance determined by coalescent simulations (n = 10,000), assuming a large panmictic population of constant size under neutral evolution.

I used neutral microsatellite markers to characterize historical changes in zebra population sizes through implementation of a coalescent-based model in LAMARC v2.1.5 (Kuhner 2006). For each population, I conducted two independent Bayesian runs under the Brownian motion approximation mutational model (Beerli & Felsenstein 2001), with a sampling routine of 40,000 parameter sets at intervals of 80 increments, and a burn-in of 4,000. Demographic parameters that reflect total population size and changes in population size were jointly estimated from the posterior sampling distributions, including: (1) Theta (θ), a measure of diversity proportional to the effective population size (*Ne*) and mutation rate (μ) such that ($\theta = 4Ne\mu$) and (2) the exponential growth rate parameter (g) indicating the direction and magnitude of change in population size. I

evaluated these parameters to distinguish between putative population expansions or bottlenecks, and compared relative magnitude across populations. A Metropolis-coupled MCMC approach was employed for each run (using one cold chain and four heated chains) to enable thorough searching of the parameter space. Acceptance rates were assessed to ensure they fell between 5 and 40%, probability density functions inspected for uni-modality, and effective sample size (ESS) were confirmed to be >200 using Tracer v1.5 (Rambaut & Drummond 2007).

Finally, I specifically tested for historical declines in population size using neutral microsatellites in the program BOTTLENECK (Piry *et al.* 1999). This analysis computed the expected heterozygosity distribution for each locus given the observed number of alleles and sample size under the assumption of mutation-drift equilibrium. I simulated the coalescent process under a step-wise mutation model with 1,000 replicates. The proportion of loci for which the expected heterozygosity (H_E) was greater than the heterozygosity estimated under mutation-drift equilibrium (H_{EQ}) was calculated and compared to the 50% proportion expected in a non-bottlenecked population (i.e. 50% of the time $H_E > H_{EQ}$, while 50% of the time $H_{EQ} > H_E$). I used a Wilcoxon sign-rank test to assess significant departures from these expectations (Luikart *et al.* 1998). I also determined whether the observed allele frequency distribution differed significantly from an L-shape distribution typical of non-bottlenecked populations using the "mode-shift" indicator.

Phylogenetic allele networks

Historical MHC allele relationships were constructed in SplitsTree v4.11.3 (Huson & Bryant 2006), using the neighbor-net algorithm which computes a set of incompatible splits (i.e. parallel branches) to form a split network from the data distance matrix. Evolutionary distances were calculated following the Kimura 3-parameter model for nucleotide substitution using the Akaike Information Criterion (AIC) model selection criterion in ModelTest 3.7 (Posada & Crandall 1998). Based on AIC results, rate variation was assumed to be equal across sites and proportion of invariable sites (Pinvar) to be 0.9018 at the *DRA* locus. Whereas, at the *DQA* locus, rate variation was gamma-distributed ($\alpha = 0.4918$) with Pinvar = 0. The simple indel coding method (Simmons & Ochoterena 2000) was used to incorporate information from the indel variation at the *β-Fibr* locus.

Under coalescent theory, the allele network of a neutral locus is expected to reflect changes in population size. Therefore, I constructed a *β-Fibr* network using the statistical parsimony method implemented in TCS 1.21 (Clement *et al.* 2000), incorporating allele frequency information and percent from each population. Unequal sample sizes were adjusted though rarefaction (Hurlbert 1971).

Results

Microsatellite exact tests and error estimates

The microsatellite loci used in these analyses were highly polymorphic, with a mean of 9.0 alleles per locus (range: 3-15) in the Etosha zebra population and 7.1 alleles per locus (range: 3-12) in Kruger (Table S3). Mean expected heterozygosity was 0.76 and 0.73 over all samples genotyped in Etosha and Kruger, respectively (Table S3).

Results from exact tests by locus and population (after application of Bonferroni corrections) indicated significant ($p < 0.001$) departures from H-W equilibrium at Hms7 and Htg15 in Etosha. Further tests revealed that there was significant heterozygote deficiency at these particular loci. Heterozygote deficiency could reflect a genuine biological phenomenon, such as inbreeding, selection against heterozygotes, or population subdivision into separate breeding units (i.e. the Wahlund effect; Wahlund 1928), but the above seemed unlikely given only a subset of loci showed excess homozygosity. Alternatively, heterozygote deficiency may be due to a locus-specific genotyping effect, such as the presence of null alleles. However, I found evidence for heterozygote deficiency and predictions of elevated rates of null alleles only in Etosha, but not in Kruger. Therefore, although unlikely, I considered the possibility that a mutation in the flanking region of these loci in the Etosha population may have interfered with primer binding and null allele frequency was assessed accordingly (see below). Note that these observations may be the result of local purifying selection on closely linked functional loci in this particular population.

After implementing the Bonferroni correction procedure, I found evidence for significant LD ($p < 0.001$) between Hmb1 and Htg14 in Etosha only, but not in Kruger. Again these patterns differed across populations and, therefore, I inferred that loci were physically independent and that the observed LD may be the result of population substructure within the Etosha.

The mean null allele frequency (NAF) over all loci and populations, was approximately 2.3% (Table S4), and ranged from 0 to 6.8% by locus. Total genotyping error varied by locus (ranging from 0 to 7%) and was found to be extremely low with a large proportion of loci exhibiting no error (7 out of 15 loci) (Table S4). The mean error rate across the 15 loci was 0.3% in blood samples and 2.5% in fecal samples genotyped. Breakdown of error contributions in fecal sample genotyping revealed false alleles (mean = 2.1%) as the most significant factor contributing to the observed error and allelic dropout accounted for only ~0.5% of the error. Additional PCRs, following the comparative approach of Frantz *et al*, 2003, lowered the error rate in fecal samples to 0%. Therefore, I am confident that this genotyping approach yielded relatively accurate results and upholds the assumptions made in subsequent analyses.

Intra- and inter-population genetic variation

`This study assessed diversity and selection in the functional antigen binding regions of MHC loci, *DRA* and *DQA*, within two natural populations of plains zebra. Intron 7 of the *β-Fibr* gene was also successfully sequenced and contrasted to MHC data to quantify neutral background genetic variability. Ten novel haplotypes were sequenced at the *β-Fibr* locus in *E. quagga* and deposited in GenBank [GenBank: XXX], and two indel mutations totaling 14 base pairs were discovered (Table 1). Nine *DRA* alleles (*Eqbu-*

*DRA*01, *03-*05, *07-*11*) [GenBank: AJ575299, EU930126, EU930121, EU930118, HQ637392 - HQ637396] were recovered in Etosha and Kruger zebra populations. Twenty-one *DQA* alleles *(Eqbu-DQA*01-02, *04-*22)* [GenBank: EU935837, EU935829, EU935834-EU935836, EU935832, EU930130, HQ637397- HQ637409, XXX] were also recovered in these two populations. All but one MHC allele, *DQA*22*, were recovered in a previous study (Kamath & Getz 2011). Estimates of allelic richness corrected for unequal sample sizes, through rarefaction, were similar to estimates made without the correction applied and revealed similar inter-population diversity patterns at all loci.

At both MHC loci, indices of sequence diversity (H_D, π and k) were relatively high when contrasted to neutral sequence data at the *β-Fibr* intron, particularly for the *DQA* locus (Table 1). Interestingly, comparisons of indices across populations revealed depressed sequence diversity at the *DRA* locus in Etosha relative to Kruger (Table 1; Figure 1). This pattern was consistent across sequence-based (H_D, π and k) and allele frequency-based (H_E) indices (Etosha: $H_D = 0.748$, $\pi = 0.0088$, $k = 2.158$, $H_E = 0.743$; *Kruger:* $H_D = 0.874$, $\pi = 0.0097$, $k = 2.393$, $H_E = 0.860$). This observation is also contradictory to what I would expect given the diversity at neutral intron and microsatellite data─ mean allelic diversity (A) and expected heterozygosity (H_E) at microsatellite loci were both greater in Etosha than in Kruger (Table 1). All diversity estimates were similarly higher in Etosha at the considered intron.

At the *DQA* locus, diversity estimates were similar across populations. Furthermore, APD, a diversity index independent of allele frequencies, corresponds with this inter-population and inter-locus pattern (Table 2; Figure 1). Non-parametric statistical analyses of APD diversity of the *DRA* locus between populations corroborated that diversity among individuals was significantly lower in Etosha than Kruger ($\chi^2 = 75.80$, $p < 0.001$; $Z = 8.701$, $p < 0.0001$), whereas the opposite was observed at the *β-Fibr* intron which exhibited significantly higher APD in Etosha ($\chi^2 = 18.45$, $p < 0.001$; $Z = -4.293$, $p < 0.001$). However, at the *DQA* locus no significant difference between population means was found ($\chi^2 = 1.15$, $p = 0.284$; $Z = -1.072$, $p = 0.284$) (Table 2). Despite the potential bias in *DQA* diversity estimates due to the inability to accurately estimate allele frequencies, the correspondence of APD patterns across loci and populations with other diversity estimates suggests these conclusions are robust.

At the *DQA* locus, there was on average 1.56 (range: 1.31-1.80) and 1.77 (range: 1.46-2.07) alleles per individual in the Etosha and Kruger zebra populations, respectively. Most individuals for both populations were homozygous with only one allele, indicating potential inefficiency in the primers to amplify the second *DQA* locus. Contingency table analyses and a likelihood ratio test revealed no evidence for significant differences in copy number frequency distributions across populations ($\chi^2 = 5.578$, $p = 0.1341$) (Figure 2).

Population differentiation

Estimates of population pair-wise F_{ST} revealed evidence for significant, albeit low, levels of genetic differentiation at MHC loci across populations (*DRA*: $F_{ST} = 0.045$, $p < 0.001$; *DQA*: $F_{ST} = 0.016$, $p = 0.042$; Table 3). When compared to neutral data, *DQA* population structure was significantly lower than that observed at neutral loci, falling outside of the microsatellite 95% CI (0.026 – 0.053). Whereas, *DRA* differentiation was

not significantly different from F_{ST} estimates observed at microsatellite loci ($F_{ST} = 0.038$, $p < 0.001$; $R_{ST} = 0.055$, $p < 0.001$). However, because the *β-Fibr* intron follows a similar mutational model, estimates of F_{ST} at this locus are more directly comparable to those of the MHC. Analyses of *β-Fibr* genetic structuring indicated far higher differentiation across populations ($F_{ST} = 0.140$, $p < 0.001$) and supported the observatiotan that low differentiation at MHC loci may be indicative of balancing selection acting on these genes. These results were corroborated by estimates of the F_{ST} analog, F′ (Table 3); Again, F′ estimates revealed considerably lower population structure at MHC loci than at the *β-Fibr* locus.

Neutrality tests on sequence loci

Tests for departures from neutrality by calculation of Tajima's *D* and Fu's F_S, were not significant ($p > 0.05$) at all loci and populations (Table 4). Although non-significant, test statistics were negative at the intron, indicating a potential weak affect of population expansion or, alternatively, purifying selection. Positive values of the test statistics were observed at MHC loci and can be indicative of either a population bottleneck or positive selection acting on the site frequency spectrum. Results from Slatkin's E-W test revealed significantly lower homozygosity (F_{obs}) than would be expected (F_{exp}) under a neutral model, an indication that balancing selection is likely acting on MHC loci. This observation could alternatively reflect a relatively recent population bottleneck, but this hypothesis can be rejected due to incongruent test results at the neutral intron (Table 4) and based on results of demographic analyses (see below). Significance of E-W tests, but not *D* or F_S, implies that selection at these loci is likely a relatively recent phenomenon.

The distribution of allele frequencies can act as signatures of demographic change and/or selection, and were used to tease apart and explain findings from neutrality tests. MHC allele distributions were inspected relative to that at the *β-Fibr* by population. The most outstanding observation was that the *DRA* allele frequencies were evenly distributed in Kruger, whereas in Etosha it was skewed (Figure 3). In particular, the *DRA*03* allele was discovered to be predominant in Etosha representing approximately 44% of all alleles observed which could potentially be driving the decreased diversity found at this locus. The *DQA* locus visually exhibited similar distributional patterns across populations (Figure 3), but contingency analyses still suggested significantly different distributions ($p < 0.001$). Also, *DQA* distributions differed in the large number of rare and private alleles found in each population. In both populations, the *DQA*01* allele was present at a significantly greater frequency than any other allele (25 to 29%). *β-Fibr* allele frequency distributions were also significantly different ($p < 0.001$) between populations, with Kruger exhibiting a more skewed distribution (Figure 3).

Molecular selection analyses

Molecular-based analyses revealed no evidence for positive selection or variable rates of selection at the molecular level of the *DRA* locus in both zebra populations (Table 5). In contrast, likelihood ratio tests (LRTs) comparing models of invariable versus variable evolutionary rates across codon sites (M3 *vs.* M0) and models of positive selection to neutral evolution (M7 *vs.* M8) were significant ($p < 0.05$) at the *DQA* locus in both populations. In Etosha, all LRTs were significant, including the M1a *vs.* M2a comparison.

Demographic inference based on neutral data

Mismatch distribution analyses at the neutral intron indicated a weak signature of recent population expansion in the Etosha zebra population, as the null model could not be rejected (SSD = 0.014, $p = 0.0612$) (Table 4; Figure 4E). In contrast, the model of sudden population growth was rejected in Kruger (SSD = 0.029, $p = 0.041$) (Table 4; Figure 4F), suggesting a stabilized population. Similarly, the intron allele network showed a pattern typical of expansion for Etosha (Figure 5). Harpending's raggedness index (RI) was significant at the *DQA* locus in both populations (Etosha: RI = 0.042, $p < 0.001$; Kruger: RI = 0.055, $p = 0.009$) and the mismatch distribution at this locus was multi-modal (Table 4; Figure 4C, D). At the *DRA* locus in Etosha, RI was also significant (RI = 0.015, $p = 0.042$) with a shift in the distribution peak to greater pair-wise differences than would be expected under recent population expansion (Table 4; Figure 4A) and, thereby, incongruent with demographic inference based on intron data. Also, the *DRA* locus in Kruger reflected a pattern typical of population expansion which is in contrast to the evidence for demographic stability inferred via neutral data (Table 4; Figure 4B).

Results from demographic analyses in LAMARC indicated population stability in both Etosha and Kruger, with little to no population growth or decline. The most probable estimate for growth rate (and 95% confidence interval) in Etosha was 1.30 (0.08 - 1.55), suggesting very low positive growth. Whereas, in Kruger, growth rate was estimated to be -37.37 (-495.62 - 10.11) and, although negative, the confidence interval included zero, suggesting no considerable demographic change. The finding of some population growth in Etosha may be consistent with the mismatch results; however the magnitude of g is extremely small and is more likely indicative of population stability. The most probable estimate of theta (θ) was comparable across populations, estimated to be 9.93 (6.34 – 10.13) and 9.84 (5.33 – 10.08) in Etosha and Kruger populations, respectively. These results suggest that the zebra populations may be similar in population size. In addition, analyses in BOTTLENECK and Wilcoxon sign-rank tests showed no significant evidence for the occurrence of historical size reductions in either population (Etosha, $p = 0.97$; Kruger, $p = 0.42$). Further, the "mode-shift" indicator suggested normal L-shape allele frequency distributions, indicating the presence of many low-frequency alleles, an observation expected for non-bottlenecked populations.

Phylogenetic relationships among alleles

Phylogenetic allele networks of MHC loci showed a lack of geographical structuring of alleles by population (Figure 6). A large number of alternative connections were found among alleles at MHC loci, potentially due to the high sequence divergence among alleles present, whereas the relationship among *β-Fibr* alleles was straight-forward. In general, long branches were found, particularly in the *DQA* allele network, which suggests divergent alleles have persisted over a considerable amount of time.

Discussion

This study juxtaposed MHC sequence data with neutral data from microsatellites and a nuclear intron in order to disentangle the co-occurring effects of selection from demography on immunogenetic variation across populations. These results suggested that selection is largely affecting MHC loci in zebra, but also that it likely acts variably across genes and populations.

Evidence for balancing selection and temporal variation in selection

The hypothesis that balancing selection is driving the maintenance of MHC gene diversity is well accepted and supported in the literature (see reviews by Bernatchez & Landry 2003; Piertney & Oliver 2006). Further, recent studies focused on long-term evolutionary patterns (through evaluation of d_N/d_S ratios and MHC gene phylogenies) have provided genus-level evidence for balancing selection acting on both the *DRA* and *DQA* loci in equids (Janova *et al.* 2009; Kamath & Getz 2011). Specifically, balancing selection in these studies was evident from observations of alleles retained through species divergences, also known as 'trans-species polymorphism' (Takahata 1990). The results here provided several lines of evidence to further support its occurrence over shorter evolutionary time scales in zebra (i.e. over the history of a population or species), including (1) elevated MHC diversity over neutral diversity within populations (Table 1 and 2; Figure 1), (2) low genetic structure across populations relative to that observed at neutral loci (Table 3) and (3) significant rejection of the null hypothesis of neutrality based on the assumption of mutation-drift equilibrium in Ewens-Watterson-Slatkin (E-W-S) tests (Table 4).

These findings combined provided convincing evidence for balancing selection operating on MHC genes in zebra. However, discrepancies were found in molecular analyses of selection as well as Tajima's *D* and Fu's F_S test results. Rates of nonsynonymous to synonymous mutations ($\omega = d_N/ d_S$) within zebra populations were low (*DRA*, $\omega \approx 0.2$; D*QA*, $\omega \approx 0.7$ - 1.0), pointing towards purifying selection at the *DRA* locus and neutral evolution at the *DQA* locus. Likelihood ratio tests revealed significant heterogeneity in ω across *DQA* codon positions and evidence for positive selection at specific codons. In contrast, I could not reject the hypothesis of one evolutionary rate across *DRA* codons. At both genes, Tajima's *D* and Fu's F_S tests were not significant, and I was unable to reject the hypothesis of neutral evolution.

At first glance these results appear be contradictory. However, the ability of these tests to detect selection differs depending on the nature of the mutational process and the duration, strength and timing of selection (Garrigan & Hedrick 2003). Demographic processes may interfere with and obscure signatures of selection, and it is unclear whether inspection of current variation at these genes has preserved this signature (further discussed below). Despite lack of significance, the sign of Tajima's *D* and Fu's F_S test statistics were consistent with positive selection at MHC genes and contrasted with values at the neutral intron, indicating a potential weak signal of selection. These test statistics reflect selection operating to change the site frequency spectrum, and accumulation of such mutations may require long time periods extending beyond the history of a population. In contrast, the significant E-W tests recovered here point to the conclusion that relatively recent selection is playing an important role in shaping allele

frequency distributions within zebra populations. While balancing selection may have maintained the diversity at MHC genes, it is possible that signatures of selection have been obscured by fluctuating selective pressures due to changing pathogen communities or demographic events over time.

Analyses of population genetic structure

At both neutral and functional loci, genetic differentiation across zebra populations was relatively low ($F_{ST} \leq 0.14$) (Table 3). This finding is in agreement with a previous phylogeographic study on five subspecies of *E. quagga* that showed structuring in zebra populations is among the lowest out of 17 different savannah-adapted ungulate species in Africa (Lorenzen *et al.* 2008). Also, F_{ST} at the *β-Fibr* intron (F_{ST} = 0. 16, Φ_{ST} = 0. 18) was higher relative to F_{ST} based on microsatellites (F_{ST} =0.038, R_{ST} = 0.055). These result are also concordant with their study (mtDNA: Φ_{ST} = 0. 173, microsatellites: θ_{ST} = 0.053). This, in addition to non-significance of global exact tests at microsatellite loci (p = 1), but significance at the intron locus ($p < 0.001$), are likely due to differences in the mutational modes and rates of these markers (Hedrick 1999).

Under the theory of over-dominance, I would expect population differentiation at MHC loci to be low due to theoretically greater effective migration rates relative to neutral loci (Schierup *et al.* 2000). This has been shown to occur even between small populations with little gene flow that would be highly susceptible to the effects of drift (e.g. Mona *et al.* 2008). Consistent with this prediction, I found that measures of F_{ST} at the *DQA* locus were seven to eight times lower than at the *β-Fibr* intron, and two to three times lower than at microsatellite loci. Populations shared many of the same alleles, though five to six private alleles were observed per population. These findings suggest that balancing selection is likely acting on the *DQA* locus under a mechanistic over-dominance model, as is also supported by significant E-W Tests. Alternatively, this lack of differentiation could also be found without over-dominance if similar pathogens were present in both environments.

At the *DRA* locus, Φ_{ST} was also two to three fold lower than Φ_{ST} calculated at the intron, but a discrepancy in this finding was found with measures of differentiation similar to that calculated for neutral microsatellite loci (Table 3). Several previous studies have also reported results that contradict this expectation, with discoveries of concordant or elevated population structure at MHC versus microsatellite loci (e.g. Landry & Bernatchez 2001; Miller *et al.* 2001; Ekblom *et al.* 2007; Miller *et al.* 2010). These results can be attributed to either (i) local adaptation, (ii) weak selection acting on the MHC locus, or (iii) an artifact of using genetic markers evolving under different mutation modes and rates (Hedrick 1999). The results found in this study could be due to any of the aforementioned causes or, alternatively, explained by the effects of historical demographic events (discussed below). However, the comparison of estimates of *DRA* structure to *β-Fibr* F_{ST} or Φ_{ST} does, in fact, follow a pattern akin to the observations at the *DQA* locus and argues for the explanation that the difference in the microsatellite mutation model may be responsible for the discrepancy I have observed. This underscores the potential problem with solely using microsatellite data for inferences about selection on functional sequence data.

Despite being low, MHC differentiation remained significant as indicated by pair-wise F_{ST} and global exact tests. Given that differentiation at the *β-Fibr* intron was much

greater than at MHC loci, particularly at the *DQA* locus, structure at the intron is assumed to reflect the processes of mutation and drift. This would be probable given that little to no recent gene flow between these two populations is expected, because the populations are contained in parks and separated by more than 2,000 km of a human-dominated matrix. Therefore, I conclude that drift may be affecting MHC differentiation, albeit weakly, by stochastically shifting allele frequency distributions within populations. This conclusion is supported by the lack of pronounced geographical structuring in intra-specific allele parsimony networks (Figure 6).

Inter-population patterns of genetic diversity and pathogen composition

The Etosha zebra population had higher genetic diversity than Kruger at all neutral loci (Table 1 and 2; Figure 1). This pattern was fairly consistent regardless of the diversity index evaluated, and may reflect differences in population size or the effect of a demographic event in one or both populations. Demographic scenarios are discussed in more detail in the following section below. In contrast, *DQA* population patterns are generally similar across populations, except at measures based on sequence diversity which agree with results from neutral data (Figure 1). This result may suggest the strength of selection is strong enough to preserve similar levels of diversity, potentially due to equivalent pathogen regimes across populations.

Interestingly, the pattern of *DRA* genetic diversity across populations is strikingly incongruent with that observed at neutral loci, revealing that variability at this locus is depressed in Etosha (Figure 1). Allele frequency distributions suggest that the paucity of diversity at this locus in Etosha is driven by the presence of one predominant allele (Figure 3). In contrast, although Kruger possessed many of the same *DRA* alleles as Etosha, allele frequencies were relatively evenly distributed. The finding of opposing *DRA* and neutral diversity across populations argues against the fact that demography plays a predominant role in shaping diversity at this locus and suggests that selection at the *DRA* is geographically heterogeneous.

The marked difference in *DRA* patterns across populations, with respect to neutral data, highlights the possibility that zebra in Etosha may be subject to strong local selective pressure by pathogens at this locus. A previous study by Matthee *et al*, 2004, found that gastrointestinal helminth communities differed significantly among African equids and even between the Etosha and Kruger zebra populations, suggesting that specialization of parasite communities has followed host population divergence. Notably, zebra of Etosha had a much lower Stronglyinae nematode species richness with only one species as opposed to six identified in Kruger (Matthee *et al.* 2004). Despite such low diversity in Etosha, nematode prevalence has been found to be extremely high (> 98% of individuals infected) (Turner & Getz 2010). It is possible that Etosha's particularly arid climate (rainfall < 500 mm/year), relative to Kruger's (rainfall 550 - 650 mm/year), may play an important role in limiting parasite diversity in Etosha (Matthee *et al.* 2004). Besides intestinal parasites, zebra in Etosha are also known to be one of the main hosts of anthrax, a lethal infectious bacterial disease (Turnbull *et al.* 1989). Anthrax outbreaks in Etosha are severe and occur annually in wildlife (Lindeque & Turnbull 1994). Interestingly, zebra in the park exhibit the highest recorded incidence of anthrax in southern Africa and the disease has been implicated as one of the primary causes of adult mortality (Ebedes & Page 1976). In contrast, Kruger zebra experience infrequent anthrax

outbreaks (de Vos 1990). Given the role of MHC class II genes in recognizing these particular pathogen types, I speculate that either the low, prevalent parasite community or repeated anthrax outbreaks may act as significant selective pressures driving the observed *DRA* diversity in Etosha. Specifically, I hypothesize that high prevalence of any one pathogen may directionally select for a particular advantageous *DRA* allele, causing it to increase in frequency, consistent with a 'selective sweep'.

The effects of population demography on MHC variation

Demography and migration may confound the ability to detect selection. Therefore, I used neutral genetic data to elucidate population demographic history and disentangle these effects from that of selection on patterns of immune gene variation. The demographic analyses based on microsatellite loci suggested that zebra abundance is similar in Etosha and Kruger. However, years of survey data suggest this is not the case; in Etosha, zebra abundance (and 95% confidence intervals) was estimated at approximately 13,000 (10,900 – 15,000) individuals (see Turner & Getz 2010), whereas in Kruger size estimates were nearly three times as large (Owen-Smith & Mason 2005) Coalescent mismatch analyses, using intron data, provided evidence for a recent expansion in Etosha and suggested that Kruger has been supporting a stable population (Table 4; Figure 4). In contrast to this, coalescent estimates of growth rates based on microsatellite loci suggested stability in both populations. This discrepancy may again be the result of the mutational mode of the genetic marker assessed reflecting processes that have occurred on different time scales. Taken together, I hypothesize that Etosha may have undergone a historical population expansion, but that both zebra populations have remained stable over recent time.

Interestingly, mismatch distributions between neutral and functional genes are discordant within each population, likely due to differences in selective pressures. The ragged, multimodal distribution at the *DQA* locus in both populations is consistent with long term stability at this locus. In contrast, the uni-modally shifted peak at the *DRA* locus argues for an older effective expansion at this gene and contradicts results from neutral data, suggesting that this result is not due to population demography. In summary, I conclude that selection has played a stronger role than demography in shaping MHC diversity across these two zebra populations.

Conclusions

MHC studies in natural populations are critical to understanding adaptation and evolutionary potential as variation in these genes reflect biologically relevant processes significant to fitness. Patterns of MHC variation are shaped by a complex interplay of selective and demographic factors, which may be challenging to disentangle, but possible to achieve through the combination of multiple lines of evidence. These data suggest that selection on MHC genes in zebra populations varies spatiotemporally, and also differs by locus. Balancing selection over evolutionary time scales may act cumulatively to retain MHC diversity, but this selection signature may not be detectable over shorter periods of time, potentially due to fluctuating and diverse pathogen communities. I found substantial evidence for balancing selection at MHC genes in zebra populations, while data also

suggested the possibility of ongoing directional selection acting locally at the *DRA* locus, two findings which may be compatible with one another when considering different time scales. Demographic events, such as population expansions, were not found to have significant effects on the observed distribution of diversity among populations. These results highlight the importance of integrating neutral and adaptive data over different time and spatial scales to elucidate the relative effects of demography and selection in shaping adaptive diversity. Future ecological studies are warranted that link host immune diversity with pathogen community structure to better understand the mechanisms driving gene adaptation.

Acknowledgments

I would like to thank W.M. Getz, R. Bowie and members of the Getz and Bowie Labs at UC Berkeley for valuable comments throughout this study's development. I am also grateful to O. Putzeys, H. Ganz, W. Turner for help collecting samples and to M. Tsyaluk, K. Hall and E. DeFranco for assistance in the lab. Field work and sampling was made possible by support from the Namibian Ministry and Environment and Tourism (MET) and South African National Parks (SANParks). I also thank the scientific staff at the Etosha Ecological Institute (EEI), particularly W. Kilian, who provided resources and support for this study. This research was supported by the National Institute of Health Ecology and Evolution of Infectious Disease (NIH-EEID) GM083863 to WMG and National Science Foundation Doctoral Dissertation Improvement Grant (NSF-DDIG) MCINS-20091291 to PLK.

Tables

Table 1. Indices of genetic diversity at neutral (*β-Fibr* intron, microsatellites) and MHC loci (*DRA, DQA*) within zebra populations (Etosha and Kruger)

Population	Locus	*N*	*A*	A_{CORR}	H_E	F_{IS}	*S*	*Indels (bp)*	H_D *(SD)*	*π (SD)*	*k (SD)*
Etosha	μsats (*n* = 13)[1]	84	8.92 (± 0.87)	7.99	0.767 (± 0.038)	0.056 (± 0.031)	---	---	---	---	---
	β-Fibr intron[2]	36	8	7.75	0.725	0.119	6	14	0.682 (0.037)	0.0015 (0.0002)	0.961 (0.110)
	ELA-DRA	72	8	7.84	0.743	0.421	9	0	0.748 (0.028)	0.0088 (0.0005)	2.158 (0.141)
	ELA-DQA	36	16	15.71	---	---	80	1	0.885 (0.025)	0.0916 (0.0073)	18.687 (1.361)
Kruger	μsats (*n* = 13)[1]	38	7.08 (± 0.65)	7.01	0.749 (± 0.033)	-0.029 (± 0.037)	---	---	---	---	---
	β-Fibr intron[2]	24	6	6	0.517	0.748	4	14	0.490 (0.082)	0.0009 (0.0002)	0.553 (0.095)
	ELA-DRA	31	8	8	0.86	0.587	7	0	0.874 (0.013)	0.0097 (0.0005)	2.393 (0.237)
	ELA-DQA	30	15	15	---	---	81	0	0.897 (0.021)	0.0837 (0.0101)	17.165 (1.338)

Measures derived from genotypic data include the allelic richness (A), rarefaction-corrected allelic richness (A_{CORR}), expected heterozygosity (H_E) and inbreeding co-efficient (F_{IS}). Sequence-based diversity measures include number of segregating sites (S), number of insertion/ deletion mutations (Indels), haplotype diversity (H_D), nucleotide diversity (π) and average pair-wise difference between sequences (k). The number of individuals (N) used in diversity analyses are reported. Standard deviations (SD) or standard errors (± SE) for diversity estimates are in parentheses.

[1] *Mean (± standard error) estimates reported, averaged over 13 loci*
[2] *A, H_E, and F_{IS} incorporate indel mutations for allele identification*

Table 2. Within population diversity based on indices independent of allele frequency

	Etosha		Kruger		K-W test	Wilcoxon
Locus	*C*	APD ± SE	*C*	APD ± SE	χ^2, *p*-value	*Z*, *p*-value
β-Fibr intron	630	61.19 ± 1.41	276	50.66 ± 2.69	**18.45, *p* <0.001**	**-4.295, *p* <0.001**
ELA-DRA	2556	68.21 ± 0.75	465	84.37 ± 1.39	**75.80, *p* <0.001**	**8.701, *p* < 0.0001**
ELA-DQA	630	86.04 ± 1.15	435	85.38 ± 1.31	1.15, *p* =0.284	-1.072, *p* =0.284

Average percent difference (APD) and standard errors (SE) were estimated by pair-wise comparisons of all individuals in a population, with *C* equal to the number of comparisons made at each locus (Yuhki & O'Brien 1990). Chi-square (χ^2) and *Z* test statistics are reported for Kruskal-Wallis (K-W) rank randomization tests and Wilcoxon tests, respectively. Significant rejection ($p < 0.05$) of the null hypothesis of group mean equality is indicated in bold.

Table 3. Measures of population differentiation

Locus	F_{ST}	Φ_{ST} / R_{ST} [1]	Exact Test *p*-value	F' ± SE
µsats (n = 13)[2]	0.038 (p <0.001)[3]	0.055 (p <0.001)	p =1.000	---
β-Fibr intron	0.140 (p <0.001)	0.129 (p <0.001)	p <0.001	0.178 ± 0.027
ELA-*DRA*	0.045 (p < 0.001)	0.051 (p <0.001)	p <0.001	0.034 ± 0.013
ELA-*DQA*	0.016 (p =0.021)	0.018 (p = 0.042)	p <0.001	0.021 ± 0.014

Conventional population *F*-statistics (F_{ST}) based on haplotype frequencies, Φ_{ST} based on a Kimura 2-parameter distance matrix and global differentiation exact tests with associated *p*-values are reported. F' is measure of population sub-division analogous to F_{ST}, but based on the similarity index (Lynch 1990). Standard error (± SE) of F' was estimated through a Taylor expansion approximation.

[1] *Φ_{ST} estimates refer to sequence loci, whereas R_{ST} refers to microsatellites.*
[2] *Estimates based on n = 13 loci*
[3] *95% confidence interval = 0.026 – 0.053*

Table 4. Tests of neutrality and changes in population size at sequence loci.

Population	Locus	*N*	*D*	F_S	E-W-S Test		Mismatch Test	
					F_{obs}	F_{exp}	SSD	RI
Etosha	*β-Fibr* intron	72	-0.532	-1.681	0.269	0.312	0.014 (*p* =0.061)	0.085 (*p* =0.116)
	ELA-*DRA*	144	0.780	1.304	**0.225***	0.359	0.056 (*p* =0.060)	0.150 (***p* =0.042**)
	ELA-*DQA*	76	0.486	10.838	**0.132***	0.149	0.022 (*p* =0.122)	0.042 (***p* <0.001**)
Kruger	*β-Fibr* intron	47	-0.895	-1.776	0.325	0.371	0.029 (***p* =0.041**)	0.128 (*p* =0.372)
	ELA-*DRA*	62	1.538	0.550	**0.183***	0.302	0.013 (*p* =0.080)	0.051 (*p* =0.194)
	ELA-*DQA*	67	0.040	9.380	**0.149***	0.154	0.031 (*p* =0.380)	0.055 (***p* =0.009**)

Neutrality assessed by Tajima's *D* (*D*), Fu's F_S (F_S) and the Ewens-Watterson-Slatkin (E-W-S) test. Population expansion models tested by analyses of mismatch distributions: sum of square deviations (SSD) between observed and expected mismatch and raggedness index (RI) of mismatch distribution are reported. *N* = the number of alleles used for each locus and population. Significant rejection of the null hypothesis ($p < 0.05$) of neutrality or constant population size is indicated in bold by the asterisk (*), and *p*-values are reported in parentheses.

Table 5. Molecular detection of selection across codon sites at the *DRA* and *DQA*.

Locus	Model comparison	df	Etosha			Kruger		
			ω	χ^2	*p- value*	ω	χ^2	*p- value*
DRA	M0 vs. M3	4	0.213	6.60E-04	1.000	0.249	3.80E-04	1.000
	M1a vs. M2a	2		2.40E-05	1.000		1.00E-04	1.000
	M7 vs. M8	2		-4.18	1.000		4.00E-04	1.000
DQA	M0 vs. M3	4	1.082	**51.68**	**0.000**	0.681	**59.14**	**0.000**
	M1a vs. M2a	2		**13.16**	**0.001**		4.93	0.085
	M7 vs. M8	2		**14.56**	**0.001**		**7.54**	**0.023**

Nonsynonymous to synonymous substitution rates (ω) at each locus and population were estimated from M8 in PAML (Yang 2007). Likelihood ratio tests (LRT) for significant heterogeneity across sites (M3 v. M0) and for positive selection (M1a v. M2a, M7 v. M8) are reported. Significance (*p*-value), determined by calculation of the chi-square test statistic (χ^2) and degrees of freedom (df), highlighted in bold.

Figures

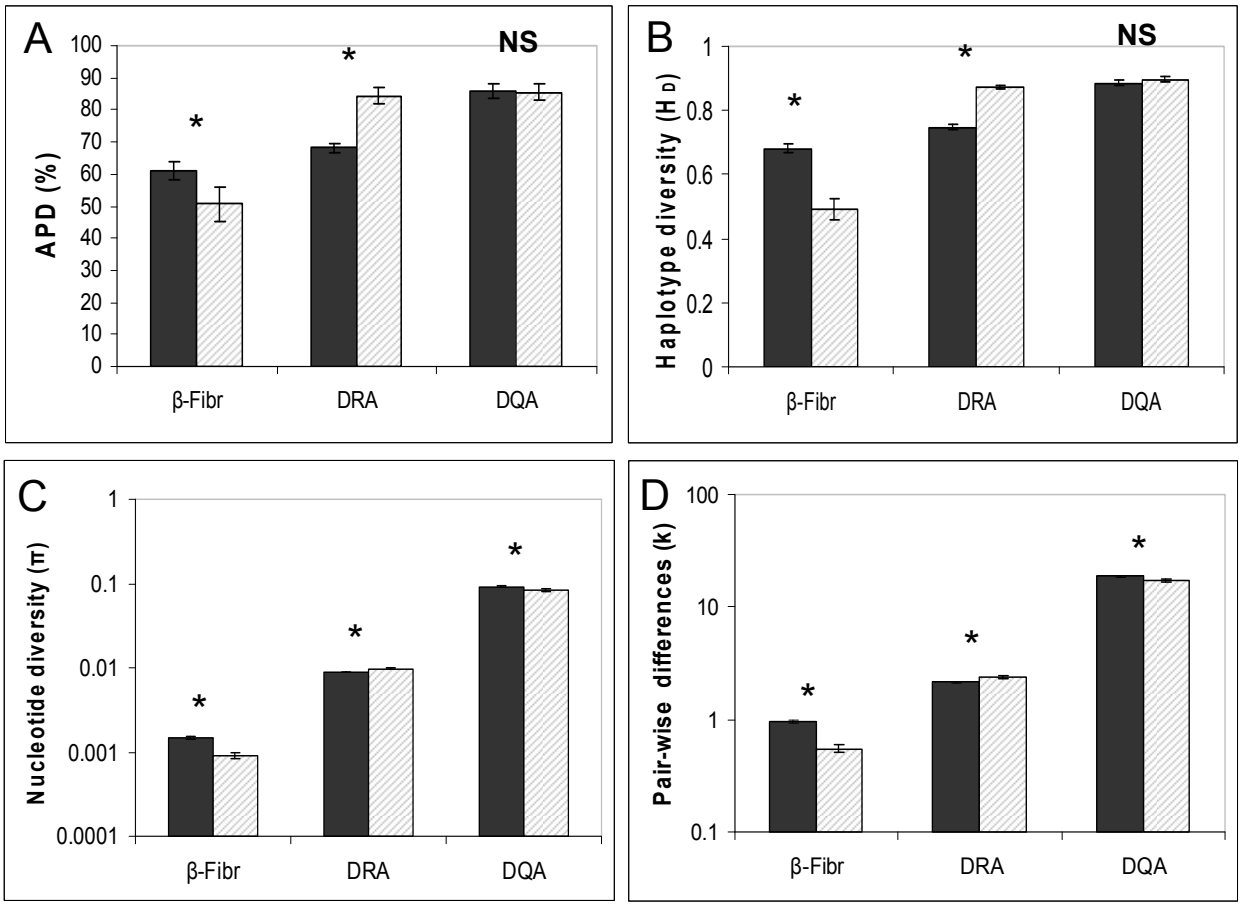

Figure 1. **Statistical comparisons of diversity indices at a neutral intron (*β-Fibr*) and MHC loci (*DRA* and *DQA*) across populations**
Etosha (black) versus Kruger (gray diagonal). Diversity indices reported are: A. average percent difference (APD), B. haplotype diversity (H_D), C. nucleotide diversity (π) and D. average pair-wise sequence difference (k). Significant difference ($p < 0.05$) between population means are indicated by asterisk (*) and non-significance by NS.

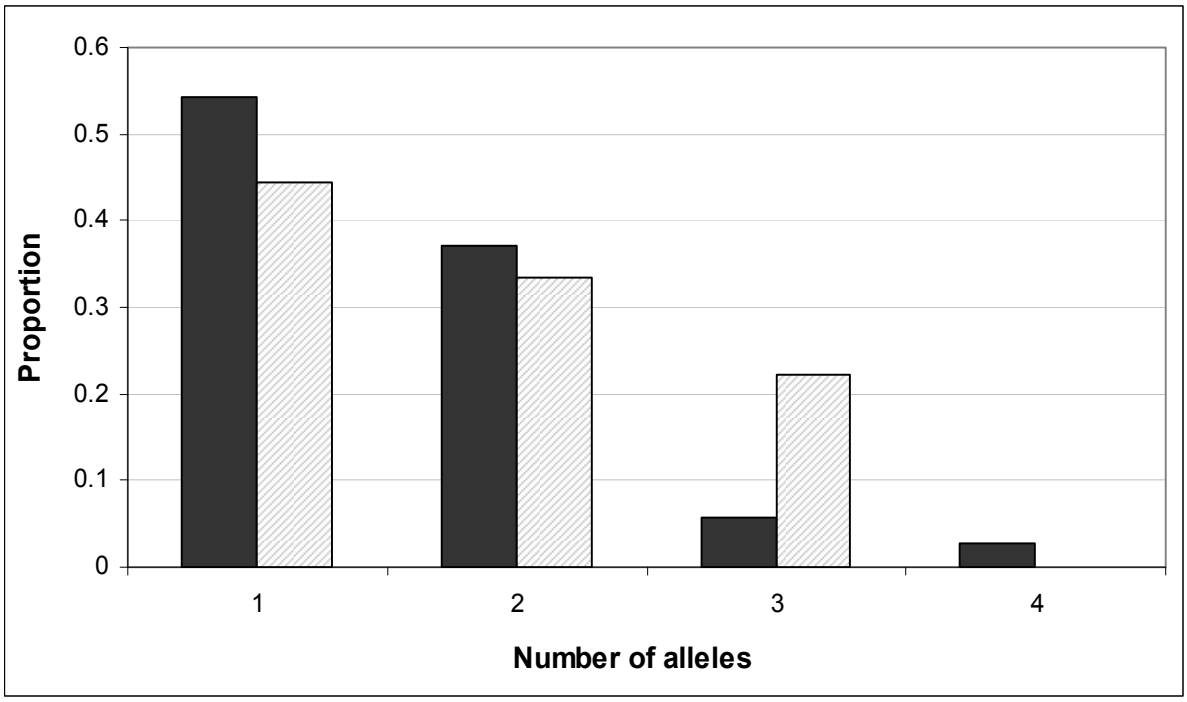

Figure 2. Frequency distribution of *DQA* copy number in Etosha versus Kruger
One to four alleles were observed in each individual: Etosha = black, Kruger = gray diagonal. Contingency goodness-of-fit analyses revealed no significant difference between population frequency distributions ($\chi^2 = 5.578$, $p = 0.1341$).

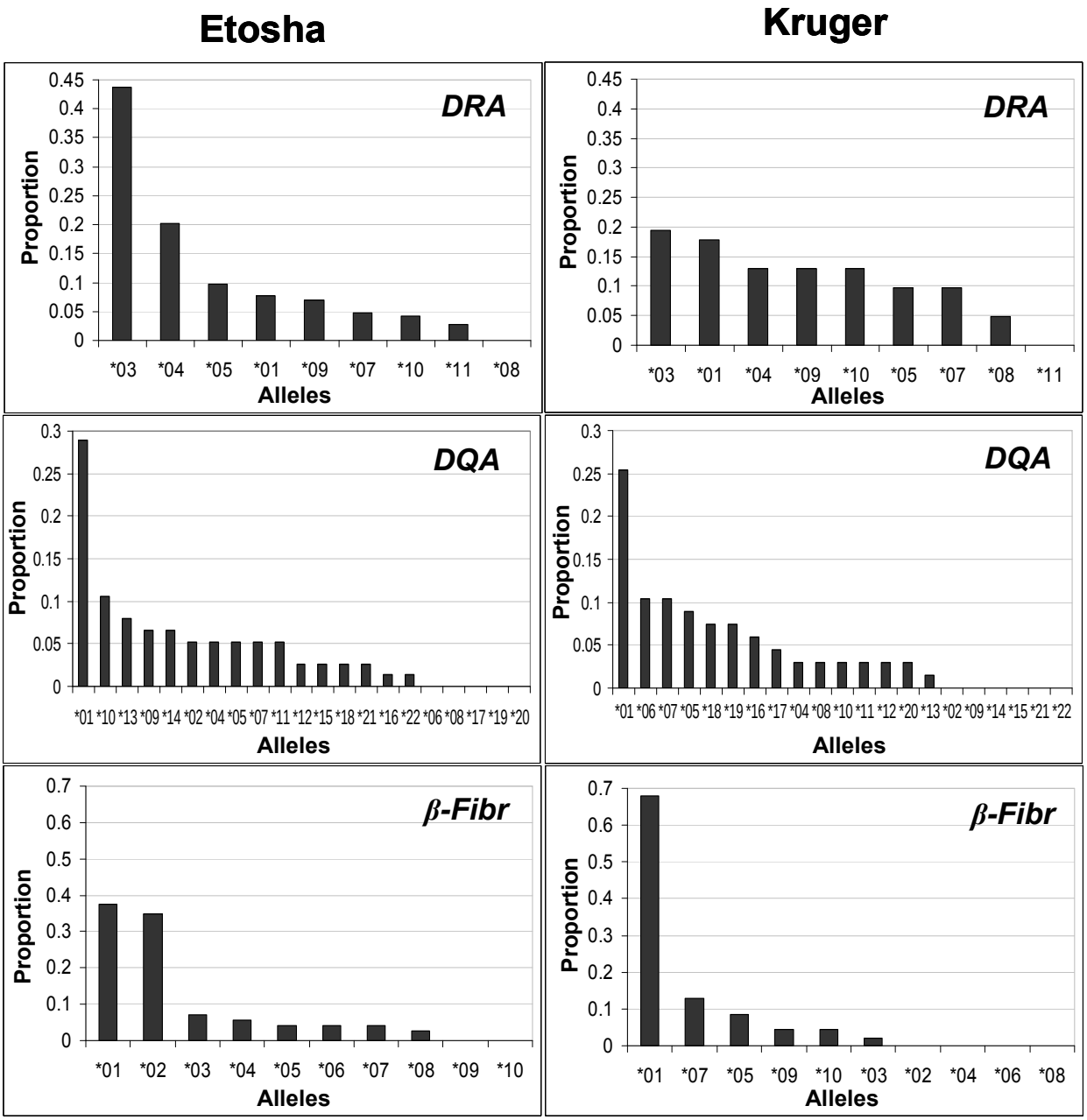

Figure 3. Allele frequency distributions by locus (*DRA, DQA, β-Fibr*) and zebra population (Etosha, Kruger)

Alleles in each graph arranged in order of descending population frequency. *DQA* allele frequency calculated under the assumption that not all alleles per individual were observed, as individuals were expected to possess two *DQA* gene copies. Thus, *DQA* allele frequency was calculated as the number of individuals that the allele was observed in divided by the total number of alleles.

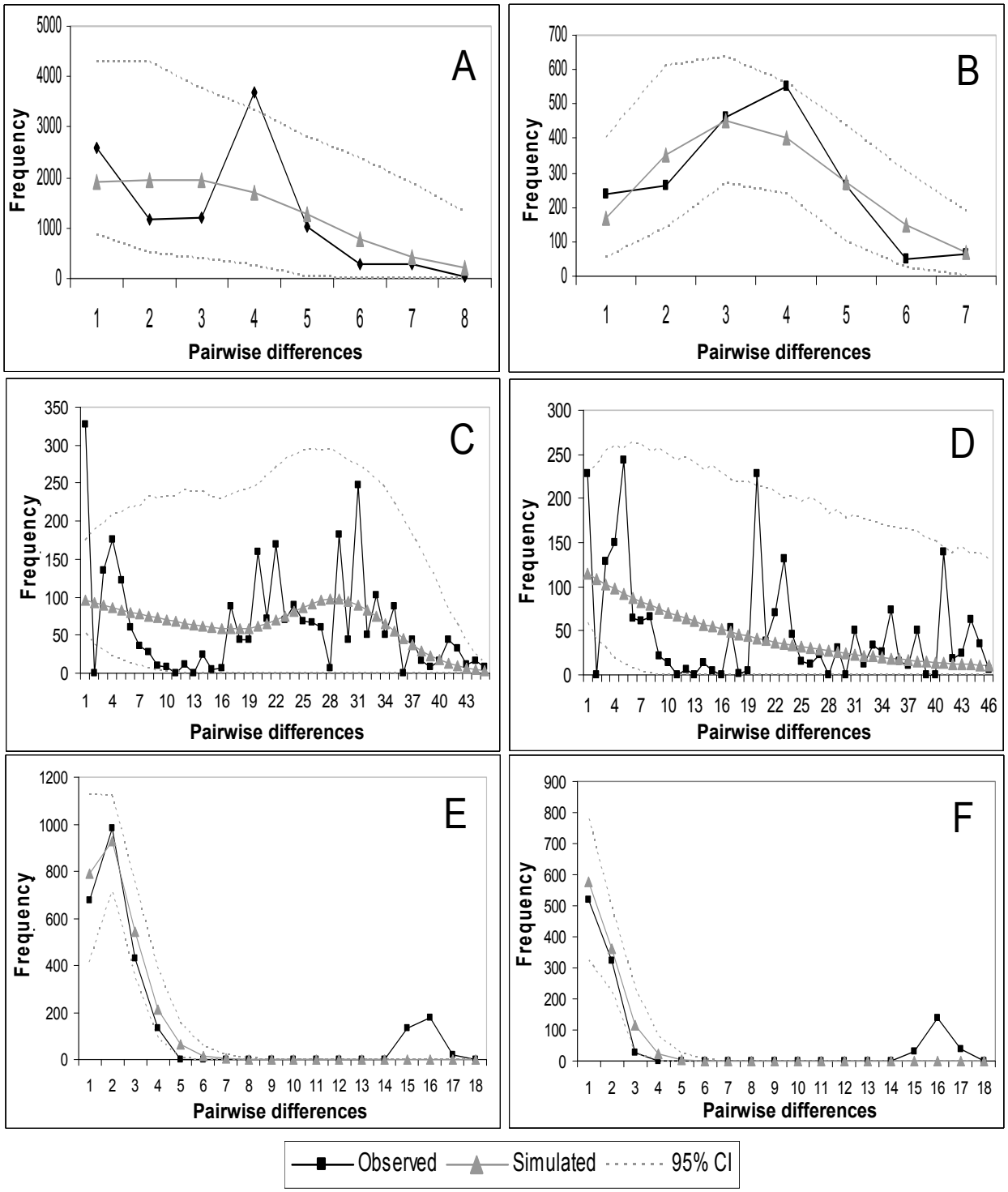

Figure 4. Mismatch distributions of functional and neutral genetic loci
DRA (A, B), *DQA* (C, D), and *β-Fibr* (E, F). Observed frequencies of pairwise nucleotide differences among sequences (black squares) and expected frequencies of pairwise nucleotide differences (gray triangles) are shown under a model of recent population expansion. Dashed lines indicate the 95% confidence interval.

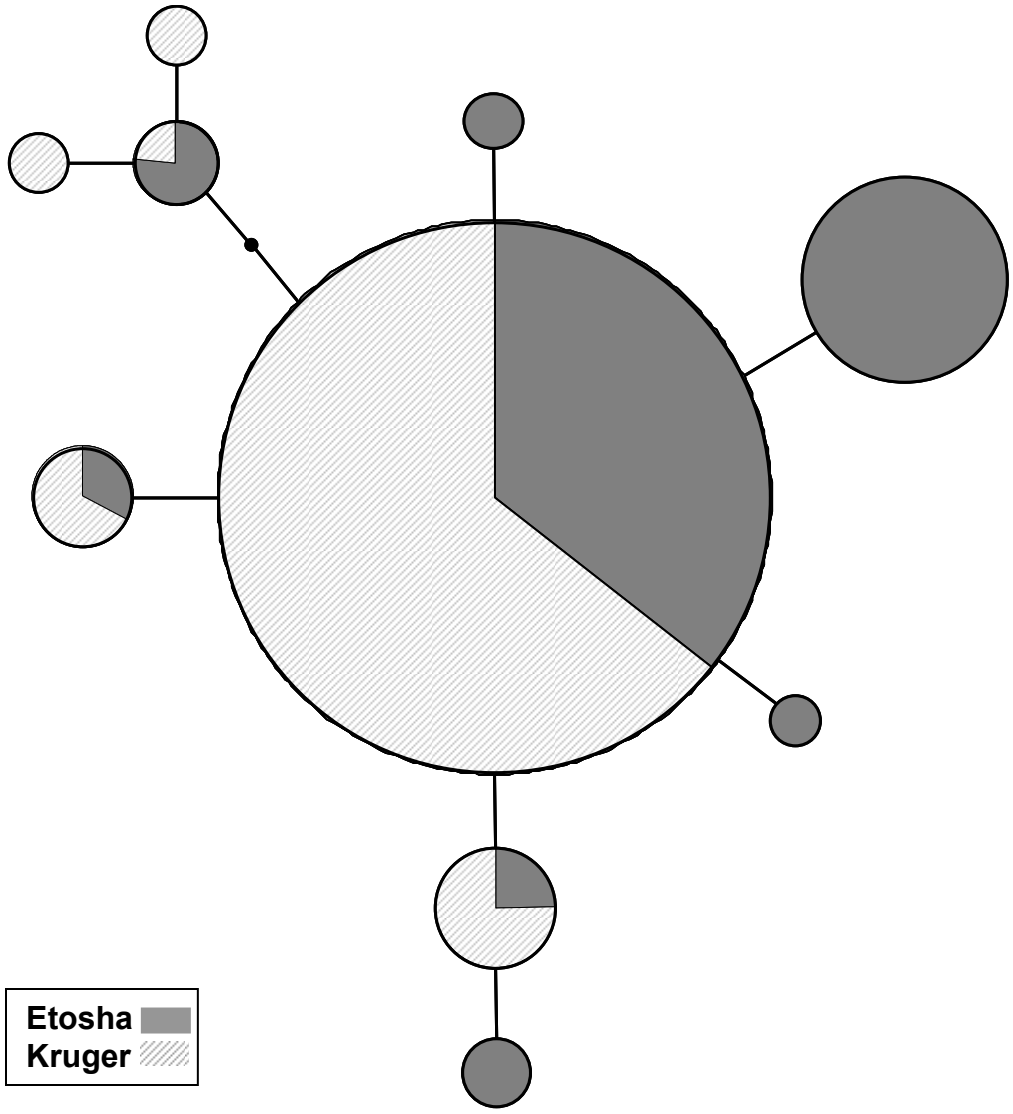

Figure 5. Allele network of the *β-Fibr* intron 7
Circle size represents allele frequency. Proportion of alleles found in Etosha (gray) and Kruger (diagonal) are shown, with sample sizes corrected by rarefaction. Lines reflect phylogenetic distance between alleles, with a putative allele that was not sampled represented by the black dot.

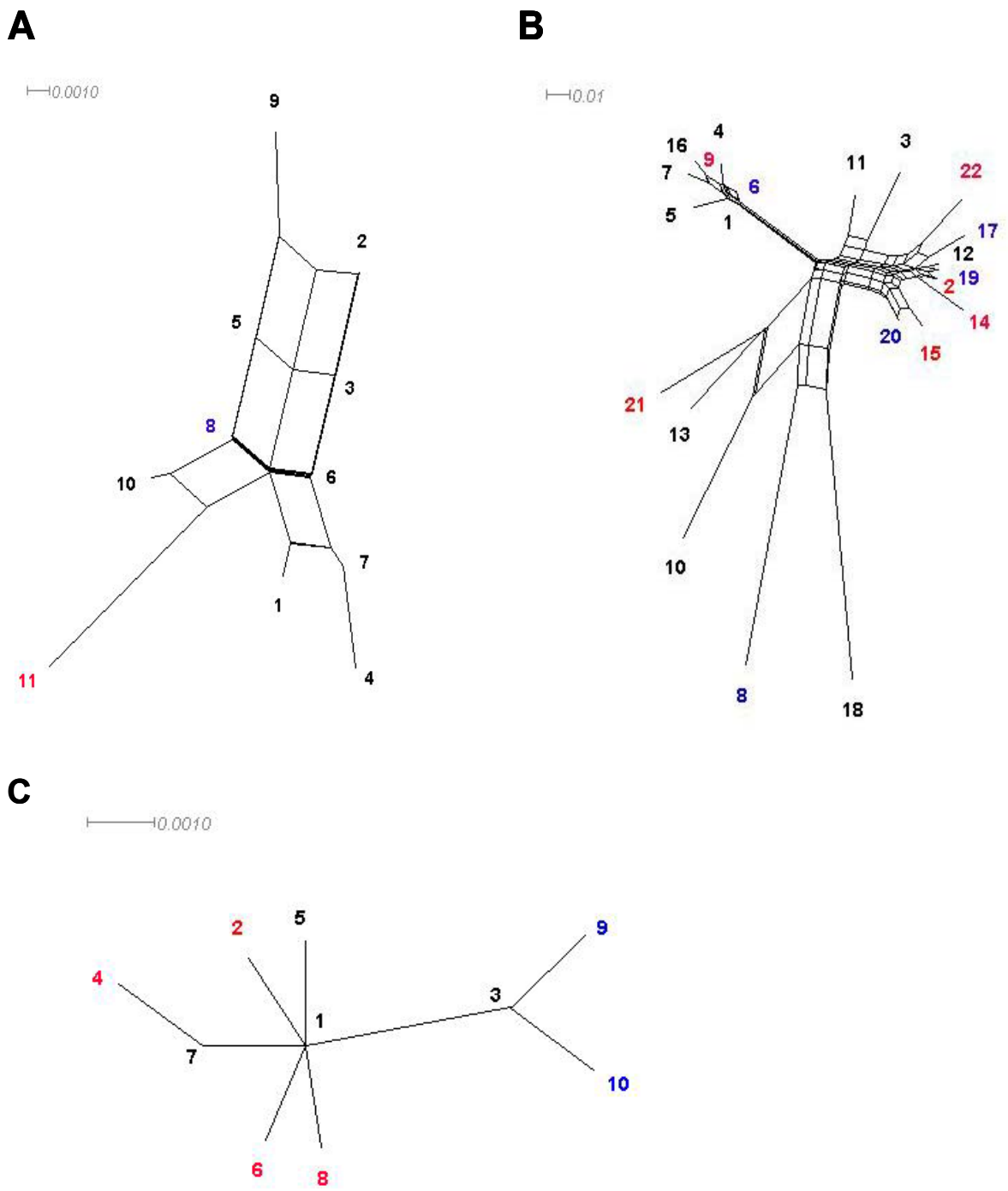

Figure 6. Maximum parsimony split networks for MHC loci, (A) *DRA*, (B) *DQA* and (C) *β- Fibr* intron in Etosha and Kruger zebra populations

Networks constructed by the Neighbor-Net algorithm. Alleles only found in Etosha are shown in red, and those only in Kruger are indicated in blue. Alleles found in both populations are in black. Distances were calculated following the Kimura 3-parameter (K3P) model for nucleotide substitution. Rate variation was assumed to be equal across sites and the proportion of invariable sites (Pinvar) to be 0 and 0.9 at *β-Fibr* and *DRA,* respectively. At the *DQA*, rate variation was gamma-distributed (α = 0.4918) with Pinvar = 0.

Supplementary materials

Table S1. Sample size and allelic richness of zebra genotyped at neutral and functional MHC loci in Etosha versus Kruger National Park

Population	No. samples typed (µsats/ *β-Fibr*/ *DRA*/ *DQA*)	No. of Alleles			
		µsats	*β-Fibr*	*DRA*	*DQA*
Etosha	84/ 37/ 72/ 36	9.0 (0.9)	8	8	16
Kruger	38/ 24/ 31/ 30	7.1 (0.7)	6	8	15

Neutral loci analyzed included 15 microsatellite loci (µsats) and the *β-Fibrinogen* intron 7 (*β-Fibr*) locus. MHC loci analyzed included the ELA-*DRA* (*DRA*) and ELA-*DQA* exon 2 (*DQA*) loci. The number of alleles found in each population is reported, with that for microsatellites reported as the mean (standard error) of 15 loci.

PCR protocols:

β-Fibrinogen, intron 7.

PCR mixes contained 1.75uL of 10x PCR buffer (100mM Tris-Cl, pH 8.3, 500mM KCl, 15mM $MgCl_2$, 0.01% (w/v) gelatin) 0.4 mM deoxynucleotide triphosphates (dNTPs), 1U Ampli*Taq* Gold DNA polymerase (Applied Biosystems), 15µg bovine serum albumin (BSA) (New England BioLabs), 0.67 µM of each primer, and 25-50ng of DNA in a 15µL total reaction volume. The following thermocycling conditions were used: an initial denaturation at 95ºC for 6 min; 48 cycles of 94ºC for 1 min, 62ºC for 1 min 30 s, and 72ºC for 2 min; final extension at 72ºC for 10 min.

Microsatellite loci.

Each locus was amplified by PCR in a 10 µL total reaction volume, comprising of approximately 25ng DNA, 1.33 µL GeneAmp 10x PCR buffer, 0.7 U Ampli*Taq* Gold DNA polymerase (Applied Biosystems), 0.4 mM dNTPs, 10µg BSA (New England BioLabs) and 0.50 µM of each primer. PCRs for Aht21, Coro14, Hmb1, Hms7, Htg15, and Lex52 were carried out under a "touch-down" thermocycling profile: initial denaturation at 95ºC for 10 min; 2 cycles of 94ºC for 1 min, 60ºC for 1 min, and 70ºC for 35 s; 18 cycles of 93ºC for 45 s, 59ºC for 45 s, and 70ºC for 45 s, with the annealing temperature decreasing by 0.5ºC with each cycle; 30 cycles of 92ºC for 30 s, 50ºC for 30 s, and 70ºC for 1 min; final extension at 72ºC for 10 min. Whereas, PCRs for the remaining microsatellite loci used the following profile: initial denaturation at 95ºC for 6 min; 40 cycles of 94ºC for 1 min, 55-60ºC for 1 min, and 72ºC for 45 s; final extension at 72ºC for 5 min. (Table S2).

Table S2. Description of microsatellite loci

Locus	No. of Alleles	Range (bp)	Flourescent Label	PCR Annealing Temp. (°C)
Aht21	10	196-214	6-FAM	TD*
Asb23	19	150-199	NED	60
Cor014	14	134-165	HEX	TD*
Hmb1	9	90-127	PET	TD*
Hms7	9	161-184	NED	TD*
Htg7	15	122-153	6-FAM	TD*
Htg9	12	114-140	PET	55
Htg14	7	135-149	HEX	60
Htg15	5	129-142	HEX	TD*
Lex20	11	197-220	6-FAM	60
Lex33	3	163-177	NED	60
Lex52	5	195-203	6-FAM	60
Ucdeq505	11	155-176	6-FAM	55
Um011	15	147-177	HEX	57
Vhl47	5	123-133	HEX	60

Includes total number of alleles observed size range in base pairs (bp), fluorescent label used for genotyping analyses, annealing temperature used in polymerase chain reaction (PCR).
*TD= “touch-down” thermocycling profile was used. This involved a shifting annealing temperature from 60°C to 50°C (see description in text).

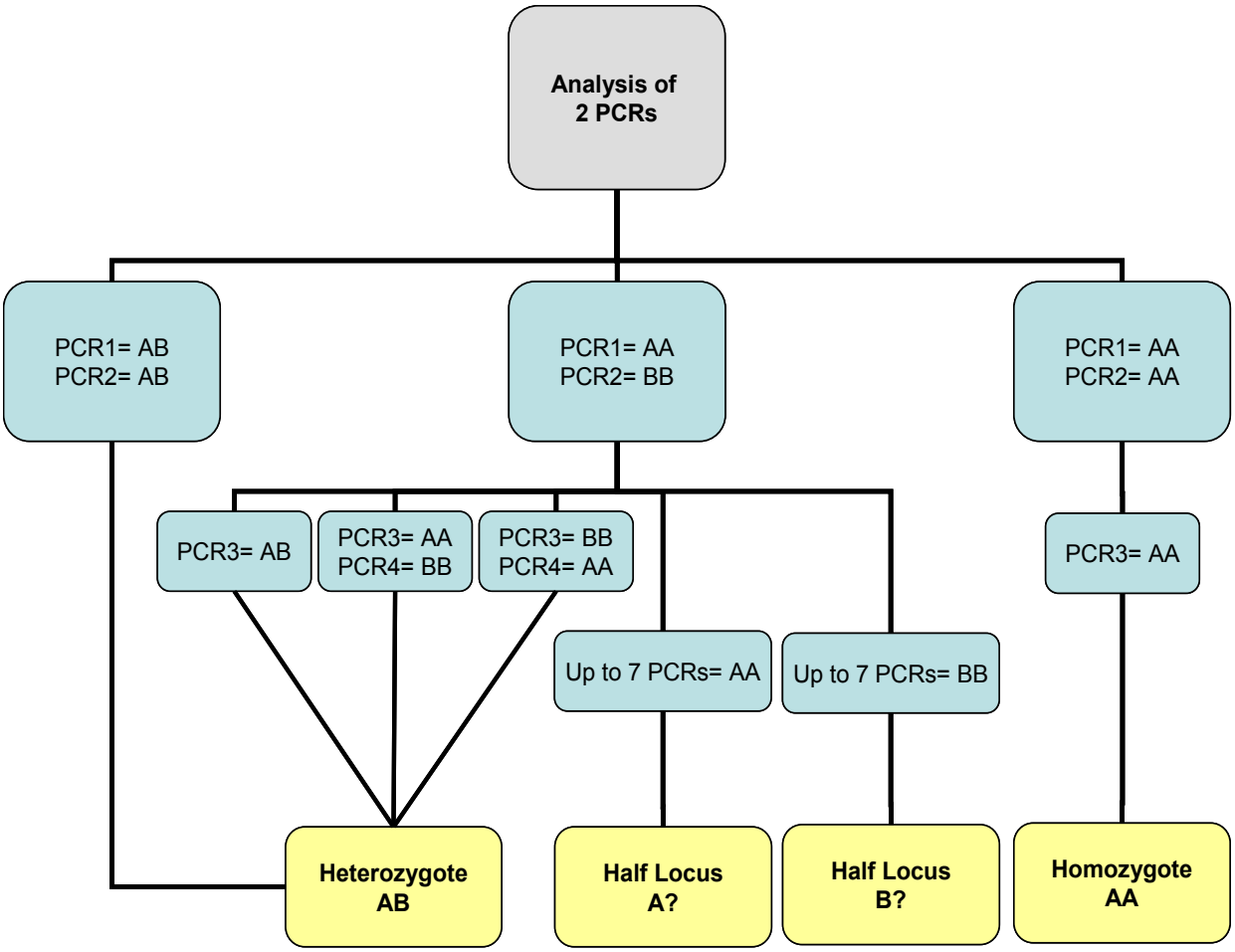

Figure S1. Flowchart of comparative microsatellite genotyping approach
Adapted from Hansen *et al*., 2008. This approach involves comparing two to three initial replicate PCRs, with heterozygotes confirmed in two and homozygotes in three PCRs. If a disagreement is found (e.g. first PCR results in heterozygote and second in homozygote for one allele) additional PCRs were performed until each allele is observed a minimum of two times. In the event that no consensus was found, an individual was either scored as having a missing genotype or given a half-locus genotype, by assigning one allele as missing data. Therefore, in summary a minimum of 2 PCRs is required to confirm a heterozygote genotype and 3 PCRs for a homozygote, with a maximum of up to 7 PCRs conducted.

Table S3. Microsatellite diversity by locus and population

	Etosha					Kruger				
Locus	*N*	*A*	H_O	H_E	F_{IS}	*N*	*A*	H_O	H_E	F_{IS}
Aht21	82	9	0.744	0.833	0.107	29	7	0.690	0.796	0.134
Asb23	82	15	0.841	0.887	0.051	14	12	0.643	0.862	0.254
Coro14	75	11	0.813	0.818	0.005	30	11	0.900	0.848	-0.062
Hmb1	79	8	0.658	0.675	0.025	37	6	0.811	0.763	-0.063
Hms7	79	8	**0.595***	0.774	0.231	34	7	0.824	0.799	-0.031
Htg7	82	13	0.866	0.861	-0.006	35	10	0.857	0.791	-0.084
Htg9	84	11	0.845	0.865	0.023	38	10	0.868	0.857	-0.013
Htg14	84	7	0.833	0.761	-0.095	35	7	0.714	0.819	0.128
Htg15	83	4	**0.446***	0.651	0.315	34	5	0.912	0.689	-0.323
Lex20	81	11	0.765	0.821	0.068	29	7	0.724	0.829	0.126
Lex33	78	3	0.615	0.605	-0.018	11	3	0.455	0.376	-0.209
Lex52	75	5	0.693	0.740	0.064	34	4	0.765	0.659	-0.160
Ucdeq505	81	11	0.802	0.765	-0.049	36	5	0.694	0.752	0.076
Um011	80	14	0.875	0.865	-0.012	34	9	0.706	0.730	0.033
Vhl47	79	5	0.519	0.547	0.051	37	4	0.459	0.404	-0.136
MEAN	*80.27*	*9.00*	*0.76*	*0.76*	*0.05*	*31.13*	*7.13*	*0.73*	*0.73*	-0.02
SE	*0.73*	*0.93*	*0.03*	*0.03*	*0.03*	*2.08*	*0.71*	*0.04*	*0.04*	0.04

For each population (Etosha and Kruger), the diversity by locus and total mean diversity are reported. Diversity is described in terms of number of alleles (*A*), observed heterozygosity (H_O), and expected heterozygosity *(*H_E). *N* = sample size analyzed for each locus. Significant departures from Hardy-Weinberg equilibrium indicated in bold.
* $p < 0.01$

Table S4. Microsatellite null allele frequency and genotyping error rates

		Blood			*Fecal*		
Locus	**NAF (FreeNA)**	**Dropout**	**False Alleles**	**Total**	**Dropout**	**False Alleles**	**Total**
Aht21	0.055	0	0	0	0	0.048	0.048
Asb23	0.038	0	0	0	0.048	0.024	0.071
Cor014	0.006	0	0.024	0.024	0	0.071	0.071
Hmb1	0.013	0	0	0	0	0	0
Hms7	0.068	0	0	0	0	0	0
Htg7	0.011	0	0	0	0	0.024	0.024
Htg9	0.024	0	0	0	0	0.024	0.024
Htg14	0.003	0	0	0	0.024	0	0.024
Htg15	0.054	0	0.024	0.024	0	0.071	0.071
Lex20	0.039	0	0	0	0	0	0
Lex33	0.000	0	0	0	0	0	0
Lex52	0.018	0	0	0	0	0.048	0.048
Ucdeq505	0.001	0	0	0	0	0	0
Um011	0.000	0	0	0	0	0	0
Vhl47	0.010	0	0	0	0	0	0
Mean	0.023	0.000	0.003	0.003	0.005	0.021	0.025

Null allele frequency (NAF) was estimated using FreeNA (Chapuis & Estoup 2007). Total genotyping error, broken down by allelic dropout and false allele generation, was calculated by paired genotyping of blood and fecal samples from individual zebra captures. All rates marked as "0" represent values of 0.000.

Chapter 4. Parasite-mediated selection drives an immunogenetic tradeoff in plains zebra (*Equus quagga*) of Etosha National Park, Namibia

"It takes all the running you can do, to keep in the same place."
— Lewis Carroll, Through the Looking Glass

Abstract

Pathogen evasion of recognition by the host immune system is believed to be the primary force driving extreme polymorphism in genes of the Major Histocompatibility Complex (MHC). Although this gene family is well characterized in structure and function, there is still much debate surrounding the mechanisms by which MHC diversity is selectively maintained. Over the past decade, many studies have investigated the relationships between MHC variation and specific pathogens in natural populations and have found variable evidence supporting as well as rejecting the primary hypotheses of heterozygote advantage or frequency-dependence, whereas other studies have asserted that these mechanisms may not be mutually exclusive. Few studies, however, have focused on the selective effects of multiple parasite types on wildlife host immunogenetic patterns. In this study, I examined the relationships between variation in the equine MHC gene, ELA-*DRA*, and both gastrointestinal (GI) and ectoparasite intensity in a free-ranging plains zebra (*Equus quagga*) population. I found that specific *DRA* alleles present at opposite frequencies in the population had antagonistic effects, such that rare alleles were associated with increased GI parasitism and, in contrast, common alleles predicted higher tick burdens. These results are in support of a frequency-dependent selection process and suggest that GI parasites exert strong selective pressure at this locus, based on the observation that maladaptive 'susceptibility alleles' are reduced in population frequency. Heterozygote advantage, in terms of genotypic allele divergence, was also discovered to play a selective role in decreasing GI parasite burden, but in conjunction with the presence of common allele. I hypothesize that a potential immunogenetic tradeoff modulates resistance/susceptibility to parasites in this system, such that with MHC-based resistance to GI parasitism, a fitness cost is incurred to the host in the form of increased ectoparasite susceptibility. This study highlights the importance of investigating the role of multiple parasites in shaping patterns of MHC variation.

Introduction

Pathogens are known to exert strong selective pressures on the fitness of natural host populations. While pathogens are constantly evolving increased virulence and host recognition avoidance mechanisms, hosts are challenged to respond by evolving a diversity of innate and adaptive immune defenses (reviewed in Slev & Potts 2002). This ongoing evolutionary "arms-race" (Van Valen 1973) is known to influence the molecular diversity in both pathogen and host genomes, particularly for immunological genes (Frank 2000). In evolutionary ecology, there has been increasing focus on this selective molecular interplay in wildlife populations (Acevedo-Whitehouse & Cunningham 2006), that has come with recognizing the importance of immune system function not only to emerging infectious disease and host-parasite co-evolution, but also population dynamics and life history strategies (Daszak *et al.* 2000; Morens *et al.* 2004).

The Major Histocompatibility Complex (MHC), a gene family comprised of immune-related genes, has been a particular focus in evolutionary ecology due to its incredible diversity and significance in mate choice, kin recognition and host immunity in vertebrates (reviewed in Edwards & Hedrick 1998). This gene complex encodes the molecules responsible for initiating host immune response, by delivering foreign peptides derived from pathogens to helper T-cells. Given this fundamental role in immune function, it has been established that observations of extreme polymorphism at the MHC are the direct result of balancing selection by pathogens (Doherty & Zinkernagel 1975; Hedrick 2002; Bernatchez & Landry 2003). Support for this hypothesis has been shown in a global study on humans that found populations with increased pathogen diversity to also have elevated MHC diversity (Prugnolle *et al.* 2005). In addition, associations between specific MHC alleles and pathogen resistance have been demonstrated in model systems (Penn *et al.* 2002; McClelland *et al.* 2003) as well as in humans (Hill *et al.* 1991), non-human primates (Schad *et al.* 2005), ungulates (Paterson *et al.* 1998), rodents (Froeschke & Sommer 2005; Meyer-Lucht & Sommer 2005) and fish (Langefors *et al.* 2001). Furthermore, balancing selection at MHC loci has been inferred from evidence of lower MHC differentiation among populations (e.g. Mona *et al.* 2008) and more evenly distributed allele frequencies within populations than would be expected under neutral theory (Takahata & Nei 1990).

Despite being well-studied, understanding the mechanism by which pathogens drive diversity in MHC genes has been the subject of much debate (reviewed in Sommer 2005). Three hypotheses have been primarily considered: (i) overdominant selection (Doherty & Zinkernagel 1975), (ii) negative frequency-dependent selection (Takahata & Nei 1990); (iii) and fluctuating selection over time and space (Hedrick 2002). Overdominant selection, or heterozygote advantage, is based on the assumption that heterozygous individuals are capable of recognizing a more diverse suite of pathogens (Doherty & Zinkernagel 1975; Hughes & Nei 1989) and has been convincingly supported by empirical evidence from multiple parasite co-infection studies (Penn *et al.* 2002; McClelland *et al.* 2003) as well as in natural systems (Richman *et al.* 2001; Evans & Neff 2009). Negative frequency-dependent selection, also known as rare-allele advantage, suggests that the advantage of a specific allele varies with its frequency in the population as the result of pathogen evasion, such that common alleles are selected against and rare alleles for (Takahata & Nei 1990). This mode of selection has been particularly difficult

to directly demonstrate in natural populations given the long time scales necessary for detecting allele frequency changes. Although, one study used genetic time-series data to demonstrate that MHC allele frequencies fluctuated over time in cohorts of Great reed warblers (*Acrocephalus arundinaceus*) (Westerdahl *et al.* 2004), but did not link their results to any parasite driving factors. Despite this challenge, however, some have argued that observations that specific MHC alleles are associated with pathogen resistance provides indirect evidence for a frequency-dependent mechanism (Froeschke & Sommer 2005; Croisetiere *et al.* 2008). Finally, under the fluctuating selection hypothesis, host-pathogen interactions are expected to vary over time and space (Hedrick 2002). This model exerts the idea that allele fitness is spatiotemporally heterogeneous, a phenomenon that has been theoretically shown to be capable of maintaining high levels of MHC polymorphism (Hedrick 2002). This has been supported by population studies finding spatial and/or temporal divergence at MHC discordant with patterns at neutral loci (e.g. Miller *et al.* 2001; Charbonnel & Pemberton 2005). Fluctuating selection is a more probable working hypothesis given that natural populations are exposed to changing environmental conditions which may consequently lead to oscillations in pathogen pressures.

The MHC remains an important model through which to test hypotheses regarding the influence of pathogen-mediated selection on host molecular diversity and, ultimately, disease susceptibility. However, resolving the relative importance of these aforementioned hypotheses in any given study system has been suggested to be a difficult task (Spurgin & Richardson 2010). One challenge is that these hypotheses may not be mutually exclusive (see Apanius *et al.* 1997). For example, it is also possible that heterozygotes may have higher fitness than homozygotes on average in a given population due to 'dominance' effects of resistance alleles (e.g. Penn *et al.* 2002), but that these few advantageous alleles may fluctuate in accordance with the relative fitness they confer, thereby confounding conclusions of heterozygote advantage and frequency-dependence. Also, heterozygote superiority may exist, but depend on the degree of molecular divergence at overlapping peptide biding regions (e.g. Richman *et al.* 2001) which is also known as the 'divergent allele advantage' hypothesis (Wakeland *et al.* 1990). Another problem often encountered is that the allele frequency distributions of both overdominance and frequency-dependence are predicted to be theoretically indistinguishable, although a simulation study recently showed that distributions under negative frequency-dependence were much less stable (Ejsmond *et al.* 2010).

This may be even further complicated by antagonistic pleiotropic effects of multiple parasites on a single host locus, such that resistance to different pathogen types requires different MHC alleles or genotypes. With these effects, increased resistance to one pathogen may come with a cost of decrease resistance to another, resulting in the rather perplexing finding of associations between MHC alleles and increased susceptibility to infection (e.g. Langefors *et al.* 2001; Froeschke & Sommer 2005; Schad *et al.* 2005; Croisetiere *et al.* 2008; Loiseau *et al.* 2008; Evans & Neff 2009). The majority of studies to date have focused on gastrointestinal parasites when investigating MHC-pathogen relationships in wild vertebrate populations (Paterson *et al.* 1998; Froeschke & Sommer 2005; Meyer-Lucht & Sommer 2005; Schad *et al.* 2005; Oppelt *et al.* 2010). Others have also looked at ectoparasite (Oliver *et al.* 2009b), bacterial (Evans & Neff 2009), protist (Loiseau *et al.* 2008) and viral infections (Deter *et al.* 2008), but few have investigated

the relationships of multiple parasite types with MHC diversity (but see Ditchkoff *et al.* 2005; Deter *et al.* 2008). Given that multiple parasites may act in concert or opposition to mediate selection on the host MHC, incorporating knowledge about different types affecting hosts may lead to conclusions than would have otherwise gone undetected regarding selection on the MHC. In this study, I investigated the relationships between multiple parasite groups and MHC diversity in a wildlife population, to elucidate the mechanisms by which immunogenetic variation is being selectively driven and maintained.

Study system background

The plains zebra (*Equus quagga*) population of Etosha National Park (ENP), Namibia, provides an excellent natural system in which to elucidate the mechanisms by which pathogens drive variation in the MHC. ENP is considered to be a eutrophic savanna ecosystem (Huntley 1982) with low annual precipitation (rainfall < 650 mm/year). This relative aridity has been suggested to play a role in limiting *E. quagga* parasite diversity relative to that found in plains zebra inhabiting savanna ecosystems with higher annual rainfall (e.g. Kruger National Park—KNP, South Africa) (Matthee *et al.* 2004). However, zebra of ENP have been shown to be particularly susceptible to gastrointestinal (GI) nematodes, with nearly all individuals in the population found with nematode infections (Turner & Getz 2010). Additionally, I note that hard-bodied ticks were observed on most individuals examined in the field, although the ecology of the host-ectoparasite relationship has yet to be characterized in this system.

Gastrointestinal nematodes have been shown to have negative fitness consequences in wild sheep populations (Gulland 1992; Gulland & Fox 1992; Paterson *et al.* 1998) and associations between the MHC and these parasites, as mentioned previously, have been demonstrated. Similarly, the vertebrate immune system is known to be key in responding to ectoparasite infections; for example, host antibodies can bind proteins in tick saliva, interfering with effective tick engorgement and nutrient absorption, thereby inhibiting ova production and viability (Trager 1939; Wikel 1996; Proctor & Owens 2000). Additionally, immune-triggered inflammatory response and increase in host skin temperature may result in tick detachment (Allen & Kemp 1982). Several studies have also found associations between the MHC and ectoparasite prevalence and intensities (Acosta-Rodrigiez *et al.* 2005; Ditchkoff *et al.* 2005; Oliver *et al.* 2009b). To my knowledge, only one natural study has investigated host immunogenetic relationships with both of these parasite types concurrently (see Ditchkoff et al. 2005). Here I take advantage of the low diversity, but high prevalence of both parasite groups in E. quagga of ENP, both of which are expected to exert strong selective pressure on the host immune system, to address the selective mechanisms shaping diversity at the MHC.

The equid MHC complex, or Equine Lymphocyte Antigen (ELA), has been molecularly and structurally characterized (Albright-Fraser *et al.* 1996; Bailey *et al.* 2000; Brown *et al.* 2004) and previous work found evidence for selection on the antigen binding region of class II MHC loci among equids (Janova *et al.* 2009; Kamath & Getz 2011), as well as among *E. quagga* populations of southern Africa (see Chapter 3). The ELA-*DRA* locus is one of the class II alpha chain paralogues and of particular importance

as it encodes the critical antigen binding domain responsible for recognition of foreign peptides. Interestingly, this genetic locus in vertebrates is considered to be much less diverse relative to other classical MHC genes, for example in humans (Chu *et al.* 1994), dogs (Wagner *et al.* 1999), cats (Yuhki & O'Brien 1997), and goats (Takada *et al.* 1998). In contrast to this lack of diversity in most taxonomic groups, recent evidence has proven equids to be an exception, exhibiting remarkably high levels of *DRA* polymorphism (Albright-Fraser *et al.* 1996; Brown *et al.* 2004; Janova *et al.* 2009; Kamath & Getz 2011). This finding has been corroborated in *E. quagga* populations of southern Africa where even greater allelic diversity was discovered (see Chapter 3), implying the need for further examination of *DRA* polymorphism in natural populations. This study also revealed low MHC differentiation among populations, typical of balancing selection acting to maintain diversity, but also a highly skewed *DRA* allele frequency distribution suggestive of pathogen-driven directional selection in ENP that is not evident in KNP (Figure 1; see also Chapter 3).

In this study I investigated the relationships between *ELA-DRA* variation and parasitism in the *E. quagga* population of ENP. I specifically focused on the immunogenetic effects of *DRA* heterozygosity and specific allele/genotype effects in predicting the burdens of both GI and ectoparasites in zebra, while controlling for potentially confounding ecological and demographic predictors that have been shown to influence parasitism in ungulate hosts of this system (see Turner & Getz 2010). My null hypothesis was that immunogenetic variables do not influence parasite burden, and thus do not significantly improve models incorporating only demographic and ecological factors. I evaluated this null against the following alternative hypotheses: (i) 'heterozygote advantage' in terms of the general *DRA* heterozygote fitness in the population; (ii) 'heterozygote advantage' under the divergent allele advantage hypothesis, assuming that heterozygote fitness increases with the number of heterozygous base positions present in an individual's *DRA* genotype; (iii) frequency-dependence in terms of the non-additive (or dominance) effects of alleles of a particular frequency class; and (iv) frequency-dependence through examination of specific single allele effects. I also considered that the hypotheses of frequency-dependence and heterozygote advantage may not be mutually exclusive by allowing for interactions of these effects in the models evaluated. Finally, I examined whether these variables predict multiple types of parasitism (GI nematodes and ectoparasite ticks) to elucidate possible pleiotropic effects of the MHC in modulating parasite resistance.

Methods

Study population

This study focused on a natural population of plains zebra (*E. quagga*) in ENP, a large (22,915 km^2) fully-enclosed nature reserve in northern Namibia. The *E. quagga* population in the park is considered to be relatively abundant, with an estimated size of approximately 13,000 (95% CI: 10,900 – 15,000) (Namibian Ministry of the Environment and Tourism, unpublished data) and previous demographic analyses based on genetic data have predicted that the population has been both recently and historically stable (see Chapter 3). ENP is classified as semi-arid mopane savanna (Huntley 1982)

and has a unique salt pan that makes up approximately one fifth of the entire park area. Rainfall generally exhibits a seasonal pattern with the majority of rain falling between the months of October to April. Rainfall data from 2007 to 2010, encompassing this study period, revealed that the mean (± standard error) annual rainfall at the Okaukuejo weather station was 461 mm (± 89 mm), and at the Halali weather station was 570 mm (± 39 mm). Also, rainfall patterns were consistent with past weather records for this area, showing a rainfall peak from October to April (Figure 2).

Data collection

Data were collected from adult zebras (n = 65: females = 55, males = 10) during a series of capture events that took place between March 2008 to August 2010 on the Okaukuejo and Halali central plains region of ENP (Figure 3). Zebra captures were conducted during five distinct periods that covered three rainy and dry seasons. During captures, zebras were anesthetized, allowing for various samples to be collected (Protocol #R217-0510B, approved by the Animal Care and Use Committee, UC Berkeley). Additionally, individuals were VHF or GPS-collared which enabled a subset to be re-captured, with between one to five captures per individual. I collected fecal samples for gastrointestinal (GI) parasite quantification (n = 140 individual zebras) and ectoparasites were picked directly from the animal (n = 119 observations). Of these sampling observations, I was able to collect both GI parasites in tandem with ectoparasites for only a portion of the data (n = 87). Age was estimated from wear patterns of permanent incisors, using the approach described by Smuts (1974) based on the size and shape of the infundibulum and the grinding surface of the upper incisors (Smuts 1974), based on dental wear patterns due to the chewing force exerted over an individual's lifetime. Finally, I also collected blood samples for immunogenetic characterization. Blood samples were collected in ethylenediaminetetraacetic acid (EDTA) tubes to minimize coagulation and kept at -20°C for long term preservation of DNA. Both blood and fecal samples were kept on ice in the field and then at 4°C for storage for short-term preservation. Ticks were stored in 70% ethanol.

Quantification of parasite burden

Gastrointestinal parasites.

Gastrointestinal (GI) parasite burden was measured in terms of fecal egg counts (FECs; eggs per gram of feces) following a modification of the McMaster flotation procedure (Food and Agricultural Organization Gordon & Whitlock 1939; Food 2005). This approach provided an appropriate non-invasive means for quantifying relative parasite burdens among individuals, and has been proven to be valuable for assessing parasitism in wild ungulate hosts (e.g. Paterson et al. 1998; Coltman et al. 1999; Cassinello et al. 2001; Schwensow et al. 2007). The modified protocol used here has been previously described in detail by Turner et al., 2010. In summary: four grams of fecal material were suspended in 56mL of a saturated salt (NaCl) solution, large plant debris was filtered and the remaining filtrate was aliquoted onto two chambers of a McMaster slide. The salt solution resulted in flotation of strongyle nematode eggs which allowed for subsequent quantification using a compound microscope. The eggs observed in the two slide chambers were counted, summed and multiplied by 50 to yield a measurement of eggs per gram of feces. All parasitological assessment took place within 48 hours

following collection to ensure fresh samples for accurate quantification and identification.

Previous work assessed GI parasites in zebra of ENP and revealed nearly 100% prevalence of strongyle nematodes from the order Strongylida (Turner & Getz 2010). In concordance with these findings, all zebra sampled in this study were infected by strongyle nematodes. The implemented McMaster flotation technique is known to be limited in identifying all GI parasite groups that may be present, specifically being incapable of detecting trematode eggs or nematode larvae (Bowman 2003). Furthermore, microscopic observation did not allow for further accurate taxonomic resolution of nematodes beyond the order level. Thus, I conservatively refrained from more specific classification. However, strongyle nematodes of equid hosts have been shown to be generally of the family Strongylidae, comprised of the subfamily Strongylinae ("large strongyles") and Cyanthostominae ("small strongyles"). Furthermore, Krecek et al., 1987, performed necropsies on plains zebra of ENP and found predominantly Strongylidae nematodes, identifying 15 species of this family (Krecek *et al.* 1987b). Also, a comparative study of intestinal helminth parasites in African equids found no single species infections, with a minimum of five strongylid species present in any given individual (Matthee *et al.* 2004). These previous findings suggest that it is likely the nematodes found in this study also fall within this taxonomic grouping and represent multiple strongylid species.

In domestic equids, Strongylidae nematodes primarily inhabit the large intestine and some species have been found to cause significant damage to the intestinal mucosa and arterial system (Bowman 2003). In particular, some members of the Strongylinae subfamily, as adult worms, are bloodsuckers and their larvae are also capable of migrating through host tissues, sometimes occluding small arteries and, thus, potentially causing arteritis, thrombosis, embolism and fatal infarction of the bowel (Bowman 2003). While less destructive, the more numerous Cyanthostominae nematodes are also known to cause lesions in the cecum and colon and negative fitness consequences can include persistent diarrhea, emaciation and consequent protein deficiency. In a wild sheep population, significant relationships were found between survivorship and nematode infections (Paterson *et al.* 1998) and post-mortem examinations following a severe population crash pointed to immunosupression due to strongyle nematodes as the culprit (Gulland 1992; Gulland & Fox 1992).

Ectoparasites.

Ectoparasite burden was assessed by inspecting zebras during capture and collecting all visible arthropods, not distinguishing between larvae, nymphs or adult stages, and focusing on the peri-anal (base of tail) and groin areas. Note that this method does not ensure complete quantification and characterization of an individual's tick burden; however, as sampling effort remained consistent and focused on the body regions known to harbor the highest densities of ticks, I believe this data provided a relative estimate of tick abundance that is comparable among individuals. Five hard bodied tick species (Family: Ixodidae) were identified from the genus *Hyalomma* and *Rhipicelphalus*: *H. rufipes, H. truncatum, R. evertsi mimeticus, R. sulcatus* and *R. turanicus* (I. Horak, *pers. comm*.). The majority of ticks observed were *R. e. mimeticus*, supporting previous research on ticks from plains zebra of this region (Horak *et al.* 1992). However, for this

study, I evaluated infection by ticks in terms of overall abundance, not distinguishing between species or life stages.

Ticks are known to be one of the chief vectors for infectious disease agents, often causing severe illness and even death to their hosts. For example, in equids, ticks are known to be able to transmit the lethal African horse sickness virus (Dardiri & Salama 1988) as well as the protozoan pathogen, babesiosis (Dwinger 1999). Beyond harboring disease agents, they can also decrease host fitness through dermatoses (inflammation, itching, swelling) and envenomization (delivered through tick saliva). In addition, the toxins transmitted can sometimes cause host paralysis. The main species observed here, *R. e. mimeticus*, is a medium-sized dark brown tick found primarily throughout Namibia in wild equids and greater kudu (Horak *et al.* 1984a; Horak *et al.* 1984b; Horak *et al.* 1992). This particular species inhabits their hosts year round and life stage activity varies seasonally ─ evidence from cattle suggests adult life stages peak from November to May, whereas immature life stages peak from both February to March and May to September (Biggs & Langenhoven 1984).

MHC genotyping

Genomic DNA was extracted from blood samples using Qiagen DNeasy extraction kits, following the manufacturer's instructions for non-nucleated blood (Qiagen, Valencia, CA). I examined the diversity of a class II MHC gene: the Equine Lymphocyte Antigen, DR alpha chain (*ELA-DRA*) exon 2 region, which contains the antigen binding sites (ABS) known for their role in foreign peptide recognition. The *DRA* provides a particularly useful study system, in comparison to many MHC loci, because it is known to have a single expressed copy, thus, individual genotypes are easily identified. In *E. quagga* of this study system, I previously found eight *DRA* alleles which exhibited a total of six single nucleotide polymorphisms (SNPs) and four amino acid replacements, two of which occur at functional ABS (see Chapter 3).

I amplified 246 bp of *DRA* exon 2 in plains zebra of ENP through polymerase chain reaction (PCR) using the equid-specific primers, Be3 and Be4 (Albright-Fraser *et al.* 1996). Individual *DRA* genotypes were determined through direct sequencing and heterozygous nucleotide positions were confirmed by sequencing in both forward and reverse directions Sequence chromatograms were aligned and edited manually using Geneious v.5 (Drummond *et al.* 2010). Detailed sequencing and PCR protocols are outlined in Kamath & Getz, 2011. Allele sequences were inferred using the haplotype phase determination algorithm implemented in PHASE v2.1 (Stephens *et al.* 2001), that has proven to be accurate in identifying alleles in highly variable MHC loci (Bos *et al.* 2007). I followed strict criteria for allele identification, setting a threshold posterior probability of 0.8 and the requirement that an allele must be observed at least twice before being considered as a "true" allele (i.e. in a minimum of one homozygote or two heterozygotes).

Statistical analyses

Statistical analyses of parasitism were conducted using generalized estimation equations (GEE), or marginal models analyzed within the generalized linear model framework (Liang & Zeger 1986). This approach is particularly useful for incorporating both random and fixed effects when model residuals violate the assumptions of

independence and normal distribution, and can account for the correlation structure in the data (Hardin & Hilbe 2003). Thus, this analytical method was appropriate for dealing with this dataset given that all parasite response variables were count data, the error structures were not normal, and repeated measurements were taken from individual zebras. Furthermore, the GEE procedure has been suggested to be valuable when the main factor of interest is between subjects (i.e. individuals) through time (Omar *et al.* 1999), as is the case here where the primary goal was to evaluate the contribution of immunogenetic factors to the variation in parasite intensity.

I specifically addressed whether there is a relationship between *DRA* genetic variation and parasitism using the following parasitological response variables: (1) GI parasite intensity defined as the estimated number of nematode eggs per gram (EPG) of feces in an infected individual (see Margolis *et al.* 1982) and (2) ectoparasite intensity in terms of the number of ticks collected from an individual (ECTO). I also tested for the effects of co-infection (by both GI and ectoparasites), specifically testing for a direct relationship between parasite types, through an independent GEE analysis that omitted the other explanatory variables of interest due to limited overlapping data from individual zebras. GEE model estimates were determined by incorporating zebra identification as a random effect that followed an exchangeable working correlation structure to account for repeated measurements from individuals (see Hardin & Hilbe 2003; Zuur *et al.* 2009). Since parasite intensity was measured as count data, a Poisson error distribution with log link function was specified; with the GEE approach, however, no specific error distribution was directly assumed.

Previous research has shown corresponding patterns in the peaks of rainfall and GI parasitism, revealing that season and age are significant predictors of gastrointestinal parasite (GI) intensity in this study population (Turner 2009; Turner & Getz 2010). Therefore, I similarly accounted for these and other potential confounding non-genetic explanatory variables, by incorporating season, sex and age in all of the statistical models. Season was defined based on rainfall observations encompassing the study time period (from 2007 to 2010), but with a one-month lag due to the previous findings that there is a time lag in parasite egg shedding behind rainfall (Turner 2009). Therefore, here, I defined the wet season as November to May, and the dry season as June to October (Figure 2). Individual age was represented as a continuous variable estimate based on incision wear patterns (Smuts 1974).

Statistical models were fit with genetic explanatory variables that would allow for testing aspects of both the heterozygote advantage and frequency-dependent selection hypotheses. Here, my definition of heterozygote advantage employed was that heterozygotes generally have higher fitness than homozygotes in the population, which has been explained to be caused by 'dominance' effects and masking of 'susceptibility alleles'. This was used in preference to the stricter individual level definition of overdominance that asserts a heterozygote has greater fitness than either homozygote of the two alleles in its genotype. This strict definition was not possible to address due to reduced sample size and statistical power given the high number of *DRA* alleles present in the population. I included heterozygosity as a binary fixed effect in the full model (i.e. heterozygote or homozygote). However, as host-parasite interactions occur at the molecular level, I also included heterozygosity as the number of single nucleotide polymorphisms (SNPs) observed in an individual to address the divergent allele

advantage hypothesis. This particular heterozygosity factor allowed for the fitness of a heterozygote to vary by the number of SNPs in its genotype. Additionally, to address whether allele frequency is related to parasite intensity, the presence or absence of rare, middle-range and common alleles in an individual's genotype were included as separate explanatory factors in these models. I defined "common" as the two most frequent alleles (>15%: *DRA*03, DRA*04)*, "rare" as the three rarest alleles (approx. <5%: *DRA*07, DRA*10, DRA*11*) and "mid-range frequency" as those alleles with frequencies falling in-between (5-10%: *DRA*01*, *DRA*09*, *DRA*05*). Allele frequencies were determined from the dataset of $n = 130$ alleles (Table 1). For a summary of both explanatory and response variables used in these statistical analyses, see Table 2.

Prior to model selection, I conducted an initial graphical exploration of the data to determine the spread and distribution of both response and explanatory variables and identify potential outlier data points. High leverage outliers were removed from use in subsequent statistical analyses. In addition, as collinearity among explanatory variables may result in unstable parameter estimates and inflated standard errors (Quinn & Keough 2002), I assessed pairwise scatterplots, correlation coefficients and variance inflation factors ($VIF_{MAX} < 3$) among the explanatory variables. As expected, I found that *DRA* heterozygosity was highly correlated with the number of SNPs. Furthermore, genotype frequency was correlated with all the allele frequency categories (rare, middle and common frequencies). Therefore, I re-fit the models by including these variables separately. This resulted in a total of four maximal starting models that included all of the non-genetic and genetic explanatory covariates to address both hypotheses of heterozygote advantage (HET or SNPS) and frequency-dependent selection (GENO.FREQ or COMM/ MID/ RARE). I also determined the best-fit null models that only included the non-genetic covariates of season, sex and age for model comparison against models including genetic terms. See Table 3 for a list of the null and maximal models assessed.

Finally, after determining the best model fit to the data, I tested the associations of specific *DRA* alleles and parasite intensities by including these as explanatory fixed effects in the models. These allelic covariates were considered as either the presence or absence of a particular allele in the *DRA* genotype. I was not able to consider additive genotype effects (or co-dominance) due to reduced statistical power derived from a large variability in sample size among specific genotypes in the sample.

As GEE is not a likelihood-based modeling method, model selection was conducted using the quasi-likelihood information criterion (QIC; Pan 2001), by choosing those with the smallest QIC values as candidate best-fit models. Given the large number of possible combinations I first evaluated the maximal models (genetic + non-genetic terms), followed by further refinement through a stepwise approach based on QIC values. Potential interactions between genetic variables were considered during this process, to account for the combined effects of heterozygosity and allele frequency. Model fit was evaluated by the difference from the best-fit overall model ($\Delta QIC = QIC_i - QIC_{min}$), percent reduction from the best-fit null model ($((QIC_{null} - QIC_i) / QIC_{null})$ and QIC weights (w) (Burnham & Anderson 2002). For a list of all GI and ectoparasite models evaluated, see Table S1 and S2. Finally, I validated candidate models by plotting Pearson's residuals against model fitted values to assess homogeneity, used residual histograms to assess normality, and plotted residuals against each explanatory variable to test for homogeneity

of error variances. Bartlett tests were also employed to confirm the latter (Snedecor & Cochran 1989). In best-fit models, the significance of parameter estimates was determined using a Wald Test, an appropriate test for GEE estimates (see Zuur *et al.* 2009).

All computations were carried out using R 2.12.2 (R Development Core Team, 2011) with GEE fitting performed using the `geepack` library (Halekoh *et al.* 2006), assessment of co-linearity using pairwise scatterplots using the `AED` package (Zuur 2010) and calculation of Pan's QIC (Pan 2001) performed using the `yags` library (Carey 2004).

Results

Data summary

Mean (± standard error) strongyle nematode egg count in the sample was 2541 (± 133) eggs per gram (EPG), ranging from 100 to 8050 EPG. I sampled an average of 4.5 (± 0.4) ticks per individual (range: 0 - 29 ticks/ individual). Of these, the majority of individuals were dominated by the tick species *R. e. mimeticus* of the Ixodidae family. Exploratory data analysis of both environmental (season) and demographic (sex and age) factors revealed GI parasitism was significantly higher in the wet than the dry season (Figure 4): Mean nematode FEC was 2903 (± 171) EPG and 1816 (± 160) EPG in the wet and dry seasons, respectively. On the contrary, sex and age did not appear to be related to GI parasitism. In exploratory data analysis, sex was only evaluated in the wet season given paucity of samples from males in the dry season, resulting in unbalanced data with which to assess patterns across seasons. With regards to ectoparasitism, there appeared to be no obvious effect of any of the evaluated ecological variables on tick abundance.

In this study, I found eight ELA-*DRA* alleles in the *E. quagga* population of ENP, all of which were previously discovered in *E. quagga*: *DRA*01,*03-*05,*07,*09-11* [GenBank: AJ575299, EU930126, EU930121, EU930118, HQ637392, HQ637392-HQ637396]. The *DRA* allele frequency distribution observed here was congruent with results based on a larger sample size from this same population (Figure 1; see also Chapter 3), revealing a strongly skewed distribution (Table 1). The alleles, *DRA*03* and *DRA*04* were the most common and were found at population proportions of approximately 0.19 and 0.28, respectively. *DRA*01, *05* and **09* were found at proportions from 0.08 to 0.10, whereas *DRA*07, *10* and **11* were the most rare, at proportions ≤ 0.05.

Initial assessment of the relationships between genetic variables and parasitism, based on raw data, suggested that mid-frequency alleles (*DRA*01, 05, *09*) conferred decreased parasitism, for both strongyle nematodes and ticks (Figure 5). Interestingly, the relationships between common and rare alleles with parasitism showed opposing patterns for the two general parasite types, with host immunity to GI parasite loads appearing to be relatively more important that immunity to ectoparasites. Individuals with a rare allele had significantly more GI parasites, whereas those with a common allele had significantly more ectoparasites, reflecting a potential MHC tradeoff between the two parasite types. However, parasite burden in heterozygotes versus homozygotes was not significantly different. At first glance, parasitism appears to have a weak, inverse relationship with the number of heterozygous positions (or SNPs) in an individual's

genotype (Figure 6). However, this interpretation is speculative since sample sizes were smaller for observations where the number of SNPs was higher (i.e. at 5 or 6 SNPs).

Model Selection

The final selected models for both GI and ectoparasite data fit my model validation criteria outlined in the methods (see above).

Gastrointestinal parasites (Nematodes):

Initial modeling of the GI parasite response variable (GI) revealed that the model scaling parameter (i.e. dispersion parameter) was consistently very high (>600), indicating greater variability in the data set than would be expected from the specified statistical model (i.e. overdispersion). Overdisperion may not only inflate standard errors but also indicates a poor fit to the selected model. Further, some explanatory variables violated the assumption of homogeneity of variance when fitted against model residuals. Therefore, I repeated the model selection process after applying a square root transformation on the GI parasite response variable, GI. This did not influence model selection or the covariates that were identified as having a significant effect; however, it limited the direct interpretation of the coefficient estimates derived from the GI parasite models. Thus, I based my interpretation on the sign and significance of model covariates to identify significant predictors of parasitism and describe the direction of the effect. Note that the GEE is a marginal model and thus, the relationships are not conditional on random effects, but only on explanatory variables (Hardin & Hilbe 2003; Zuur *et al.* 2009).

The best-fit non-genetic model of GI parasitism included season, age and gender as explanatory variables, but only a significant effect for season (SEASON (wet): 0.220 ± 0.051, Wald statistic = 18.66, p = 1.6e-06; Table S2) predicting increased parasitism in the wet season and corroborating results from exploratory analyses (Figure 4). The inclusion of genetic variables considerably improved the model fit, based on QIC values (ΔQIC = 45.68) and the maximal model with the lowest QIC included the additional genetic variables: SNPS, COMM, MID, RARE. The percent reduction in the QIC was also assessed and found to be small (ΔQIC / QIC < 0.001). Again, the GI parasite maximal model revealed a significant effect of season (0.220 ± 0.054, Wald statistic = 16.43, p = 5.1e-06), but also of rare allele presence in the *DRA* genotype (0.123 ± 0.051, Wald statistic = 5.92, p = 0.015) (Table S2). After further refinement, the final model for GI parasitism was fitted to a Poisson error structure and is defined as follows:

$$E(\sqrt{Y_{is}}) = \log(\mu_{is}) = 3.711 + 0.203 * SEASON + 0.123 * SEX - 0.10 * AGE + 0.04 * SNPS + 0.105 * COMM + 0.147 * RARE - 0.08 * SNPS * COMM,$$

where Y_{is} is the number of strongyle nematode eggs per gram in an individual i at time s, assuming a Poisson distribution with a mean μ_{is}. The variance and correlation between sampling time s and t are defined as:

$$var(\sqrt{Y_{is}}) = 3.78 * (\mu_{is})$$
$$cor(\sqrt{Y_{is}}, \sqrt{Y_{it}}) = 0.014^{|s-t|}$$

The modeling was based on data from 62 clusters (i.e. individuals) with a maximum cluster size of five and the α-correlation parameter, which estimates the correlation among observations within a cluster, was -0.014 ± 0.104. The estimated scale/dispersion parameter was 3.78 ± 0.443. The best-fit candidate model predicting GI parasitism again revealed season as a significant non-genetic explanatory variable (0.203 ± 0.057, Wald Statistic = 12.84, $p < 2e\text{-}16$; Table 4), consistent with both null and maximal model results. Also in agreement with the maximal model, the presence of a rare allele in the *DRA* genotype was found to significantly predict increased GI parasite burden (0.147 ± 0.042, Wald Statistic = 12.14, $p = 4.9e\text{-}04$), whereas there was no significant effect of common or mid-frequency alleles (Figure 7). This model included a significant interaction between the presence of a common allele and the number of SNPS at the *DRA* locus (-0.08 ± 0.034, Wald Statistic = 4.99, $p = 0.026$; Table 4). This suggests that heterozygous individuals with both a common allele and higher number of SNPs are more likely to have fewer GI parasites.

Ectoparasites (Ticks):

Initial data exploration revealed one high leverage outlier and therefore, I removed this observation from the data prior to model selection. The non-genetic null model for ectoparasitism only included season as a predictor, although the parameter estimate was not significant (SEASON (wet): -0.163 ± 0.110, Wald Statistic = 2.2, $p = 0.14$; Table S2). The maximal model of parasitism by ticks had a lower QIC than the null (ΔQIC /QIC= 0.034), but there were no significant effects for any of the included explanatory variables. Similarly, this result was also found in the best-fit model including all possible genetic and non-genetic covariates. The best-fit model was defined as:

$$E(Y_{is}) = \log(\mu_{is}) = 1.469 - 0.161 * SEASON + 0.006 * AGE - 0.061 * HET + 0.030 * COMM - 0.246 * MID + 0.337 * HET * COMM,$$

where Y_{is} is the number of ticks found on an individual i at time s, assuming a Poisson distribution with a mean μ_{is}. The variance and correlation between sampling time s and t were defined as:

$$var(Y_{is}) = 1.77 * (\mu_{is})$$
$$cor(Y_{is}, Y_{it}) = 0.045^{|s-t|}$$

The estimated scale/dispersion parameter was 1.77 ± 0.337. There were 64 clusters with a maximum cluster size of three and the α-correlation parameter estimate was 0.045 ± 0.079. The lack of significant coefficients in all models suggests either weak effects and/or missing explanatory variables that were not considered here.

DRA locus allele effects

The inclusion of allelic variables into the model significantly improved the previously described models for both GI (ΔQIC / QIC = -0.33) and ectoparasitism (ΔQIC / QIC = -0.32). Explanatory variables were substituted for the allele frequency covariates due to redundancy and collinearity. These allelic models corroborated results that the presence of a rare allele in an individual's genotype predicts an increased level of GI parasite intensity (Table 5). Specifically, the best-fit allelic model indicated significant effects of the alleles, *DRA*07* (0.180 ± 0.078, Wald Statistic = 5.35, $p = 0.021$) and *DRA*11* (0.305 ± 0.127, Wald Statistic = 5.81, $p = 0.016$) (Figure 8). The *DRA*10* allele was also significant in the maximal allele model, but this finding did not hold up in the model that best fit the data (Table 5). Once more, season was found to be a significant predictor of parasitism with a similar coefficient estimate as found in the previously described models.

In contrast, the ectoparasite model suggested that presence of the common alleles (*DRA*03* and *DRA*04*) as a significant predictor of tick burden (*DRA*03*: 0.681 ± 0.303, Wald Statistic = 5.06, $p = 0.025$; *DRA*04*: 0.769 ± 0.360, Wald Statistic = 4.57, $p = 0.033$; Table 5, Figure 9). The best-fit allelic model for ectoparasitism was equivalent to the maximal model. The finding that common alleles (in heterozygous genotypes) correlated with reduced GI parasitism, but also with elevated ectoparasite burden, is consistent with the hypothesis that GI parasitism and ectoparasitism are negatively correlated, and suggests that comparatively high GI parasitism has higher fitness costs to hosts than comparatively high ectoparasitism.

Relationships between parasite types

The opposing results from statistical models of GI parasitism and ectoparasitism warranted further post-hoc investigation into whether a direct inverse relationship (i.e. a co-infection effect) exists between these broadly grouped parasite types. I used the joint parasite data from a subset of the total data set ($n = 86$, after the removal of a high leverage outlier) to conduct an independent statistical analysis using both GI parasite and tick data, adding each reciprocally as explanatory and response variables. Again, I square-root transformed the variable, GI, due to overdispersion. In both cases, I found that the alternate parasite effect variable was not significant (GI: 1.53e-05, Wald Statistic = 0.16, $p = 0.69$; ECTO: 0.007 ± 0.012, Wald Statistic = 0.37, $p = 0.54$). Furthermore, when I added each parasite variable reciprocally as a covariate to maximal and best-fit models for each parasite response, I similarly found that co-infection effects were not significant. These results suggested that a direct relationship between the type major parasite groups does not exist and strengthens my argument that parasitism may be modulated by the host immune system.

Discussion

Statistical models of parasitism revealed significant immunogenetic effects predicting parasite intensity in *E. quagga*, improving upon purely ecological models of parasitism. These data suggest that the MHC locus, ELA-*DRA*, is centrally involved in a complex interplay between host and parasite and strongly support the occurrence of parasite-mediated frequency-dependent selection acting on this locus. Alleles present at opposing frequencies in the population were found to have non-additive effects, increasing susceptibility to different parasite types, such that rare alleles were associated with increased GI parasitism and common alleles with ectoparasitism. This intriguing finding highlights the potential for host immunogenetic tradeoffs, with multiple parasite groups competitively driving patterns of diversity at the MHC. Although I found little conclusive evidence for heterozygote advantage, the interaction between allelic effects and heterozygosity, in terms of the degree of divergence between alleles, implies that both heterozygote advantage and frequency-dependence mechanisms are likely acting in concert to shape selection at this locus.

Non-genetic effects on parasitism

The finding that season strongly predicts strongyle nematode intensity in plains zebra corroborated results from a previous study in this population (Turner & Getz 2010), with increased parasitism in the wet versus the dry season. In further agreement, an individual's sex did not significantly affect parasite intensity, although best-fit genetic and allelic models of GI parasitism did include this factor as an explanatory variable, suggesting that it does explain some of the variation in parasite intensity. As these data were biased with an increased sample size from females, and subsequent lack of data from males in the dry season, it is difficult to draw any solid conclusions from this result as this unbalance is reflected in large confidence intervals of the estimated coefficients. However, it is unlikely that sex plays a role, given results from Turner & Getz, 2010, which is based on a very large and balanced sample size from this population. This previous research also found a significant age effect when contrasting yearlings and adults to juveniles (< 1 year old). However, these results only consider adult individuals between the estimated ages of 2.5 to 12 years old. So, although not comparable, these results suggest that among adults, parasite intensity does not increase or decrease in any consistent manner through an individual's lifetime. The lack of a relationship between age and parasite intensity suggests that an individual's adaptive response does not confer increased resistance over time due to continual exposure. Alternatively, given the previous results of Turner & Getz, 2010, showing juveniles had higher parasite intensities than adults, it is possible that acquired immunity plays a predominant role at an early age, from birth through initial exposure, but thereafter does not change with time.

Models of ectoparasitism, while including ecological and demographic effects in best-fit models, suggested that these variables do not significantly explain patterns of parasite intensity. Tick burdens appeared to be higher in the dry season, contrasting the results found for GI parasites, but the differences were not significant. However, incorporating knowledge on ectoparasite life stages is warranted given that larval, nymphal and adult activity has been known to peak in different seasons (Biggs & Langenhoven 1984; Walker *et al.* 2000), and including such information in these models

may reveal different relationships between environmental variables and abundance of tick life stages.

Hypothesis of frequency dependence

All models of parasitism were significantly improved with the addition of allelic explanatory variables, with greater than 32% reduction in QIC from the best-fit null (i.e. non-genetic) models. Most notably, results strongly supported the hypothesis of frequency-dependent selection acting on the *DRA* locus. This conclusion was primarily based on the results showing significant effects of specific alleles on parasite patterns, but even more so by the significance of alleles from particular frequency classes. Most intriguingly, I found that the frequency-dependent predictors of GI parasites were in opposition with predictors of ectoparasite intensity, potentially reflecting an immune system tradeoff (see below). In particular, individuals with rare alleles had significantly higher nematode FECs, whereas individuals with common alleles had significantly higher tick burden. In addition, to support this hypothesis, zebra with mid-frequency alleles had lower parasite burdens in general, regardless of the parasite type.

In previous work, I found that the *DRA* locus exhibited an unusually skewed allele frequency distribution in *E. quagga* of ENP, differing from that found in zebra of KNP (Figure 1; see also Chapter 3) which exhibited a more even allele frequency distribution typically attributed to balancing selection (Takahata & Nei 1990). This observed difference in allelic patterns over zebra populations is consistent with spatially fluctuating selection occurring at the *DRA* locus and suggests that there are critical differences in drivers of selection present in ENP versus KNP host populations. The more arid climate of ENP, when contrasted to KNP, has been hypothesized to be responsible for the relatively lower Strongylinae nematode species richness observed in plains zebra of ENP (Matthee *et al.* 2004). Congruent with this hypothesis, tick species richness is lower in ENP zebra─ a total of 5 Ixodidae *spp.* have been found both in this study and previously (Horak *et al.* 1992), whereas 7 Ixodidae *spp.* have been identified in plains zebra of KNP (Horak *et al.* 1984b). This difference in parasite species richness across populations has been suggested as a potential explanation for the directional allele frequency skew observed at the *DRA* locus (see Chapter 3). Here I also found evidence to suggest rare *DRA* alleles in ENP zebra are less fit with respect to nematode resistance and I speculate that GI parasites are likely driving the directional skew in the *DRA* allele frequency distribution observed at this locus. Given these results, I argue that both parasite richness and intensity may be important factors behind the observed skew in ENP zebra due to the combination of low species richness, but high prevalence and abundance of GI nematodes. Tick abundance also appears to be lower in ENP than in KNP when comparing previous studies in both systems (Horak *et al.* 1984b; Horak *et al.* 1992), and further suggests the possibility that ticks may play a weaker selective role on host immunogenetics in KNP. However, additional research comparing MHC-parasite relationships between these populations is necessary to confirm these postulations.

Incongruent association patterns of *DRA* alleles with tick versus GI nematode burden, implies that the *DRA* has pleiotropic effects on different parasite types and adds a layer of complexity to the story of selection at this locus. Although GI parasites may be significant drivers of selection at this MHC locus (see below), it is also possible I have only observed a snapshot of the selective process in time and that this MHC-parasite

system exists in a state of flux. The reduction in both GI and ectoparasitism by mid-frequency alleles supports this idea as these alleles would be expected to be most stable, as opposed to alleles at either end of the spectrum that have putative negative fitness effects with respect to one or the other parasite types.

Hypothesis of heterozygote advantage

Comparison of candidate models excluded *DRA* heterozygosity (in terms of having two different alleles at this locus) as a significant predictor of either type of parasitism. This was further confirmed by the lack of significant differences in parasite load between homozygotes and heterozygotes in the ENP zebra population. Heterozygosity, defined more explicitly at the molecular level as the number of heterozygous base positions, was included in best-fit models, but also was not a significant predictor in the majority of these models. However, the best-fit GI parasite model did include a significant interaction between alleles found at high frequency (COMM) and heterozygosity (SNPS), together incurring a negative effect on parasitism. This result suggested that GI parasitism decreases in common allele heterozygotes as the mutational differences between the two alleles in the *DRA* genotype increases (i.e. with a higher number of SNPs). This finding is congruent with the divergent allele advantage hypothesis (Wakeland *et al.* 1990) which asserts that more divergent alleles will increase functionality in the peptide-binding repertoire and thus have higher adaptive value.

Taken together, these results argue for both the models of heterozygote advantage and frequency-dependence, both of which are thought to be able to occur due to pathogen evasion of the MHC recognition process (reviewed in Apanius *et al.* 1997). In this case, pathogen avoidance of host recognition by specific allelic variants may be expected to drive a fluctuating frequency-dependent selective process, while at the same time, hosts with greater diversity in antigen binding sites will be able to recognize more diverse parasite types. Oppelt *et al.* 2010, also observed similar complex interactions in a study that found one MHC allele of the *DRB* locus associated with resistance to intestinal coccidian parasites in European rabbits (*Oryctolagus cuniculus*), but only when found in a heterozygous genotype (i.e. homozygotes for this allele did not have higher fitness). However, they asserted that allele divergence could not explain their observations, given that one disadvantageous genotype had more amino acid residue differences than an advantageous genotype. Finally, these results suggest both frequency-dependent and heterozygote advantage mechanisms are acting in concert at the *DRA* locus, and further support the divergent allele hypothesis.

'Susceptibility alleles' and antagonistic parasite effects

In this study, I found specific *DRA* alleles associated with increased parasitism, thereby conferring susceptibility rather than resistance to either parasite type. There have been several studies that have similarly reported genetic variants associated with susceptibility, in humans (Segal & Hill 2003) as well as wildlife (Schad *et al.* 2005; Schwensow *et al.* 2007; Loiseau *et al.* 2008). But, what mechanisms would allow for maladaptive genetic alleles to be maintained in a population? Pleiotropy, the phenomenon that a single gene can affect multiple traits, is believed to be widespread in nature, and pleiotropic effects have been shown theoretically to reduce the ability of beneficial alleles to be achieve fixation (Otto 2004). Furthermore, in a natural study on house sparrows,

Loiseau *et al*., 2008, found convincing evidence for antagonistic effects of a MHC class I gene on multiple malarial parasite strains, suggesting that these pleiotropic effects allowed for the persistence of apparently deleterious 'susceptibility alleles' in the population. Here, I found that allelic associations differed with respect to parasite type, with common alleles conferring susceptibility to ectoparasites as opposed to rare alleles increasing susceptibility to strongyle nematode infections. This suggests that an MHC antagonistic cost-benefit trade-off exists in conferring resistance/susceptibility to these two parasite types. Whereas, alleles inferred to be beneficial for reducing GI parasites have been driven up to high frequencies, alleles associated with susceptibility have been apparently selected against and hence, are rare in the population. With this, there is a cost to strongyle nematode resistance incurred to the host, in the form of an increased tick burden. The implications of resistance costs in determining an equilibrium level of resistance have been widely discussed and exemplified in predator/pathogen-prey model systems of *Escherichia coli* (Lenski 1988) and plants (Simms 1992). Hence, I hypothesize that the observed skew in the frequency distribution at the *DRA* locus implies that GI parasites play a more significant role in shaping the patterns of variation at this locus, and pleiotropic antagonistic effects modulate resistance/susceptibility to multiple parasites in this system.

Alternatively, the lack of specific alleles that confer resistance leads us to also consider the possibility that linkage effects may be at play, such that a gene within close proximity to the *DRA* locus may be related to the gene diversity and allelic distribution patterns I have observed in this study population. Equine Lymphocyte Antigen genes have been found to be tightly clustered, localized to chromosome 20q14-q22 in the horse genome (Ansari *et al.* 1988) and, therefore, I cannot rule out the possibility that parasites are mediating selection on another proximate locus of this gene family. Under this alternate hypothesis that a linked MHC gene is related to increased or decreased parasitism, one may also expect to similarly find associations between allele frequency classes and parasitism in nearby MHC loci.

Epistatic gene interactions on fitness, caused by either linkage disequilibrium of neighboring MHC genes or with genes outside of the complex, could further complicate conclusions drawn from these results. These effects have been shown to be a common occurrence among human MHC genes in explaining patterns of disease susceptibility, for example in multiple sclerosis, psoriasis and diabetes (reviewed in Traherne 2008). Furthermore, epistatic gene interactions have been shown to alter the cost of host resistance in a multi-parasite model system (Bohannan *et al.* 1999). Regardless of possible gene linkage, my results indicate that a general immune system tradeoff is influencing the MHC variation observed in the data and, moreover, demonstrate that the *DRA* locus is a useful genetic marker reflecting parasite-mediated selective patterns in the MHC gene region. However, I assert that future research examining associations between neighboring MHC genes in this study system is necessary to identify the immunogenetic variation that is specifically responsible for host immune response to macro-parasites.

Finally, lack of significant co-infection effects suggested that a direct relationship between parasite types may not exist. This can be explained by the fact that these parasites inhabit different areas of the host's body and, thus, are not physically interacting. These results also strengthen my argument of a possible indirect antagonistic

relationship occurring at the molecular level through modulation by the host immune system as a plausible explanation for the observed selection patterns.

Conclusions

In this study, I found evidence for parasite-mediated selection occurring on an MHC locus in a free-ranging zebra population. These results strongly supported directional selection by GI parasites on zebra hosts, modulated primarily by a frequency-dependent selective mechanism, a conclusion which explains the severely skewed *DRA* allele frequency distribution previously observed in this study system. These data also suggested that heterozygote advantage, under a divergent allele hypothesis plays a role in conferring resistance to GI parasites in this population, but may occur within a frequency-dependent framework. Most significantly, I observed what appeared to be pleiotropic antagonistic effects of a MHC gene affecting susceptibility/resistance to multiple parasites in this system. I hypothesize that the high frequency of ectoparasite 'susceptibility alleles' in this zebra population reflects an immunogenetic trade-off, and therefore a cost incurred with having increased resistance to GI parasites. These findings underscore the importance of considering multiple parasites when investigating the selective mechanisms driving host immune gene variation.

Acknowledgements

I would like to thank the staff at the Etosha Ecological Institute for their invaluable support with the sample collection, particularly W. Kilian, W. Versfeld, G. Shatumba, M. Kasona, J. Kapner, S. Kötting and B. Kötting, for assistance during capture events. I would also like to specifically acknowledge the veterinarians (Nad Brain, M. Jago and O. Aschenborn) who were critical for conducting the zebra captures and I. Horak for identifying the tick species used in this study. In addition, I am grateful to members and associates of the Getz Lab who similarly played an important role during captures: C. Cizauskas, H. Ganz, S. Bellan, M. Tsyaluk, L. Polansky, R. Zidon, O. Spiegel, Z. Havarua, M. Shikango. I thank the Ministry of Environment and Tourism, Namibia, for supporting and granting us permission to conduct this study. Finally, I thank R. Bowie and members of his lab group for helpful comments on the manuscript. This research was supported by the National Institute of Health Ecology and Evolution of Infectious Disease (NIH-EEID) GM083863 to WMG and National Science Foundation Doctoral Dissertation Improvement Grant (NSF-DDIG) MCINS-20091291 to PLK.

Tables

Table 1. ELA-*DRA* allele and genotype frequency data in Etosha National Park

			Homozygote			Heterozygote		
Allele	*n*	*f*	*Genotype*	*n*	*f*	*Genotype*	*n*	*f*
****03***	55	0.423	****03*03***	18	0.277	****03*04***	7	0.108
						****03*05***	3	0.046
						**03*01*	0	0
						****03*09***	3	0.046
						****03*07***	2	0.031
						****03*10***	2	0.031
						****03*11***	2	0.031
****04***	25	0.192	****04*04***	6	0.092	****04*05***	1	0.015
						**04*01*	0	0
						****04*09***	1	0.015
						****04*07***	3	0.046
						**04*10*	0	0
						****04*11***	1	0.015
****05***	13	0.100	****05*05***	3	0.046	****05*01***	2	0.031
						****05*09***	1	0.015
						**05*07*	0	0
						**05*10*	0	0
						**05*11*	0	0
****01***	11	0.085	****01*01***	4	0.062	****01*09***	1	0.015
						**01*07*	0	0
						**01*10*	0	0
						**01*11*	0	0
****09***	10	0.077	****09*09***	2	0.031	**09*07*	0	0
						**09*10*	0	0
						**09*11*	0	0
****07***	7	0.054	****07*07***	1	0.015	**07*10*	0	0
						**07*11*	0	0
****10***	5	0.038	****10*10***	1	0.015	****10*11***	1	0.015
****11***	4	0.031	**11*11*	0	0.000			

Reported allele and genotype frequencies (*f*) observed in this study out of a samples size (*n*) of 130 alleles. All potential genotypes are listed, but only those observed in the dataset are highlighted in bold.

Table 2. List of response and explanatory variables used in statistical analyses with short descriptions

Variable	Abbreviation	Description
Response: Parasitism		
Gastrointestinal parasite burden	GI	GI parasite intensity (nematode eggs/gram of feces)
Ectoparasite burden	ECTO	Ectoparasite intensity (total number of ticks)
Explanatory: Non-genetic		
Season	SEASON	Wet (Nov - May) or Dry (Apr - Oct)
Sex of individual	SEX	Female or male
Age of individual	AGE	Years of age estimated from dental wear (ref).
Explanatory: Genetic diversity		
Heterozygosity	HET	Heterozygote or homozygote
Single nucleotide polymorphisms	SNPS	The number of heterozygous base positions
Rare alleles	RARE	Presence or absence of a rare allele (approx. < 5% population freq.)
Mid-frequency alleles	MID	Presence or absence of a mid-frequency allele (5 - 10% population freq.)
Common alleles	COMM	Presence or absence of a common allele (>15% population freq.)
Genotype frequency	GENO.FREQ	Observed genotype frequency in dataset
Explanatory: Alleles		
*DRA*01* allele	D1	Presence or absence of DRA*01 allele
*DRA*03* allele	D3	Presence or absence of DRA*03 allele
*DRA*04* allele	D4	Presence or absence of DRA*04 allele
*DRA*05* allele	D5	Presence or absence of DRA*05 allele
*DRA*07* allele	D7	Presence or absence of DRA*07 allele
*DRA*09* allele	D9	Presence or absence of DRA*09 allele
*DRA*10* allele	D10	Presence or absence of DRA*10 allele
*DRA*11* allele	D11	Presence or absence of DRA*11 allele

Table 3. List of maximal null and genetic statistical models evaluated

Null Models
Parasitism[1] ~ SEASON + SEX + AGE
Maximal Models
Parasitism[1] ~ SEASON + SEX + AGE + HET + GENO.FREQ
Parasitism[1] ~ SEASON + SEX + AGE + HET + COMM + MID + RARE
Parasitism[1] ~ SEASON + SEX + AGE + SNPS + GENO.FREQ
Parasitism[1] ~ SEASON + SEX + AGE + SNPS + COMM + MID + RARE

[1] Parasite response variable used was either GI or ECTO as described in Table 2

Table 4. Model coefficient estimates (± standard errors) in best-fit genetic models of parasitism in zebra.

Response	Coefficients	Estimate ± S.E.	Wald Statistic	*P*-value	% reduction from QIC_{null}
GI[1]	**Intercept**	**3.711 ± 0.122**	**922.93**	**< 2e-16*****	0.08%
	SEASON (wet)	**0.203 ± 0.057**	**12.84**	**3.4e-04*****	
	SEX(male)	0.123 ± 0.105	1.48	0.224	
	AGE	-0.010 ± 0.009	1.24	0.265	
	SNPS	0.040 ± 0.032	1.56	0.212	
	COMM (presence)	0.105 ± 0.092	1.29	0.257	
	RARE (presence)	**0.147 ± 0.042**	**12.14**	**4.9e-04*****	
	SNPS:COMM	**-0.080 ± 0.034**	**4.99**	**0.026***	
ECTO	**Intercept**	**1.469 ± 0.283**	**27.02**	**2.0e-07*****	3.42%
	SEASON (wet)	-0.161 ± 0.112	2.07	0.150	
	AGE	0.006 ± 0.026	0.06	0.810	
	HET (yes)	-0.061 ± 0.244	0.06	0.800	
	COMM (presence)	0.030 ± 0.223	0.02	0.890	
	MID (presence)	-0.246 ± 0.154	2.56	0.110	
	HET:COMM	0.337 ± 0.290	1.35	0.240	

Percent reductions in quasi-information criteria (QIC) values from best-fit null models (QICnull) are shown. Significance of model coefficients determined by the Wald tests and significant terms highlighted in bold and significance level (p) indicated by asterisks: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*).

Comparison of the candidate models ruled out additional factors as predictors of parasitism; therefore, only explanatory variables in best-fit models are shown.

Table 5. Relationships between specific *DRA* alleles and parasitism as predicted by both the maximal and best-fit allele models

Response		Coefficients	Maximal allele model			Best fit allele model			% reduction from QIC_{null}
			Estimate ± S.E.	Wald Statistic	*p*-value	Estimate ± S.E.	Wald Statistic	*p*-value	
GI[1]		**Intercept**	3.565 ± 0.153	545.15	**<2e-16 *****	3.71 ± 0.127	874.41	**<2e-16*****	33.48%
		SEASON (wet)	0.201 ± 0.053	14.36	**1.5e-04 *****	0.206 ± 0.052	15.66	**7.6e-05*****	
		SEX (male)	0.151 ± 0.099	2.31	0.128	0.150 ± 0.105	2.04	0.154	
		AGE	-0.010 ± 0.008	1.41	0.235	-0.011 ± 0.008	1.78	0.182	
		SNPS	-0.088 ± 0.044	4.07	**0.044***	-0.049 ± 0.030	2.70	0.100	
	common	*03	0.265 ± 0.136	3.81	0.051	0.137 ± 0.089	2.36	0.124	
		*04	0.153 ± 0.143	1.16	0.282	0.015 ± 0.102	0.02	0.882	
	mid-freq	*01	0.129 ± 0.152	0.48	0.489				
		*05	0.129 ± 0.152	0.72	0.396				
		*09	0.206 ± 0.173	1.42	0.233	0.063 ± 0.126	0.25	0.616	
	rare	***07**	0.246 ± 0.088	7.74	**0.005****	0.180 ± 0.078	5.35	**0.021***	
		***10**	0.273 ± 0.126	4.70	**0.030***	0.150 ± 0.092	2.66	0.103	
		***11**	0.486 ± 0.190	6.54	**0.011***	0.305 ± 0.127	5.81	**0.016***	
Ecto		**Intercept**	0.712 ± 0.384	3.43	0.064				31.63%
		SEASON (wet)	- 0.183 ± 0.120	2.32	0.128				
		SEX (male)	0.047 ± 0.191	0.06	0.807				
		AGE	0.014 ± 0.024	0.32	0.572				
		SNPS	-0.137 ± 0.094	2.15	0.142				
	common	***03**	0.765 ± 0.301	6.48	**0.011***	*same as maximal model*			
		***04**	0.870 ± 0.355	6.02	**0.014***				
	mid-freq	*01	0.524 ± 0.409	1.64	0.200				
		*05	0.369 ± 0.231	2.55	0.110				
		*09	0.402 ± 0.381	1.11	0.290				
	rare	*07	0.219 ± 0.266	0.68	0.410				
		*10	0.346 ± 0.383	0.82	0.366				
		*11	0.655 ± 0.469	1.95	0.162				

GI = the intensity of gastrointestinal parasitism in eggs per gram; ECTO = intensity of ectoparasitism in terms of individual tick count. Coefficients estimates and their standard errors (Estimate ± S.E.) are shown, significance of effects was tested using the Wald Statistic and associated *p*-values ($p < 0.05$ highlighted in bold). The percent reduction in the quasi-information criterion (QIC) value from the best-fit null model (QIC_{null}) is shown.

Figures

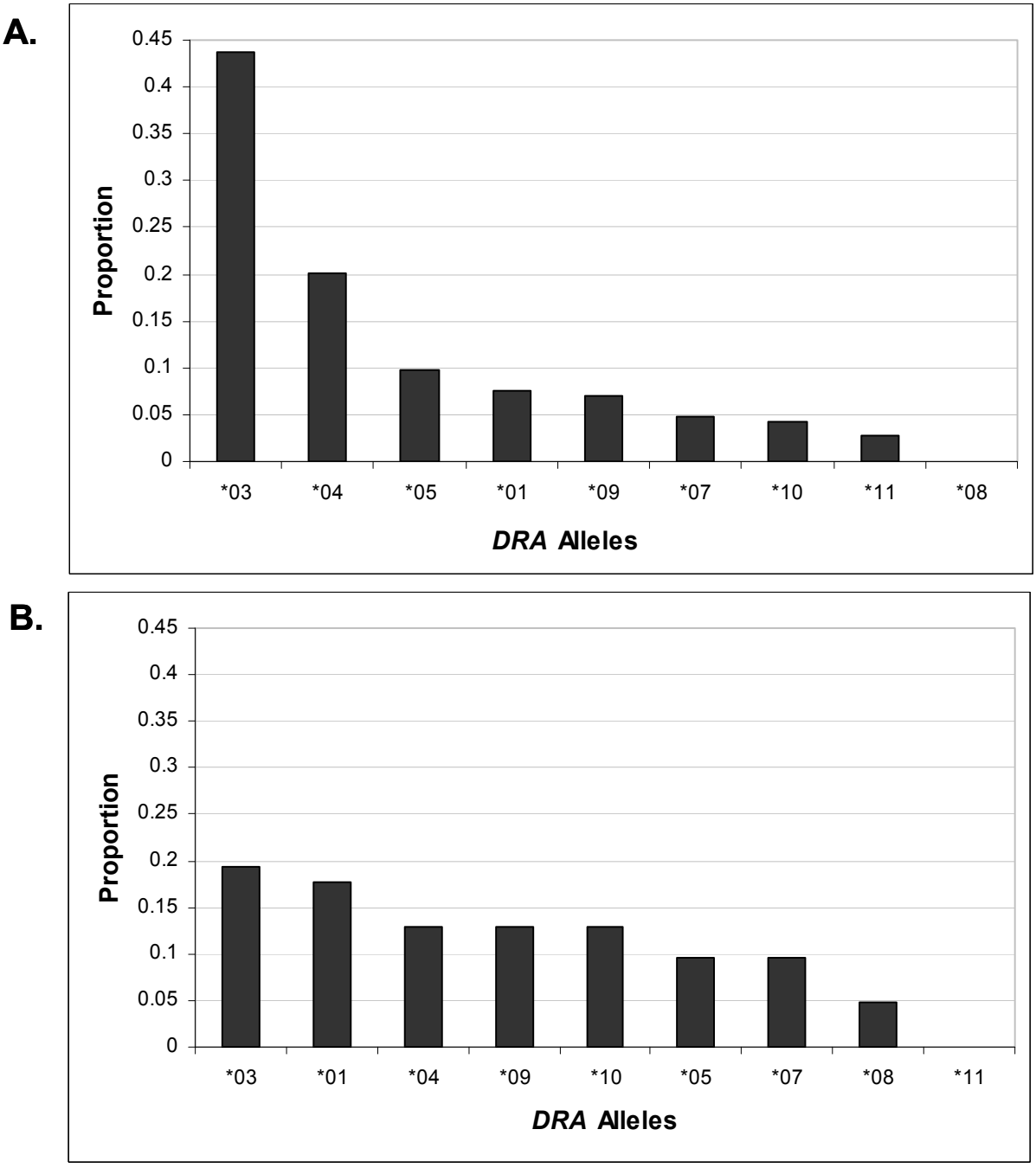

Figure 1. *ELA-DRA* exon 2 allele frequency distributions in *Equus quagga* populations
(A) Etosha National Park, Namibia, and (B) Kruger National Park, South Africa. Adapted from Chapter 3.

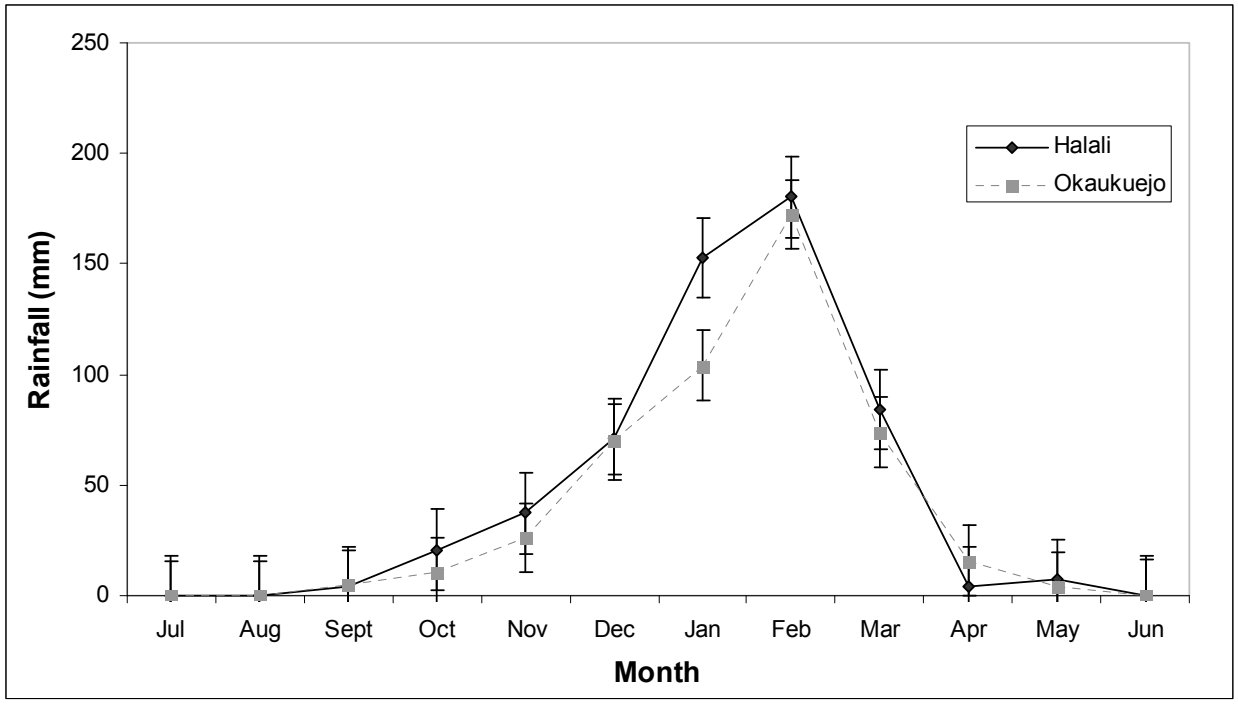

Figure 2. Mean monthly rainfall over 2007-2010 from the Halali and Okaukuejo weather stations in central Etosha

Rainfall data at each weather station are reported as the mean and standard error by month, and were collected from July 2007 to December 2010 covering the duration of the study period.

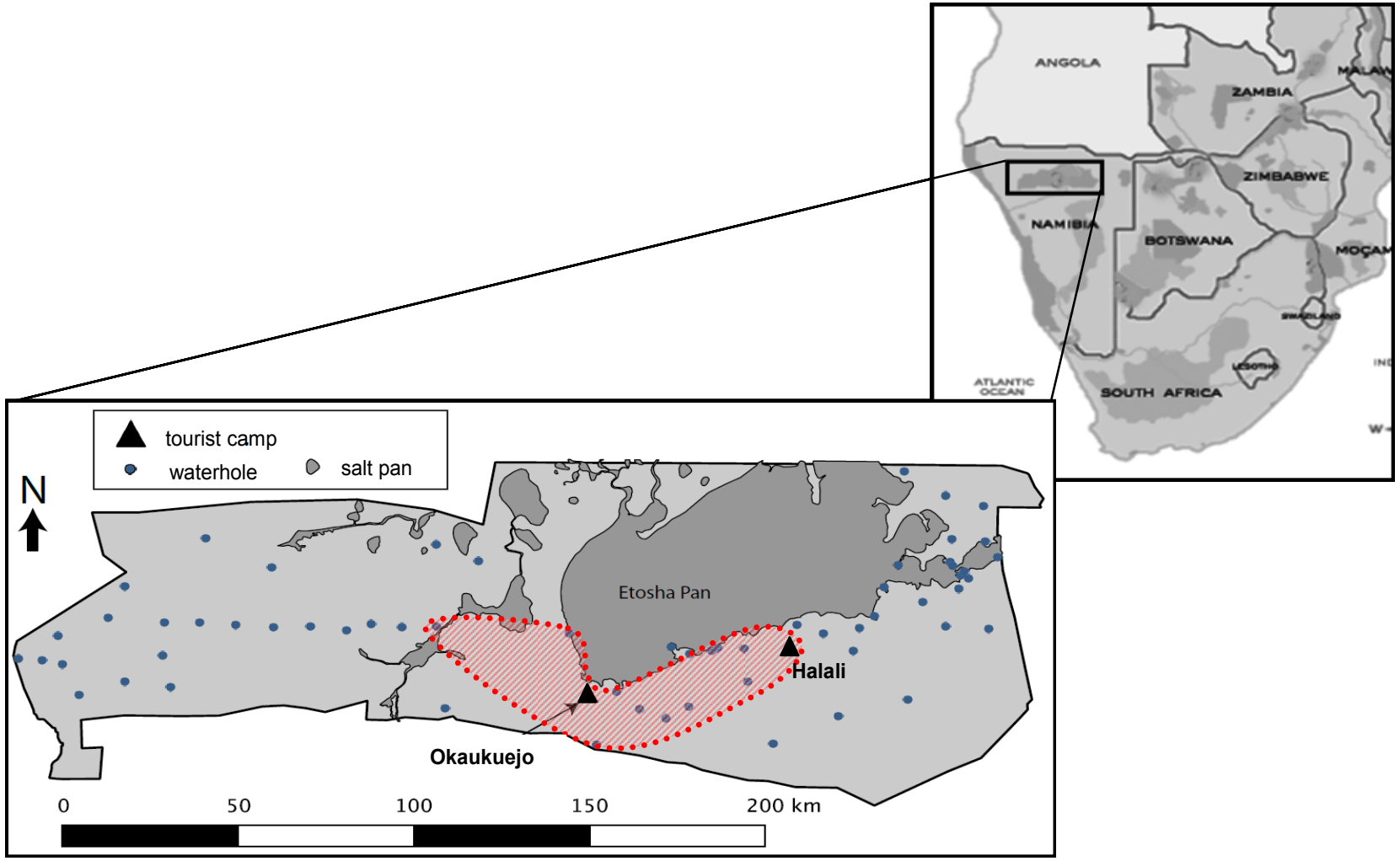

Figure 3. Map of Etosha National Park, Namibia showing sampling area of study
Sampling area from which zebra were captured (in red) chiefly covered the Okaukuejo and Halali central plains region of the park. Okaukuejo and Halali tourist camps, where the weather stations are located, are indicated by the black triangles.

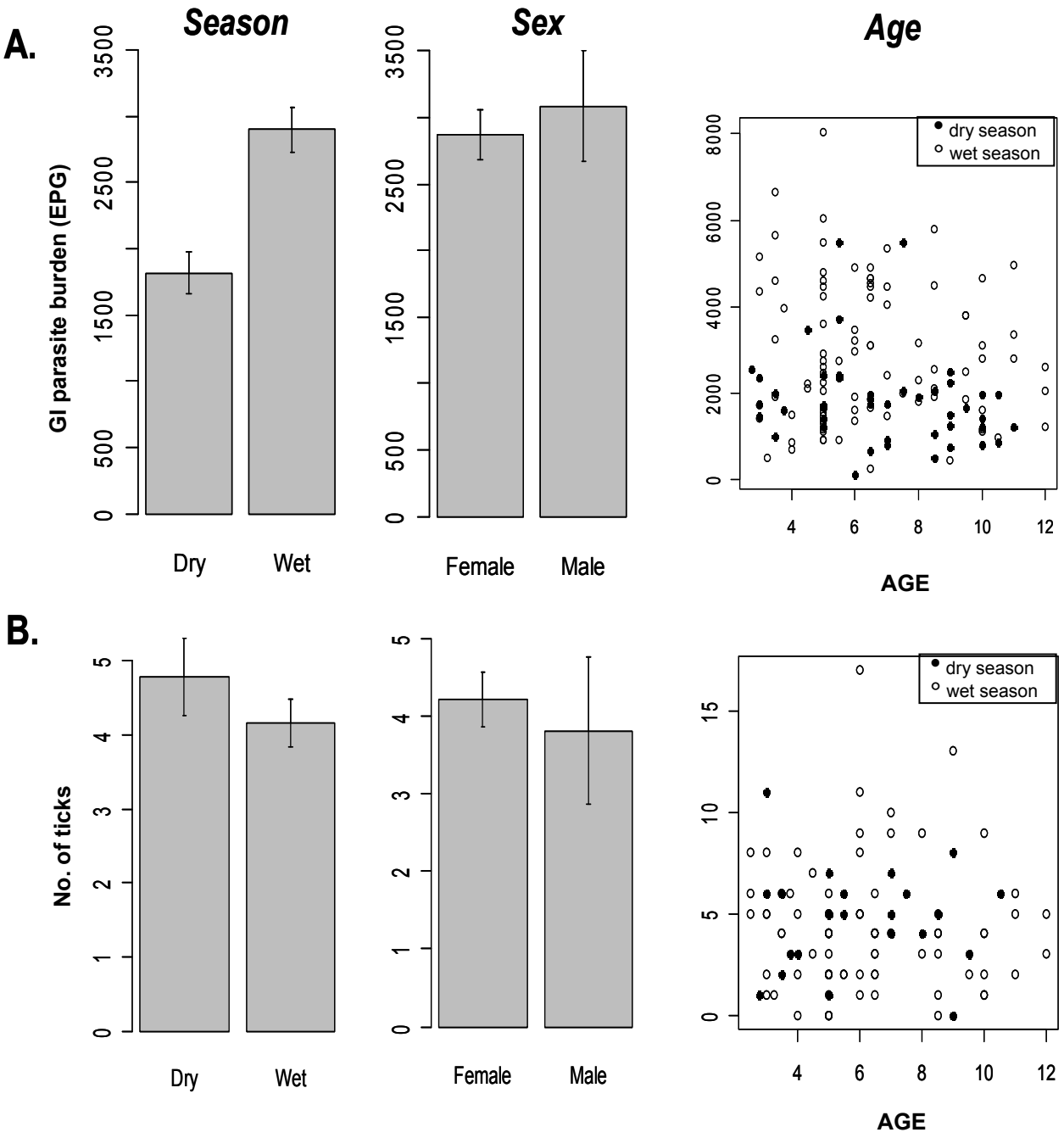

Figure 4. Relationships between non-genetic explanatory variables (season, sex, age) and parasitism

(A) Gastrointestinal (GI) parasite burden in eggs per gram (EPG) and (B) ectoparasite burden in terms of number of ticks. Samples sizes are as follows: (A) *Season*: wet = 90, dry = 54; *Sex*: female= 77, male=13; *Age*: *n* = 143. (B) *Season*: wet = 86, dry = 23; *Sex*: female = 75, male = 11; *Age*: *n* = 118. Barplot of the sex-parasite relationship displays only data from the wet season to control for the large effects of season (with data from the dry season removed). Mean and 95% confidence intervals are shown on barplots of season and sex.

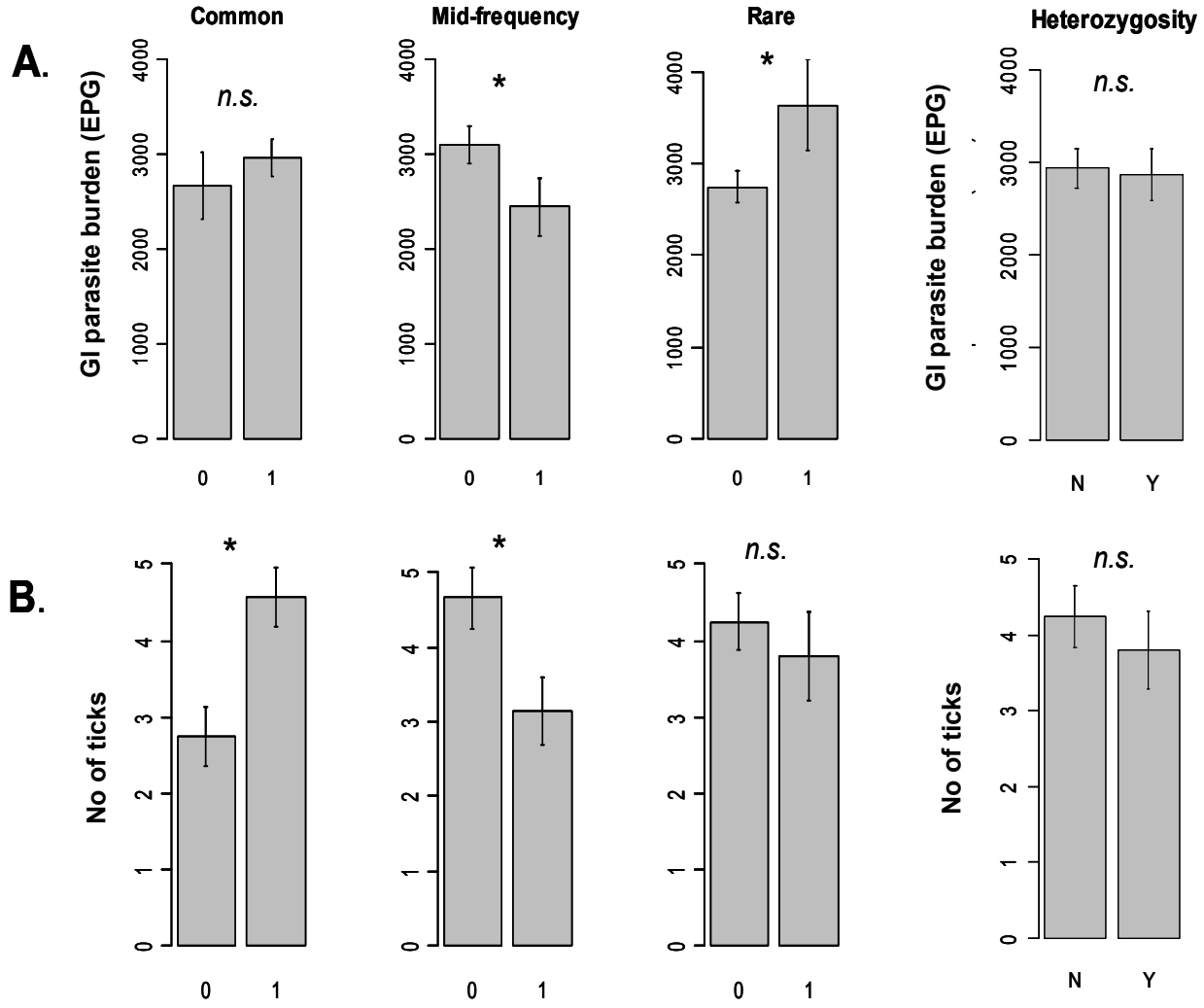

Figure 5. Relationships between genetic explanatory variables (common/mid/rare frequency alleles and heterozygosity) and parasitism

(A) Gastrointestinal (GI) parasite burden in eggs per gram (EPG) and (B) ectoparasite burden as the number of ticks. All measurements shown are from the wet season only. Absence (0) or presence (1) of a particular allele class is indicated, whereas heterozygosity is designated as either no (N) or yes (Y). Samples sizes are as follows: (A) *Common:* absence= 18, presence= 72; *Mid:* absence= 62, presence= 28; *Rare:* absence= 74, presence=16; *Heterozygosity:* no= 48, yes= 42 (B) *Common:* absence= 19, presence= 67; *Mid:* absence= 58, presence = 28; *Rare:* absence= 71, presence=15; *Heterozygosity:* no= 44, yes= 42. Mean and 95% confidence intervals are shown, and significance ($p < 0.05$) indicated by an asterisk (∗) or non-significance (n.s.).

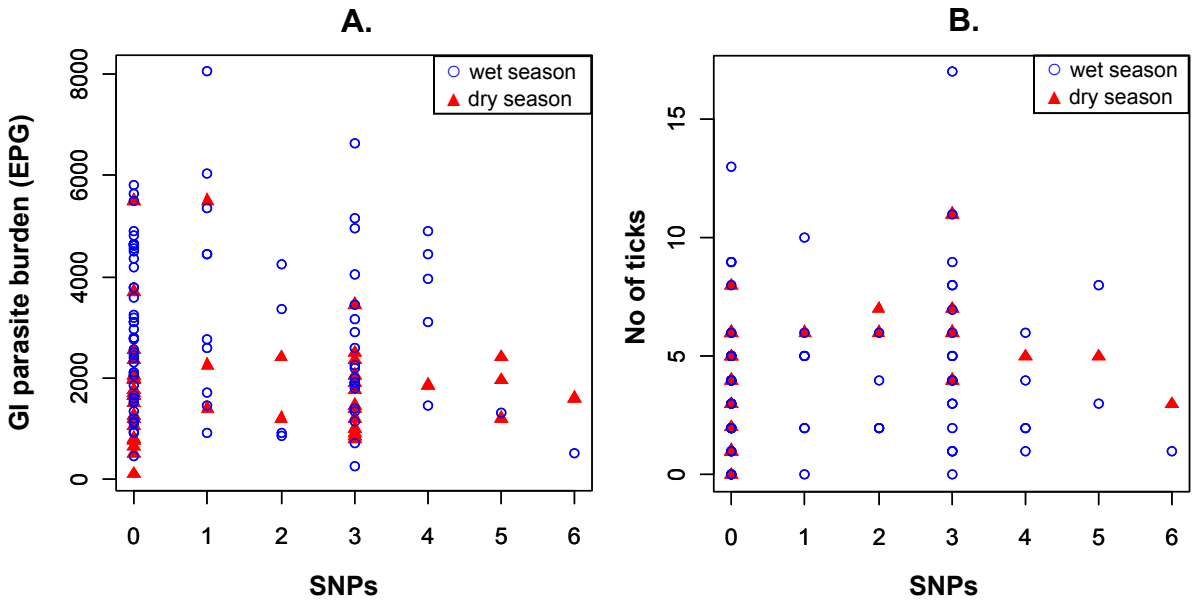

Figure 6. Relationships between heterozygosity in terms of the number of single nucleotide polymorphisms (SNPs) and parasitism

(A) Gastrointestinal (GI) parasite burden in eggs per gram (EPG) and (B) ectoparasite burden as the number of ticks.

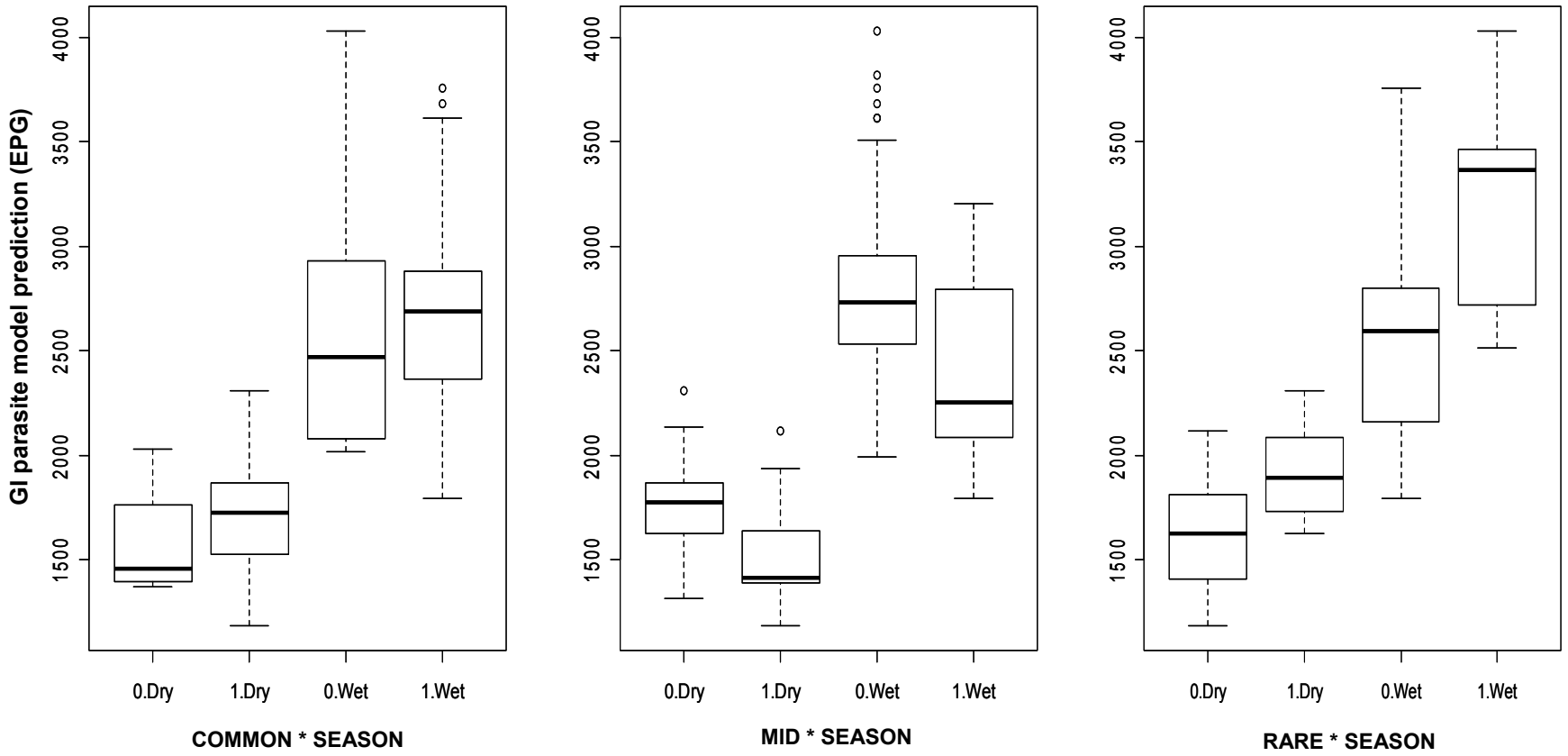

Figure 7. Gastrointestinal parasitism model predictions for frequency-dependent genetic effects of ELA-*DRA*
Showing effects of common (>20% alleles), mid-frequency (approx.10-15%) and rare (<5%) alleles on parasite burden. Factors are defined as the presence (1) or absence (0) of an allele of a particular frequency category in an individual's *DRA* genotype by season (dry, wet). Only rare alleles were found to significantly ($p < 0.001$) predict increased GI parasitism and the common and mid-frequency alleles had overall non-significant effects. Ninety-five percent confidence intervals are indicated by the dashed lines.

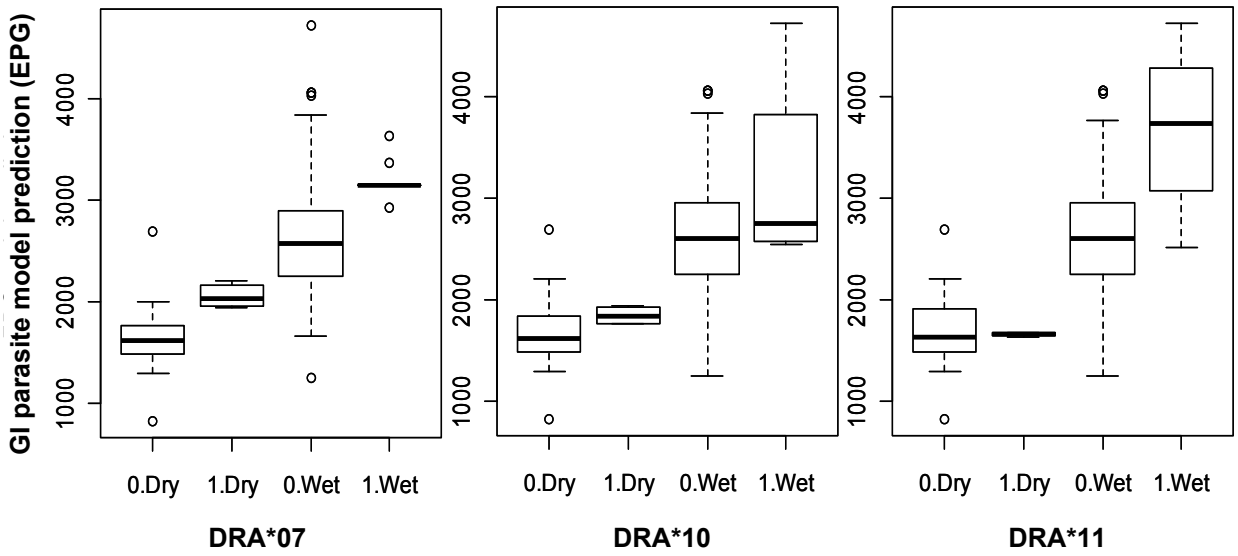

Figure 8. Allele-specific effects predicting GI parasitism, showing susceptibility effects of rare alleles, ELA-*DRA *07*, **10* and **11*

DRA allele- parasite relationships depicted in terms of the presence (1) or absence (0) of the allele in the *DRA* genotype, by season (dry versus wet). Here maximal model predictions are shown where all three alleles had a significant ($p < 0.05$) susceptibility effect. All effects predicted increased nematode load, however only *DRA*07* and **11* were significant in the best-fit allele model. Ninety-five percent confidence intervals are indicated by the dashed lines.

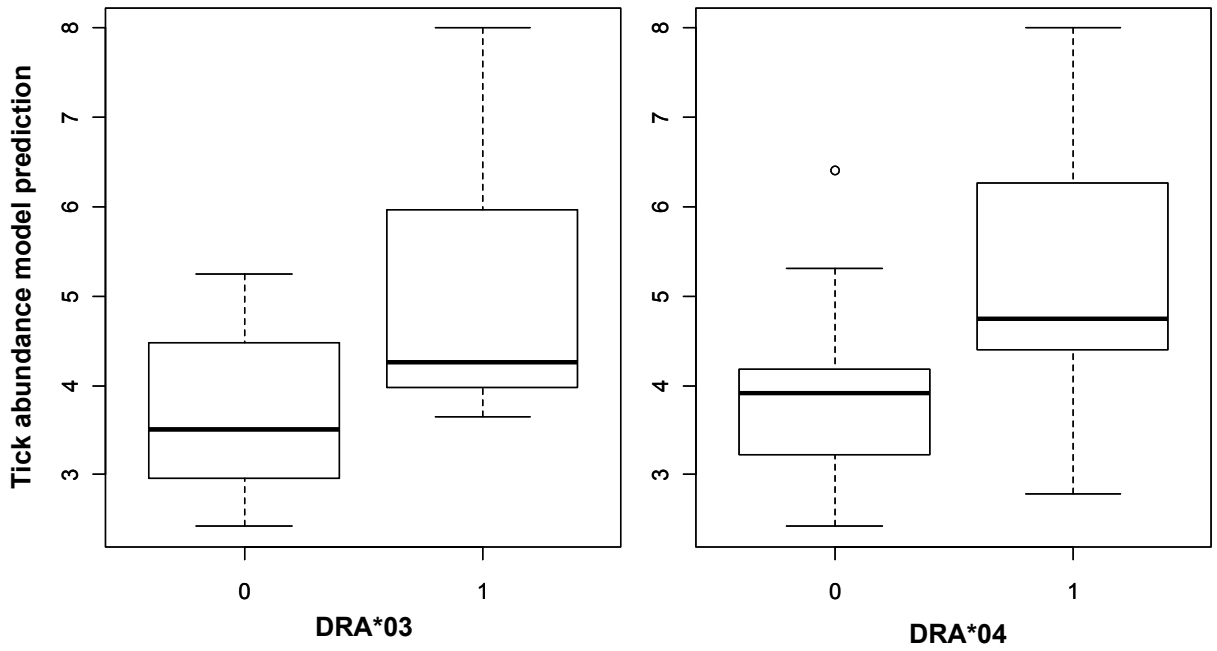

Figure 9. ELA-*DRA* alleles predicting ectoparasitism, showing susceptibility effects of common alleles, ELA-*DRA03 and *DRA**04**

DRA allele- parasite relationships depicted in terms of the presence (1) or absence (0) of the allele in the *DRA* genotype. Only predictions of the best-fit allelic ectoparasite model are shown here. Ninety-five percent confidence intervals are indicated by the dashed lines.

Supplementary materials

Table S1. Model selection results for the evaluating the effects of non-genetic and genetic variables on GI parasite intensity (GI)

MODEL: GI[1]	QIC	ΔQIC	% reduction from QIC_{null}	*w*	Comments
GI~ SEX + AGE + SEASON + SNPS + D3 + D4 + D7 + D9 + D10 + D11	**-149697**	**0.0**	**-0.3348**	**0.983**	***Best-fit Allele***
GI~ SEX + AGE + SEASON + SNPS + D1 + D3 + D4+ D5+ D7+ D9+ D10+ D11	**-149688**	**8.6**	**-0.3347**	**0.013**	***Max. Allele***
GI~ SEX + AGE + SEASON + SNPS + D3 + D4 + D5 + D7 + D9 + D10 + D11	-149686	11.1	-0.3347	0.004	
GI~ SEX + AGE + SEASON + SNPS * COMM + RARE	**-112238**	**37458.9**	**-0.0008**	**0.000**	***Best-fit Genetic***
GI~ AGE + SEASON + SEX + D4 + D7 + D9 + D11	-112229	37467.8	-0.0007	0.000	
GI~ SEX + AGE + SEASON + SNPS * COMM + MID + RARE	-112227	37469.6	-0.0007	0.000	
GI~ AGE + SEASON + SEX + D4 + D5 + D7 + D9 + D11	-112225	37472.2	-0.0006	0.000	
GI~ AGE + SEASON + D4 + D7 + D9 + D11	-112222	37474.7	-0.0006	0.000	
GI~ AGE + SEASON + SEX + D4 + D7 + D11	-112217	37479.8	-0.0006	0.000	
GI~ AGE + SEASON + SEX + D1 + D4 + D5 + D7 + D9 + D11	-112214	37482.8	-0.0005	0.000	
GI~ AGE + SEASON + SEX + D1 + D4 + D5 + D7 + D9 + D10 + D11	-112213	37483.9	-0.0005	0.000	
GI~ AGE + SEASON + SEX + D4 + D9 + D11	-112207	37490.1	-0.0005	0.000	
GI~ AGE + SEASON + SEX + D1 + D3 + D4 + D5 + D7 + D9 + D10 + D11	-112206	37491.0	-0.0005	0.000	
GI~ SEX + AGE + SEASON + SNPS + MID+ RARE + COMM	**-112199**	**37498.3**	**-0.0004**	**0.000**	***Max.Genetic***
GI~ MID + SEASON + RARE + COMM + SNPS + AGE	-112183	37514.1	-0.0003	0.000	
GI~ SEX + AGE + SEASON + HET + GENO.FREQ	-112182	37514.9	-0.0003	0.000	
GI~ SEASON + SEX + AGE + HET + GENO.FREQ	-112182	37514.9	-0.0003	0.000	
GI~ SEX + AGE + SEASON + MID+ RARE + COMM	-112181	37515.3	-0.0003	0.000	
GI~ AGE + RARE	-112179	37517.5	-0.0002	0.000	
GI~ SEASON + RARE + COMM + SNPS	-112179	37517.8	-0.0002	0.000	
GI~ SEX + AGE + SEASON + HET + MID + RARE + COMM	-112178	37518.9	-0.0002	0.000	
GI~ SEASON + D4 + D1 + D9 + D11	-112178	37518.9	-0.0002	0.000	
GI~ SEX + AGE + SEASON + GENO.FREQ	-112177	37519.9	-0.0002	0.000	
GI~ SEASON + AGE + HET + RARE	-112172	37524.8	-0.0002	0.000	
GI~ SEASON + D4 + D1 + D9	-112170	37526.3	-0.0002	0.000	
GI~ SEX + AGE + SEASON + SNPS + GENO.FREQ	-112170	37527.3	-0.0001	0.000	
GI~ AGE + SEASON + SEX + D7	-112169	37527.7	-0.0001	0.000	
GI~ SEASON + D4 + D1 + D7	-112168	37529.0	-0.0001	0.000	
GI~ AGE + SEX + SEASON + D3	-112167	37530.1	-0.0001	0.000	
GI~ SEASON + HET + COMM + MID + SEX + AGE	-112163	37533.3	-0.0001	0.000	
GI~ MID + SEASON + RARE + COMM + SNPS	-112163	37533.5	-0.0001	0.000	
GI~ SEASON + HET + COMM + MID + SEX	-112161	37535.5	-0.0001	0.000	
GI~ AGE + SEASON + SEX + D9	-112161	37535.6	-0.0001	0.000	
GI~ SEX + AGE + SEASON + SNPS	-112160	37537.1	-0.0001	0.000	
GI~ AGE + SEASON + SEX + D11	-112158	37538.3	-0.0001	0.000	
GI~ SEASON + AGE + HET + GENO.FREQ	-112158	37539.1	0.0000	0.000	
GI~ SEASON + AGE + SEX + D4 + D11	-112158	37539.1	0.0000	0.000	
GI~ SEASON + AGE + SEX + D7 + D9	-112157	37540.2	0.0000	0.000	
GI~ SEASON + AGE + SEX + D7 + D11	-112154	37542.8	0.0000	0.000	
GI~ SEX + AGE + SEASON	**-112153**	**37544.0**	**0.0000**	**0.000**	***Best-fit & Max.Null***

Over 100 models were fit to the data, including combinations of the variables described in Table 2. Only the top 40 models with the lowest Quasi-Information Criteria (QIC; Pan 2001) are shown. Difference from best-fit model ($\Delta QIC = QIC_i - QIC_{min}$), percent reduction from the null QIC (QIC_{null}) and QIC weights (*w*) are reported. Models are indicated in the comments.

[1] *A square-root transformation was applied to the GI parasite response variable, GI.*

Table S2. Model selection results for the evaluating the effects of non-genetic and genetic variables on ectoparasite burden (ECTO)

MODEL: ECTO	QIC	ΔQIC	% reduction from QIC_{null}	*w*	Comments
ECTO ~ SEX+AGE+SEASON+SNPS+D1+D3+D4+D5 +D7 +D9 + D10 + D11	**-2252**	**0**	**-0.316**	**1.00**	***Best-fit & Max. Allele***
ECTO ~ AGE + SEASON + SNPS + D1 + D3 + D4 + D5 + D7 + D9 + D10 + D11	-1797	455	-0.050	0.00	
ECTO ~ SEX + SEASON + SNPS + D1 + D3 + D4 + D5 + D7 + D9 + D10 + D11	-1796	456	-0.050	0.00	
ECTO ~ SEX + AGE + SEASON + SNPS + D3 + D4	-1794	458	-0.049	0.00	
ECTO ~ SEASON + D1 + D3 + D4 + D5 + D9 + D7 + D10 + D11	-1793	460	-0.048	0.00	
ECTO ~ SEASON + AGE + D1 + D3 + D4 + D5 + D7 + D9 + D10 + D11	-1791	461	-0.047	0.00	
ECTO ~ SEASON + SEX + D1 + D3 + D4 + D5 + D7 + D9 + D10 + D11	-1790	463	-0.046	0.00	
ECTO ~ D1 + D3 + D4 + D5 + D7 + D9 + D10 + D11	-1785	467	-0.043	0.00	
ECTO ~ AGE + SEASON + HET * COMM + MID	**-1766**	**486**	**-0.032**	**0.00**	***Best-fit Genetic***
ECTO ~ SEASON + HET + COMM + MID	-1764	488	-0.031	0.00	
ECTO ~ SEX + AGE + SEASON + HET * COMM + MID	-1763	489	-0.031	0.00	
ECTO ~ SEASON + HET + COMM + MID + AGE	-1762	490	-0.030	0.00	
ECTO ~ SEASON + HET + COMM + MID + SEX	-1761	491	-0.029	0.00	
ECTO ~ AGE + HET * COMM + MID	-1760	492	-0.029	0.00	
ECTO ~ MID + SEASON + RARE + COMM + SNPS	-1759	493	-0.028	0.00	
ECTO ~ SEASON + HET + COMM + MID + SEX + AGE	-1759	493	-0.028	0.00	
ECTO ~ HET + COMM + MID	-1757	495	-0.027	0.00	
ECTO ~ MID + SEASON + RARE + COMM + SNPS + AGE	-1757	495	-0.027	0.00	
ECTO ~ MID + SEASON + RARE + COMM + SNPS + SEX	-1756	496	-0.026	0.00	
ECTO ~ MID + RARE + COMM + SNPS	-1754	498	-0.025	0.00	
ECTO ~ SEX + AGE + SEASON + SNPS + MID+ RARE + COMM	**-1754**	**498**	**-0.025**	**0.00**	***Maximal Genetic***
ECTO ~ SEASON + HET + COMM	-1753	499	-0.025	0.00	
ECTO ~ MID + SEASON + RARE + COMM + SNPS	-1752	500	-0.024	0.00	
ECTO ~ SEASON + SNPS + COMM + MID	-1751	501	-0.023	0.00	
ECTO ~ SEASON + HET + COMM + AGE	-1751	501	-0.023	0.00	
ECTO ~ SEASON + HET + COMM + SEX	-1750	502	-0.023	0.00	
ECTO ~ MID + SEASON + RARE + COMM	-1749	503	-0.022	0.00	
ECTO ~ MID + SEASON + COMM	-1748	504	-0.022	0.00	
ECTO ~ MID + SEASON + SNPS	-1746	506	-0.021	0.00	
ECTO ~ SEASON + COMM	-1746	506	-0.020	0.00	
ECTO ~ MID + COMM + SNPS	-1746	507	-0.020	0.00	
ECTO ~ SEASON + RARE + COMM + SNPS	-1745	507	-0.020	0.00	
ECTO ~ SEASON + RARE + COMM	-1745	507	-0.020	0.00	
ECTO ~ MID + RARE + COMM	-1743	509	-0.019	0.00	
ECTO ~ SEX + AGE + SEASON + MID+ RARE + COMM	-1743	509	-0.019	0.00	
ECTO ~ MID + COMM	-1742	510	-0.018	0.00	
ECTO ~ MID + SNPS	-1741	511	-0.018	0.00	
ECTO ~ MID + SEASON + RARE	-1741	512	-0.017	0.00	
ECTO ~ SEX + AGE + SEASON + SNPS + D4	-1740	512	-0.017	0.00	
ECTO ~ AGE + COMM + MID + RARE	-1740	512	-0.017	0.00	
ECTO ~ SEX + AGE + SEASON + SNPS + D3	-1738	515	-0.016	0.00	
ECTO ~ SEX + AGE + SEASON + SNPS + D3 + D5	-1736	516	-0.015	0.00	
ECTO ~ MID + SEASON	-1736	516	-0.015	0.00	
ECTO ~ SEX + AGE + SEASON + SNPS + D9	-1735	517	-0.014	0.00	
ECTO ~ MID + RARE	-1735	517	-0.014	0.00	
ECTO ~ SEX + AGE + SEASON + HET + GENO.FREQ	-1730	522	-0.011	0.00	
ECTO ~ MID	-1730	522	-0.011	0.00	
ECTO ~ AGE + MID	-1727	525	-0.009	0.00	
ECTO ~ SEX + AGE + SEASON + HET	-1723	529	-0.007	0.00	
ECTO ~ SEX + AGE + SEASON + SNPS + D1	-1722	530	-0.007	0.00	
ECTO ~ AGE + HET	-1721	531	-0.006	0.00	
ECTO ~ SEASON + SNPS	-1718	534	-0.004	0.00	
ECTO ~ AGE + SEX + HET	-1718	534	-0.004	0.00	
ECTO ~ SEASON + RARE + SNPS	-1717	535	-0.003	0.00	
ECTO ~ SEASON + SNPS	-1716	536	-0.003	0.00	
ECTO ~ SEX + AGE + SEASON + SNPS + D10	-1713	539	-0.001	0.00	
ECTO ~ SEASON	**-1711**	**541**	**0.000**	**0.00**	***Best-fit Null***

Over 100 models were fit to the data, including combinations of the variables described in Table 2. Only the top 57 models with the lowest Quasi-Information Criteria (QIC; Pan 2001) are shown. Difference from best-fit model ($\Delta QIC = QIC_i - QIC_{min}$), percent reduction from the best-fit null QIC (QIC_{null}) and QIC weights (*w*) are reported. Models indicated in comments.

Table S3. Maximal genetic and null non-genetic models predicting parasitism

Response	% reduction from QIC_{null}	Coefficients	*Maximal genetic model*			*Null non-genetic model*		
			Estimate ± S.E.	**Wald Statistic**	***P*-value**	**Estimate ± S.E.**	**Wald Statistic**	***P*-value**
GI[1]	0.04%	Intercept	3.798 ± 0.139	746.68	**< 2e-16*** **	3.782 ± 0.077	2400.37	**< 2e-16*** **
		SEASON (wet)	0.220 ± 0.054	16.43	**5.1e-05*** **	0.220 ± 0.051	18.66	**1.6e-05*** **
		SEX(male)	0.116 ± 0.104	1.24	0.265	0.102 ± 0.099	1.07	0.3
		AGE	-0.012 ± 0.009	1.83	0.176	-0.011 ± 0.010	1.16	0.28
		SNPS	-0.026 ± 0.015	2.83	0.093			
		COMM (presence)	0.012 ± 0.109	0.01	0.914			
		MID (presence)	-0.042 ± 0.102	0.17	0.679			
		RARE (presence)	0.123 ± 0.051	5.92	**0.015***			
ECTO	2.49%	Intercept	1.469 ± 0.304	23.28	**1.4e-06*** **	1.562 ± 0.102	235.5	**<2e-16*** **
		SEASON (wet)	-0.159 ± 0.117	1.86	0.17	-0.163 ± 0.110	2.2	0.14
		SEX (male)	0.030 ± 0.201	0.02	0.88			
		AGE	0.009 ± 0.025	0.13	0.72			
		SNPS	0.064 ± 0.047	1.81	0.18			
		COMM (presence)	0.112 ± 0.203	0.31	0.58			
		MID (presence)	-0.303 ± 0.212	2.04	0.15			
		RARE (presence)	-0.243 ± 0.183	1.77	0.18			

Table shows coefficient estimates for variables in the best-fit non-genetic (null) and maximal models. GI = the intensity of gastrointestinal parasitism in eggs per gram; ECTO = intensity of ectoparasitism in terms of individual tick count. Coefficients estimates and their standard errors (Estimate ± S.E.) are shown, significance of effects was tested using the Wald Statistic and associated *p*-values ($p < 0.05$ highlighted in bold). The percent reduction in the quasi-information criterion (QIC) value from the best-fit null model (QIC_{null}) is shown.

[1] *Response variable for GI parasitism was square- root transformed*

Chapter 5. Conclusions

Overview

The co-evolutionary arms race between host and pathogen has undoubtedly shaped the diversity of genes involved in host immunity, of which Major Histocompatibility Complex (MHC) genes are considered to be paramount. This research investigated the patterns of pathogen-driven selection on MHC genes to elucidate immune system adaptation in equid host populations and species. To do this, I used a multi-level approach─ by examining MHC gene evolution at the species, population and individual levels. The results of this study not only revealed selective processes over the history of *Equus*, but also illuminated recent MHC gene history in two naturally occurring plains zebra (*Equus quagga*) populations of southern Africa: Etosha National Park (ENP), Namibia, and Kruger National Park (KNP), South Africa. Furthermore, examination of relationships between host genetics and macro-parasite intensity not only identified putative immunogenetic factors predicting susceptibility, but also uncovered the mechanisms by which parasites shape the variation in a MHC gene within zebra of ENP.

Among species of the genus *Equus*, high levels of genetic diversity were discovered in both MHC genes inspected (ELA-*DRA* and *DQA* exon 2). The *DQA* locus, in particular, was extremely variable with pair-wise sequence divergence as high as 52%. Despite exhibiting lower levels of variation, the *DRA* locus was discovered to have higher nucleotide and protein allelic richness in equids than has previously been reported in any other vertebrate taxon. Further, both gene phylogenies displayed trans-species polymorphisms, suggesting diversity has been maintained over Equidae evolution. Selection analyses revealed heterogeneity in selective pressures across codon sites and indicated that sites predicted to be under significant positive selection were also primarily those believed to be involved in foreign antigen recognition. In summary, I concluded that balancing selection is evident at the *DQA* locus over the history of *Equus*. Whereas at the *DRA* locus there was evidence for selection, but the nature of this selection appeared to be unclear. However, diversity patterns of the *DRA* locus indicated its probable significance in equid immune system function.

Investigation of MHC gene diversity within and between two *E. quagga* populations revealed that the mode and intensity of selection on MHC genes likely varies over time and space. Although not tested directly here, this was speculated to be due to fluctuations in pathogen pressures. I also found that selection likely plays a stronger role than demography in shaping the inter- and intra-population patterns of diversity in both genes. In particular, evidence was found for balancing selection at the population-level acting on the *DQA* locus, indicated by high diversity within, but low differentiation between zebra populations. Investigation of the diversity patterns at the *DRA* indicated that selection on this locus varied geographically. Specifically, results showed *DRA* diversity was likely being maintained in KNP in contrast to evidence for directional selection and local adaptation acting on this locus in ENP. Analyses of selection and diversity suggested some degree of balancing selection was evident at the *DRA* locus, and that the skewed

allele frequency distribution in ENP may be due to strong pathogen-driven selective pressure acting specifically within this population.

In the final phase of this research, I examined the relationships between parasite intensity (gastrointestinal (GI) nematodes and ectoparasite ticks) and host immunogenetic factors to explain the unusual pattern of *DRA* variation observed in *E. quagga* of ENP. Results provided strong evidence for pathogens acting through a frequency-dependent mechanism on this locus. However, heterozygote advantage was also evident, particularly with increased nucleotide divergence between the two alleles of an individual's *DRA* genotype. Most notably, I discovered that specific alleles belonging to different frequency classes predicted parasitism, but to opposing parasite types (i.e. GI versus ectoparasites). This result implied that multiple parasites may have antagonistic effects on the same MHC locus, providing a potential explanation for the discovery of 'susceptibility alleles' in this system. Given that rare *DRA* alleles were associated with increased susceptibility to GI nematodes, I hypothesized that these parasites may negatively affect zebra fitness, exerting a strong selective role in shaping *DRA* gene diversity. Most significantly, these results suggested that an immunogenetic tradeoff may be occurring at this locus, such that a fitness benefit of GI parasite resistance incurs a cost to an individual host with increased ectoparasite burden.

Significance

In this thesis, I emphasized the value of studying functional genetic variation, due to its significance in population and/or species adaptations. Much of the research on the Equine Lymphocyte Antigen (ELA), to date, has been conducted using samples from captive or domestic individuals (e.g. Hedrick *et al.* 1999; Brown *et al.* 2004; Janova *et al.* 2009) and has focused solely on its structural and molecular characterization. To my knowledge, this was the first study to address the adaptive significance of these genes in wild equids. These results were not only significant for illuminating patterns of selection across equid populations and species, but also for uncovering the mechanisms by which pathogens promote and maintain MHC diversity. Here I found that the nature of selection was extremely complex, varying over different spatial and temporal scales. This provided support for the hypothesis of heterogeneity in pathogen-mediated selective pressures on MHC molecules. This study also highlighted the importance of considering multiple pathogens when investigating the influence of particular immunogenetic factors on host fitness consequences, due to the potential for pleiotropic gene effects. Specifically, it showed that neglecting a portion of the pathogen community may lead to a misinterpretation of the patterns of selection and diversity observed; for example, without examining multiple parasite types, the presence of 'susceptibility alleles' may have remained unexplained.

The identification of functional amino acid residues under significant positive selection as well as the associated fitness effects of specific MHC alleles on parasitism has important management implications. For example, this information may be valuable for assessing individual and population-level risk to disease. Furthermore, understanding the mechanisms by which pathogens interact with the host immune system may have implications for managing the genetic diversity of threatened or endangered species; in

particular, knowing the relative importance of heterozygote advantage versus frequency dependent selection on immune genes may be informative for the management approach implemented (e.g. for increasing heterozygosity, maintaining specific resistant alleles or removing specific susceptibility alleles). In equids, these approaches could potentially be valuable for managing some of the most endangered wild equids: the African wild ass (*E. africanus*), Grevy's zebra (*E. grevyi*) and Cape mountain zebra (*E. zebra zebra*). Beyond equids, these findings may also have eventual implications for endangered Perissodactyl species (e.g. the black rhinoceros) and ultimately could be generally extrapolated to other vertebrates. Given the significance of these genes in immune function, this study provides an approach that may be applied to other natural systems, exemplifying a means by which to evaluate adaptive processes and the evolutionary potential of a population or species.

Future directions

One major conclusion of this thesis was that MHC genes may have pleiotropic effects on multiple parasites. This finding warrants the need for further characterization of the pathogen community in ENP, in order to fully understand the story of selection at the *DRA* locus. Therefore, further work aimed at identifying unknown selective pressures, such as babesiosis or African horse sickness, is required in the ENP *E. quagga* population. These particular diseases may not cause fatality, but nonetheless may have negative fitness consequences (e.g. reduced condition) to hosts; thus, they may be involved in additional host immunogenetic tradeoffs. Furthermore, in this study, the resolution by which the relationship between host immune genes parasitism is evaluated is relatively coarse with macro-parasite species grouped into major 'types' (i.e. endo- versus ectoparasites). Given that the extreme diversity at MHC loci has been hypothesized to be driven by pathogen diversity, fine-scale resolution of parasite species richness is necessary for a better understanding of this relationship. Another significant consideration is that zebra in ENP are known to be considerably affected by anthrax (*Bacillus anthracis*), a bacterial disease causing significant mortalities annually in the *E. quagga* population (Turnbull *et al.* 1989; Lindeque & Turnbull 1994). Therefore, I intend to extend this research by investigating the role that *B. anthracis* may play in shaping MHC diversity patterns, in conjunction with macro-parasites. I expect that anthrax will have strong selective effects on immune genes and will also help to explain the lack of beneficial *DRA* alleles conferring host resistance to macro-parasites in ENP.

Future directions for this research also include the evaluation of host and pathogen relationships in KNP, in order to explain the variable immunogenetic patterns observed across zebra populations. Furthermore, this work could be expanded to *E. quagga* populations in eastern Africa, to evaluate how variation pathogen environments may influence MHC gene diversity over the extent of their geographic range. Given the contrasting patterns that were observed at the *DRA* across zebra populations, linking these patterns to the parasite communities over a broader geographic scale may help to provide a potential explanation for the observed contradictory patterns at ENP versus KNP. Also, besides spatially extending these analyses, studies inspecting this relationship over time would also be valuable. For example, a study in ENP that incorporates time

series data on *DRA* locus variation and individual parasite burden may reveal whether the skewed *DRA* allele frequency distribution is stable or fluctuating due to oscillations in pathogen pressures through time. Finally, including multiple MHC genes as predictors of parasitism in statistical analyses would be useful for understanding whether epistatic gene effects exist, a phenomenon that has been shown to be occurring at the human MHC.

It has been well-established that MHC studies are critical for understanding the genetic basis of variation in pathogen susceptibility among vertebrate host populations and species and, therefore, for elucidating adaptive co-evolutionary processes. In this study, I examined two previously identified MHC class II genes, the ELA-*DRA* and *DQA* locus; however, little work has focused on other ELA genes and data is particularly lacking on class I MHC genes in equids. It is also known that immune response is driven by a complexity of genetic factors, and therefore genes outside of the MHC may also play a large role in disease susceptibility (reviewed in Acevedo-Whitehouse & Cunningham 2006). In fact, one study investigating the relative contributions of MHC and non-MHC genes in immune response to tuberculosis and malaria discovered that non-MHC genes had a greater cumulative contribution to variation in disease susceptibility than MHC genes (Jepson *et al.* 1997). In a natural study, Jensen *et al.* 2008 found evidence for spatiotemporal variation in selection in markers associated with TAP (transporter associated with antigen processing) genes of brown trout (*Salmo trutta*, L.) (Jensen *et al.* 2008). Yet, still few studies on natural vertebrate populations have focused on immune or disease-related genes outside of the MHC. The recent completion of the horse genome (Wade *et al.* 2009) will facilitate the development of genetic markers for non-MHC disease-related candidate genes and provides an exceptional opportunity to advance our understanding of host-pathogen adaptations in naturally occurring equid populations.

Literature Cited

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

Acosta-Rodrigiez R, Alonso-Morales R, Balladares S, Flores-Aguilar H, Garcia-Vazquez Z, Gorodezky C (2005) Analysis of BoLA class II microsatellites in cattle infested with Boophilus microplus ticks: class II is probably associated with susceptibility. *Veterinary Parasitology* **127**, 313-321.

Aguilar A, Roemer G, Debenham S, Binns M, Garcelon D, Wayne RK (2004) High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 3490-3494.

Akashi H (1999) Within- and between-species DNA sequence variation and the 'footprint' of natural selection. *Gene* **238**, 39-51.

Albright-Fraser DG, Reid R, Gerber V, Bailey E (1996) Polymorphism of DRA among equids. *Immunogenetics* **43**, 315-317.

Alcaide M, Edwards SV, Negro JJ, Serrano D, Tella JL (2008) Extensive polymorphism and geographical variation at a positively selected MHC class IIB gene of the lesser kestrel (Falco naumanni). *Molecular Ecology* **17**, 2652-2665.

Allen JR, Kemp DH (1982) Observations on the behavior of Dermacentor andersoni larvae infesting normal and tick resistant guinea pigs. *Parasitology* **84**, 195-&.

Allison AC (1954) Protection afforded by sickle cell trait against subtertian malarial infection. *British Medical Journal* **1**, 290-294.

Altuvia Y, Margalit H (2004) A structure-based approach for prediction of MHC-binding peptides. *Methods* **34**, 454-459.

Anisimova M (2003) Detecting positive selection with likelihood ratio tests and empirical Bayesian approach: An example study of the hepatitis delta antigen gene. *Infection Genetics and Evolution* **2**, 259.

Anisimova M, Bielawski JP, Yang ZH (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular Biology and Evolution* **18**, 1585-1592.

Anisimova M, Bielawski JP, Yang ZH (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Molecular Biology and Evolution* **19**, 950-958.

Anisimova M, Nielsen R, Yang ZH (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* **164**, 1229-1236.

Ansari HA, Hediger R, Fries R, Stranzinger G (1988) Chromosomal localization of the Major Histocompatibility Complex of the horse (ELA) by insitu hybridization. *Immunogenetics* **28**, 362-364.

Apanius V, Penn D, Slev PR, Ruff LR, Potts WK (1997) The nature of selection on the major histocompatibility complex. *Critical Reviews in Immunology* **17**, 179-224.

Avise JC (2004) *Molecular Markers, Natural History, and Evolution* Sinauer Associates, Inc., Sunderland, MA.

Bailey E (1994) Variation within the antigen binding site of the major histocompatibility complex gene of domestic horses, 123-126.
Bailey E, Marti E, Fraser DG, Antczak DF, Lazary S (2000) Immunogenetics of the horse. In: *The Genetics of the Horse* (eds. Bowling A, Ruvinsky A), pp. 123-156. CAB International Publishing, New York.
Balakrishnan CN, Ekblom R, Volker M, Westerdahl H, Godinez R, Kotkiewicz H, Burt DW, Graves T, Griffin DK, Warren WC, Edwards SV (2010) Gene duplication and fragmentation in the zebra finch major histocompatibility complex. *Bmc Biology* **8**, 19.
Balloux F, Lugon-Moulin N (2002) The estimation of population differentiation with microsatellite markers. *Molecular Ecology* **11**, 155-165.
Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 4563-4568.
Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *Journal of Evolutionary Biology* **16**, 363-377.
Bielawski JP, Yang Z (2003) Maximum likelihood methods for detecting adaptive evolution after gene duplication. *Journal of Structural and Functional Genomics* **3**, 201-212.
Biggs HC, Langenhoven JW (1984) Seasonal prevalence of Ixodid ticks on cattle in the Windhoek district of Southwest Africa Namibia. *Onderstepoort Journal of Veterinary Research* **51**, 175-182.
Binns MM, Holmes NG, Holliman A, Scott AM (1995) The identification of polymorphic microsatellite loci in the horse and their use in thoroughbred parentage testing. *British Veterinary Journal* **151**, 9-15.
Black FL, Salzano FM (1981) Evidence for heterosis in the HLA System. *American Journal of Human Genetics* **33**, 894-899.
Blouin MS, Parsons M, Lacaille V, Lotz S (1996) Use of microsatellite loci to classify individuals by relatedness. *Molecular Ecology* **5**, 393-401.
Bohannan BJM, Travisano M, Lenski RE (1999) Epistatic interactions can lower the cost of resistance to multiple consumers. *Evolution* **53**, 292-295.
Bondinas GP, Moustakas AK, Papadopoulos GK (2007) The spectrum of HLA-DQ and HLA-DR alleles, 2006: a listing correlating sequence and structure with function. *Immunogenetics* **59**, 539-553.
Bos DH, Gopurenko D, Williams RN, DeWoody JA (2008) Inferring population history and demography using microsatellites, mitochondrial DNA, and major histocompatibility complex (MHC) genes. *Evolution* **62**, 1458-1468.
Bos DH, Turner SM, Dewoody JA (2007) Haplotype inference from diploid sequence data: evaluating performance using non-neutral MHC sequences. *Hereditas* **144**, 228-234.
Bowman DD (2003) *Georgis' parasitology for veterinarians.*, 8th edn. W. B. Saunders, Philadelphia, PA.

Boyce WM, Hedrick PW, MuggliCockett NE, Kalinowski S, Penedo MCT, Ramey RR (1997) Genetic variation of major histocompatibility complex and micrcosatellite loci: A comparison in bighorn sheep. *Genetics* **145**, 421-433.

Bradley RD, Hillis DM (1997) Recombinant DNA sequences generated by PCR amplification. *Molecular Biology and Evolution* **14**, 592-593.

Brinkmeyer-Langford CL, Murphy WJ, Childers CP, Skow LC (2010) A conserved segmental duplication within ELA. *Animal Genetics* **41**, 186-195.

Brown JJ, Thomson W, Clegg P, Eyre S, Kennedy LJ, Matthews J, Carter S, Ollier WER (2004) Polymorphisms of the equine major histocompatibility complex class II DRA locus. *Tissue Antigens* **64**, 173-179.

Bruford MW, Wayne RK (1993) Microsatellites and their application to population genetic studies. *Current Opinion in Genetics and Development* **3**, 939-943.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution* **18**, 249-256.

Bryja J, Galan M, Charbonnel N, Cosson JF (2006) Duplication, balancing selection and trans-species evolution explain the high levels of polymorphism of the DQA MHC class II gene in voles (Arvicolinae). *Immunogenetics* **58**, 191-202.

Buitkamp J, Filmether P, Stear MJ, Epplen JT (1996) Class I and class II major histocompatibility complex alleles are associated with faecal egg counts following natural, predominantly *Ostertagia circumcincta* infection. *Parasitology Research* **82**, 693-696.

Burnham K, Anderson D (2002) *Model selection and multimodel inference: a practical information-theoretic approach.* Springer, New York.

Carey V (2004) yags: Yet Another GEE Solver. *R package version 4.0-2.1*.

Cassinello J, Gomendio M, Roldan ERS (2001) Relationship between coefficient of inbreeding and parasite burden in endangered gazelles. *Conservation Biology* **15**, 1171-1174.

Chapuis MP, Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution* **24**, 621-631.

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

Chardon P, Renard C, Vaiman M (1999) The major histocompatibility complex in swine. *Immunological Reviews* **167**, 179-192.

Chen JJ, Hollenbach JA, Trachtenberg EA, Just JJ, Carrington M, Ronningen KS, Begovich A, King MC, McWeeney S, Mack SJ, Erlich HA, Thomson G (1999) Hardy-Weinberg testing for HLA class II (DRB1, DQA1, DQB1, AND DPB1) loci in 26 human ethnic groups. *Tissue Antigens* **54**, 533-542.

Chen YY, Zhang YY, Zhang HM, Ge YF, Wan QH, Fang SG (2010) Natural selection coupled with intragenic recombination shapes diversity patterns in the Major Histocompatibility Complex class II genes of the giant panda. *Journal of Experimental Zoology Part B-Molecular and Developmental Evolution* **314B**, 208-223.

Chu ZTE, Carswellcrumpton C, Cole BC, Jones PP (1994) The minimal polymorphism of class-II E-alpha chains is not due to the functional neutrality of mutations. *Immunogenetics* **40**, 9-20.
Clement M, Posada D, Crandall KA (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology* **9**, 1657-1659.
Coltman DW, Pilkington JG, Smith JA, Pemberton JM (1999) Parasite-mediated selection against inbred Soay sheep in a free-living, island population. *Evolution* **53**, 1259-1267.
Coogle L, Bailey E (1997) Equine dinucleotide repeat loci LEX049-LEX063. *Animal Genetics* **28**, 378-378.
Coogle L, Reid R, Bailey E (1996) Equine dinucleotide repeat loci from LEX025 to LEX033. *Animal Genetics* **27**, 289-290.
Crandall KA, Kelsey CR, Imamichi H, Lane HC, Salzman NP (1999) Parallel evolution of drug resistance in HIV: Failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Molecular Biology and Evolution* **16**, 372-382.
Croisetiere S, Tarte PD, Bernatchez L, Belhumeur P (2008) Identification of MHC class II beta resistance/susceptibility alleles to Aeromonas salmonicida in brook charr (Salvelinus fontinalis). *Molecular Immunology* **45**, 3107-3116.
Cutrera AP, Lacey EA (2006) Major histocompatibility complex variation in Talas tuco-tucos: The influence of demography on selection. *Journal of Mammalogy* **87**, 706-716.
Cutrera AP, Lacey EA (2007) Trans-species polymorphism and evidence of selection on class II MHC loci in tuco-tucos (Rodentia : Ctenomyidae). *Immunogenetics* **59**, 937-948.
Cutrera AP, Lacey EA, Mora MS, Lessa EP (2010) Effects of contrasting demographic histories on selection at major histocompatibility complex loci in two sympatric species of tuco-tucos (Rodentia: Ctenomyidae). *Biological Journal of the Linnean Society* **99**, 260-277.
Dardiri AH, Salama SA (1988) African horse sickness - an overview. *Journal of Equine Veterinary Science* **8**, 46-49.
Daszak P, Cunningham AA, Hyatt AD (2000) Wildlife ecology - Emerging infectious diseases of wildlife - Threats to biodiversity and human health. *Science* **287**, 443-449.
de Vos V (1990) The ecology of anthrax in Kruger National Park, South Africa. *Salisbury Medical Bulletin (Supplement)* **68**, 19-23.
Deter J, Bryja J, Chaval Y, Galan M, Henttonen H, Laakkonen J, Voutilainen L, Vapalahti O, Vaheri A, Salvador AR, Morand S, Cosson JF, Charbonnel N (2008) Association between the DQA MHC class II gene and Puumala virus infection in Myodes glareolus, the bank vole. *Infection Genetics and Evolution* **8**, 450-458.
Ditchkoff SS, Hoofer SR, Lochmiller RL, Masters RE, Van Den Bussche RA (2005) MHC-DRB evolution provides insight into parasite resistance in white-tailed deer. *Southwestern Naturalist* **50**, 57-64.
Dixon CJ (2010) OLFinder-a program which disentangles DNA sequences containing heterozygous indels. *Molecular Ecology Resources* **10**, 335-340.
Doherty PC, Zinkernagel RM (1975) Enhanced immunological surveillance in mice heterozygous at H-2 gene complex. *Nature* **256**, 50-52.

Drummond A, Ashton B, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A (2010) Geneious v5.0. Available from http://www.geneious.com.
Dwinger RH (1999) Ticks and tick-borne diseases of equids, 306-310.
Ebedes H, Page LA (1976) Anthrax epizootics in wildlife in the Etosha National Park, south west Africa. *Wildlife diseases.*, 519-526.
Edwards SV, Chesnut K, Satta Y, Wakeland EK (1997) Ancestral polymorphism of MHC class II genes in mice: Implications for balancing selection and the mammalian molecular clock. *Genetics* **146**, 655-668.
Edwards SV, Hedrick PW (1998) Evolution and ecology of MHC molecules: from genomics to sexual selection. *Trends in Ecology & Evolution* **13**, 305-311.
Eggleston-Stott ML, DelValle A, Bautista M, Dileanis S, Wictum E, Bowling AT (1997) Nine equine dinucleotide repeats at microsatellite loci UCDEQ136, UCDEQ405, UCDEQ412, UCDEQ425, UCDEQ437, UCDEQ467, UCDEQ487, UCDEQ502 and UCDEQ505. *Animal Genetics* **28**, 370-371.
Ejsmond MJ, Babik W, Radwan J (2010) MHC allele frequency distributions under parasite-driven selection: A simulation model. *Bmc Evolutionary Biology* **10**, 9.
Ekblom R, Saether SA, Jacobsson P, Fiske P, Sahlman T, Grahn M, Kalas JA, Hoglund J (2007) Spatial pattern of MHC class II variation in the great snipe (Gallinago media). *Molecular Ecology* **16**, 1439-1451.
Ellis SA, Bontrop RE, Antczak DF, Ballingall K, Davies CJ, Kaufman J, Kennedy LJ, Robinson J, Smith DM, Stear MJ, Stet RJM, Waller MJ, Walter L, Marsh SGE (2006) ISAG/IUIS-VIC Comparative MHC Nomenclature Committee report, 2005. *Immunogenetics* **57**, 953-958.
Evans ML, Neff BD (2009) Major histocompatibility complex heterozygote advantage and widespread bacterial infections in populations of Chinook salmon (Oncorhynchus tshawytscha). *Molecular Ecology* **18**, 4716-4729.
Ewens WJ (1972) Sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87-&.
Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics* **1**, 47-50.
Fabiani A, Hoelzel AR, Galimberti F, Muelbert MMC (2003) Long-range paternal gene flow in the southern elephant seal. *Science* **299**, 676-676.
Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368-376.
Figueroa F, Gunther E, Klein J (1988) MHC polymorphism predating speciation. *Nature* **335**, 265-267.
Food and Agriculture Organization (2005) The Royal Veterinary College/Food and Agricultural Organization of the United Nations: Guide to Veterinary Diagnostic Parasitology. http://www.fao.org/ag/againfo/resources/documents /Parasitology/Index/Index.htm. June 2005.
Ford MJ (2002) Applications of selective neutrality tests to molecular ecology. *Molecular Ecology* **11**, 1245-1262.
Frank SA (2000) Polymorphism of attack and defense. *Trends in Ecology & Evolution* **15**, 167-171.

Frantz AC, Pope LC, Carpenter PJ, Roper TJ, Wilson GJ, Delahay RJ, Burke T (2003) Reliable microsatellite genotyping of the Eurasian badger (Meles meles) using faecal DNA. *Molecular Ecology* **12**, 1649-1661.
Fraser DG, Bailey E (1998) Polymorphism and multiple loci for the horse DQA gene. *Immunogenetics* **47**, 487-490.
Froeschke G, Sommer S (2005) MHC class II DRB variability and parasite load in the striped mouse (Rhabdomys pumilio) in the southern Kalahari. *Molecular Biology and Evolution* **22**, 1254-1259.
Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915-925.
Garrigan D, Hedrick PW (2003) Perspective: Detecting adaptive molecular polymorphism: Lessons from the MHC. *Evolution* **57**, 1707-1722.
Gaudieri S, Dawkins RL, Habara K, Kulski JK, Gojobori T (2000) SNP profile within the human major histocompatibility complex reveals an extreme and interrupted level of nucleotide diversity. *Genome Research* **10**, 1579-1586.
George M, Ryder OA (1986) Mitochondrial-DNA evolution in the genus Equus. *Molecular Biology and Evolution* **3**, 535-546.
Goodman SJ (1997) R-ST Calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. *Molecular Ecology* **6**, 881-885.
Gordon HM, Whitlock HV (1939) A new technique for counting nematode eggs in sheep faeces. *J. Counc. Sci. ind. Res. Melbourne* **12**, pp. 50-52.
Goudet J (1995) FSTAT (Version 1.2): A computer program to calculate F-statistics. *Journal of Heredity* **86**, 485-486.
Gould SJ, Hildreth JEK, Booth AM (2004) The evolution of alloimmunity and the genesis of adaptive immunity. *Quarterly Review of Biology* **79**, 359-382.
Guerin G, Bertaud M, Amigues Y (1994) Characterization of 7 New Horse Microsatellites - Hms1, Hms2, Hms3, Hms5, Hms6, Hms7 and Hms8. *Animal Genetics* **25**, 62.
Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**, 696-704.
Gulland FMD (1992) The role of nematode parasites in Soay Sheep (Ovis aries L) mortality during a population crash. *Parasitology* **105**, 493-503.
Gulland FMD, Fox M (1992) Epidemiology of nematode infections of Soay Sheep (Ovis aries L) on St-Kilda. *Parasitology* **105**, 481-492.
Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48**, 361-372.
Gustafson AL, Tallmadge RL, Ramlachan N, Miller D, Bird H, Antczak DF, Raudsepp T, Chowdhary BP, Skow LC (2003) An ordered BAC contig map of the equine major histocompatibility complex. *Cytogenetic and Genome Research* **102**, 189-195.
Halekoh U, Hojsgaard S, Yan J (2006) The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software* **15**, 1-11.
Hall TA (1999) BioEdit: a user friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**, 95-98.

Hambuch TM, Lacey EA (2002) Enhanced selection for MHC diversity in social tuco-tucos. *Evolution* **56**, 841-845.
Hansen H, Ben-David M, McDonald DB (2008) Effects of genotyping protocols on success and errors in identifying individual river otters (Lontra canadensis) from their faeces. *Molecular Ecology Resources* **8**, 282-289.
Hardin JW, Hilbe JM (2003) *Generalized Estimating Equations* Chapman & Hall/CRC, Boca Raton.
Hardy GH (1908) Mendelian proportions in a mixed population. *Science* **28**, 49-50.
Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST (1998) Genetic traces of ancient demography. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 1961-1967.
Harrigan RJ, Mazza ME, Sorenson MD (2008) Computation vs. cloning: evaluation of two methods for haplotype determination. *Molecular Ecology Resources* **8**, 1239-1248.
Harris H (1966) Enzyme polymorphisms in man. *Proceedings of the Royal Society Series B-Biological Sciences* **164**, 298-&.
Hedrick P, Kim T (1998) Genetics of complex polymorphisms: parasites and maintenance of MHC variation. In: *Genetics, Evolution, and Society* (eds. Singh RS, Krimbas CB), pp. 205-233. Harvard University Press, Cambridge.
Hedrick PW (1994) Evolutionary genetics of the Major Histocompatibility Complex. *American Naturalist* **143**, 945-964.
Hedrick PW (1999) Perspective: Highly variable loci and their interpretation in evolution and conservation. *Evolution* **53**, 313-318.
Hedrick PW (2002) Pathogen resistance and genetic variation at MHC loci. *Evolution* **56**, 1902-1908.
Hedrick PW, Lee RN, Parker KM (2000a) Major histocompatibility complex (MHC) variation in the endangered Mexican wolf and related canids. *Heredity* **85**, 617-624.
Hedrick PW, Parker KM, Gutierrez-Espeleta GA, Rattink A, Lievers K (2000b) Major histocompatibility complex variation in the Arabian oryx. *Evolution* **54**, 2145-2151.
Hedrick PW, Parker KM, Miller EL, Miller PS (1999) Major histocompatibility complex variation in the endangered Przewalski's horse. *Genetics* **152**, 1701-1710.
Hill AVS, Allsopp CEM, Kwiatkowski D, Anstey NM, Twumasi P, Rowe PA, Bennett S, Brewster D, McMichael AJ, Greenwood BM (1991) Common West African Hla Antigens Are Associated with Protection from Severe Malaria. *Nature* **352**, 595-600.
Hill RE, Hastie ND (1987) Accelerated evolution in the reactive center regions of serine protease inhibitors. *Nature* **326**, 96-99.
Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65-70.
Hopman TJ, Han EB, Story MR, Schug MD, Aquadro CF, Bowling AT, Murray JD, Caetano AR, Antczak DF (1999) Equine dinucleotide repeat loci COR001-COR020. *Animal Genetics* **30**, 225-226.
Horak IG, Anthonissen M, Krecek RC, Boomker J (1992) Arthropod parasites of springbok, gemsbok, kudus, giraffes and Burchell and Hartmann zebras in the

Etosha and Hardap Nature-Reserves, Namibia. *Onderstepoort Journal of Veterinary Research* **59**, 253-257.
Horak IG, Biggs HC, Reinecke RK (1984a) Arthropod parasites of Hartmann mountain zebra, Equus zebra hartmannae, in Southwest Africa Namibia. *Onderstepoort Journal of Veterinary Research* **51**, 183-187.
Horak IG, Devos V, Deklerk BD (1984b) Parasites of domestic and wild animals in South-Africa .17. Arthropod parasites of Burchell zebra, Equus burchelli, in the Eastern Transvaal Lowveld. *Onderstepoort Journal of Veterinary Research* **51**, 145-154.
Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA-dequence data. *Genetics* **132**, 583-589.
Hughes AL, Hughes MK (1995) Natural selection on the peptide-binding regions of Major Histocompatibility Complex molecules. *Immunogenetics* **42**, 233-243.
Hughes AL, Nei M (1988) Pattern of nucleotide substitution at Major Histocompatibility Complex class-I loci reveals overdominant selection. *Nature* **335**, 167-170.
Hughes AL, Nei M (1989) Nucleotide substitution at Major Histocompatibility Complex class-II loci - Evidence for overdominant selection. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 958-962.
Hughes AL, Nei M (1992) Maintenance of MHC polymorphism. *Nature* **355**, 402-403.
Huntley BJ (1982) Southern African savannas. In: *Ecology of Tropical Savannas* (eds. Huntely BJ, Walker BH), pp. 101-119. Springer, Berlin, Germany.
Hurlbert SH (1971) Nonconcept of species diversity - Critique and alternative parameters. *Ecology* **52**, 577-&.
Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**, 254-267.
Irvin Z, Giffard J, Brandon R, Breen M, Bell K (1998) Equine dinucleotide repeat polymorphisms at loci ASB 21, 23, 25 and 37-43. *Animal Genetics* **29**, 67-67.
Janova E, Matiasovic J, Vahala J, Vodicka R, Van Dyk E, Horin P (2009) Polymorphism and selection in the major histocompatibility complex DRA and DQA genes in the family Equidae. *Immunogenetics* **61**, 513-527.
Jansen R, Ledley FD (1990) Disruption of phase during PCR amplification and cloning of heterozygous tareget sequences. *Nucleic Acid Research* **18**, 5153-5156.
Jensen LF, Hansen MM, Mensberg KL, Loeschcke V (2008) Spatially and temporally fluctuating selection at non-MHC immune genes: evidence from TAP polymorphism in populations of brown trout (Salmo trutta, L.). *Heredity* **100**, 79-91.
Jepson A, Banya W, Sisay-Joof F, Hassan-King M, Nunes C, Bennett S, Whittle H (1997) Quantification of the relative contribution of major histocompatibility complex (MHC) and non-MHC genes to human immune responses to foreign antigens. *Infection and Immunity* **65**, 872-876.
Jordan WC, Bruford MW (1998) New perspectives on mate choice and the MHC. *Heredity* **81**, 127-133.
Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: *Mammalian protein metabolism, III* (ed. Munro HN), pp. 21-132. Academic Press, New York.
Kalinowski ST (2005) HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness. *Molecular Ecology Notes* **5**, 187-189.

Kamath PL, Getz WM (2011) Adaptive molecular evolution of the Major Histocompatibility Complex genes, DRA and DQA, in the genus Equus. *Bmc Evolutionary Biology*.
Kimura M (1983) The neutral theory of molecular evolution. In: *The neutral theory of molecular evolution.*, p. x + 367 pp. Cambridge University Press, Cambridge UK.
Klein J (1980) *Generation of diversity* Academic Press, London.
Klein J (1986) *Natural History of the Major Histocompatibility Complex* Wiley & Sons, New York.
Klein J (1987) Origin of major histocompatibility complex polymorphism: the trans-species hypothesis. *Human Immunology* **19**, 155-162.
Klein J, Satta Y, Takahata N, Ohuigin C (1993) Trans-specific MHC polymorphism and the origin of species in primates. *Journal of Medical Primatology* **22**, 57-64.
Kloch A, Babik W, Bajer A, Sinski E, Radwan J (2010) Effects of an MHC-DRB genotype and allele number on the load of gut parasites in the bank vole Myodes glareolus. *Molecular Ecology* **19**, 255-265.
Kojima K (1971) Is there a constant fitness value for a given genotype? No! *Evolution* **25**, 281-285.
Krecek RC, Malan FS, Reinecke RK, de Vos V (1987a) Nematode parasites from Burchell's zebras in South Africa. *Journal of Wildlife Diseases* **23**, 401-411.
Krecek RC, Reinecke RK, Malan FS (1987b) Studies on the parasites of zebras .V. Nematodes of the Burchell and Hartmann mountain zebras from the Etosha National Park, South West Africa/ Namibia. *Onderstepoort Journal of Veterinary Research* **54**, 71-78.
Kruger K, Gaillard C, Stranzinger G, Rieder S (2005) Phylogenetic analysis and species allocation of individual equids using microsatellite data. *Journal of Animal Breeding and Genetics* **122**, 78-86.
Kuhner MK (2006) LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* **22**, 768-770.
Kundu S, Faulkes CG (2007) A tangled history: patterns of major histocompatibility complex evolution in the African mole-rats (Family : Bathyergidae). *Biological Journal of the Linnean Society* **91**, 493-503.
Landry C, Bernatchez L (2001) Comparative analysis of population structure across environments and geographical scales at major histocompatibility complex and microsatellite loci in Atlantic salmon (Salmo salar). *Molecular Ecology* **10**, 2525-2539.
Langefors A, Lohm J, Grahn M, Andersen O, von Schantz T (2001) Association between major histocompatibility complex class IIB alleles and resistance to Aeromonas salmonicida in Atlantic salmon. *Proceedings of the Royal Society of London Series B-Biological Sciences* **268**, 479-485.
Lenski RE (1988) Experimental studies of pleiotropy and epistasis in Escherichia coli .1. Variation in competitive fitness among mutants resistant to virus T4. *Evolution* **42**, 425-432.
Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. *Genetics* **74**, 175-195.
Liang KY, Zeger SL (1986) Longitudinal data analysis using Generalized Linear Models. *Biometrika* **73**, 13-22.

Lindeque PM, Turnbull PCB (1994) Ecology and epidemiology of anthrax in the Etosha National Park, Namibia. *Onderstepoort Journal of Veterinary Research* **61**, 71-83.

Linn S (1981) Deoxyribonucleases: a survey and perspectives. In: *Enzymes* (ed. Boyer ED), pp. 131-145. Academic Press, New York.

Loiseau C, Richard M, Garnier S, Chastel O, Julliard R, Zoorob R, Sorci G (2009) Diversifying selection on MHC class I in the house sparrow (Passer domesticus). *Molecular Ecology* **18**, 1331-1340.

Loiseau C, Zoorob R, Garnier S, Birard J, Federici P, Julliard R, Sorci G (2008) Antagonistic effects of a MHC class I allele on malaria-infected house sparrows. *Ecology Letters* **11**, 258-265.

Lorenzen ED, Arctander P, Siegismund HR (2008) High variation and very low differentiation in wide ranging plains zebra (Equus quagga) insights from mtDNA and microsatellites. *Molecular Ecology* **17**, 2812-2824.

Lowe A, Harris S, Ashton P (2004) *Ecological genetics: Design, Analysis, and Application* Blackwell Publishing Ltd., Oxford, U.K.

Luikart G, Sherwin WB, Steele BM, Allendorf FW (1998) Usefulness of molecular markers for detecting population bottlenecks via monitoring genetic change. *Molecular Ecology* **7**, 963-974.

Luis C, Cothran EG, Oom MM, Bailey E (2005) Major histocompatibility complex locus DRA polymorphism in the endangered Sorraia horse and related breeds. *Journal of Animal Breeding and Genetics* **122**, 69-72.

Lynch M (1990) The similarity index and DNA fingerprinting. *Molecular Biology and Evolution* **7**, 478-484.

Lynch M, Crease TJ (1990) The analysis of population survey data on DNA sequence variation. *Molecular Biology and Evolution* **7**, 377-394.

Lynch M, Pfrender M, Spitze K, Lehman N, Hicks J, Allen D, Latta L, Ottene M, Bogue F, Colbourne J (1999) The quantitative and molecular genetic architecture of a subdivided species. *Evolution* **53**, 100-110.

Madden DR (1995) The 3-dimensional structure of peptide-MHC complexes. *Annual Review of Immunology* **13**, 587-622.

Mäkinen A, Chowdhary B, Mahdy E, Andersson L, Gustavsson I (1989) Localization of the equine Major Histocompatibility Complex (ELA) to chromosome-20 by insitu hybridization. *Hereditas* **110**, 93-96.

Margolis L, Esch GW, Holmes JC, Kuris AM, Schad GA (1982) The use of ecological terms in parasitology (report of an ad hoc committee of the American Society of Parasitologists). *Journal of Parasitology* **68**, 131-133.

Marklund S, Ellegren H, Eriksson S, Sandberg K, Andersson L (1994) Parentage testing and linkage analysis in the horse using a set of highly polymorphic microsatellites. *Animal Genetics* **25**, 19-23.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P (2010) RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**, 2462-2463.

Martinsohn JT, Sousa AB, Guethlein LA, Howard JC (1999) The gene conversion hypothesis of MHC evolution: a review. *Immunogenetics* **50**, 168-200.

Matthee S, Krecek RC, McGeoch MA (2004) A comparison of the intestinal helminth communities of equidae in Southern Africa. *Journal of Parasitology* **90**, 1263-1273.
McClelland EE, Penn DJ, Potts WK (2003) Major histocompatibility complex heterozygote superiority during coinfection. *Infection and Immunity* **71**, 2079-2086.
Meyer-Lucht Y, Sommer S (2005) MHC diversity and the association to nematode parasitism in the yellow-necked mouse (*Apodemus flavicollis*). *Molecular Ecology* **14**, 2233-2243.
Meyer AH, Valberg SJ, Hillers KR, Schweitzer JK, Mickelson JR (1997) Sixteen new polymorphic equine microsatellites. *Animal Genetics* **28**, 69-70.
Meyer D, Thomson G (2001) How selection shapes variation of the human major histocompatibility complex: a review. *Annals of Human Genetics* **65**, 1-26.
Miller HC, Allendorf F, Daugherty CH (2010) Genetic diversity and differentiation at MHC genes in island populations of tuatara (Sphenodon spp.). *Molecular Ecology* **19**, 3894-3908.
Miller KM, Kaukinen KH, Beacham TD, Withler RE (2001) Geographic heterogeneity in natural selection on an MHC locus in sockeye salmon. *Genetica* **111**, 237-257.
Mona S, Crestanello B, Bankhead-Dronnet S, Pecchioli E, Ingrosso S, D'Amelio S, Rossi L, Meneguz PG, Bertorelle G (2008) Disentangling the effects of recombination, selection, and demography on the genetic variation at a major histocompatibility complex class II gene in the alpine chamois. *Molecular Ecology* **17**, 4053-4067.
Morens DM, Folkers GK, Fauci AS (2004) The challenge of emerging and re-emerging infectious diseases. *Nature* **430**, 242-249.
Nei M (1987) Molecular evolutionary genetics. In: *Molecular evolutionary genetics.*, pp. i-x, 1-512. Columbia University Press, New York & Guildford, England.
Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics* **39**, 121-152.
Nei M, Tajima F (1981) DNA polymorphism detectable by restriction endonucleases. *Genetics* **97**, 145-163.
Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**, 641-647.
Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics* **39**, 197-218.
Nielsen R, Yang ZH (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929-936.
O'Connor SL, Blasky AJ, Pendley CJ, Becker EA, Wiseman RW, Karl JA, Hughes AL, O'Connor DH (2007) Comprehensive characterization of MHC class II haplotypes in Mauritian cynomolgus macaques. *Immunogenetics* **59**, 449-462.
Oakenfull EA, Clegg JB (1998) Phylogenetic relationships within the genus Equus and the evolution of alpha and theta globin genes. *Journal of Molecular Evolution* **47**, 772-783.
Oliver MK, Lambin X, Cornulier T, Piertney SB (2009a) Spatio-temporal variation in the strength and mode of selection acting on major histocompatibility complex diversity in water vole (Arvicola terrestris) metapopulations. *Molecular Ecology* **18**, 80-92.

Oliver MK, Telfer S, Piertney SB (2009b) Major histocompatibility complex (MHC) heterozygote superiority to natural multi-parasite infections in the water vole (Arvicola terrestris). *Proceedings of the Royal Society B-Biological Sciences* **276**, 1119-1128.

Omar RZ, Wright EM, Turner RM, Thompson SG (1999) Analysing repeated measurements data: A practical comparison of methods. *Statistics in Medicine* **18**, 1587-1603.

Oppelt C, Starkloff A, Rausch P, Von Holst D, Rodel HG (2010) Major histocompatibility complex variation and age-specific endoparasite load in subadult European rabbits. *Molecular Ecology* **19**, 4155-4167.

Otting N, de Groot NG, Doxiadis GGM, Bontrop RE (2002) Extensive MHC-DQB variation in humans and non-human primate species. *Immunogenetics* **54**, 230-239.

Otto SP (2004) Two steps forward, one step back: the pleiotropic effects of favoured alleles. *Proceedings of the Royal Society of London Series B-Biological Sciences* **271**, 705-714.

Ottova E, Simkova A, Martin JF, de Bellocq JG, Gelnar M, Allienne JF, Morand S (2005) Evolution and trans-species polymorphism of MHC class II beta genes in cyprinid fish. *Fish & Shellfish Immunology* **18**, 199-222.

Owen-Smith N, Mason DR (2005) Comparative changes in adult vs. juvenile survival affecting population trends of African ungulates. *Journal of Animal Ecology* **74**, 762-773.

Paetkau D, Waits LP, Clarkson PL, Craighead L, Strobeck C (1997) An empirical evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations. *Genetics* **147**, 1943-1957.

Pan W (2001) Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120-125.

Paterson S, Wilson K, Pemberton JM (1998) Major histocompatibility complex variation associated with juvenile survival and parasite resistance in a large unmanaged ungulate population (*Ovis aries* L.). *Proceedings of the National Academy of Sciences of the United States of America* **95**, 3714-3719.

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**, 288-295.

Penn DJ, Damjanovich K, Potts WK (2002) MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 11260-11264.

Piertney SB, Oliver MK (2006) The evolutionary ecology of the major histocompatibility complex. *Heredity* **96**, 7-21.

Piry S, Luikart G, Cornuet JM (1999) BOTTLENECK: A computer program for detecting recent reductions in the effective population size using allele frequency data. *Journal of Heredity* **90**, 502-503.

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817-818.

Potts WK, Slev PR (1995) Pathogen-based models favoring MHC genetic diversity. *Immunological Reviews* **143**, 181-197.

Proctor H, Owens I (2000) Mites and birds: diversity, parasitism and coevolution. *Trends in Ecology & Evolution* **15**, 358-364.
Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F (2005) Pathogen-driven selection and worldwide HLA class I diversity. *Current Biology* **15**, 1022-1027.
Quinn GP, Keough MJ (2002) *Experimental Design and Data Analysis for Biologists* Cambridge University Press, Cambridge, UK.
Rambaut A, Drummond A (2007) Tracer v1.4. Available from http://beast.bio.ed.ac.uk/Tracer.
Raymond M, Rousset F (1995) Genepop (Version-1.2) - Population-Genetics Software for Exact Tests and Ecumenicism. *Journal of Heredity* **86**, 248-249.
Reche PA, Reinherz EL (2003) Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms. *Journal of Molecular Biology* **331**, 623-641.
Reusch TBH, Haberli MA, Aeschlimann PB, Milinski M (2001) Female sticklebacks count alleles in a strategy of sexual selection explaining MHC polymorphism. *Nature* **414**, 300-302.
Richman AD, Herrera LG, Nash D (2001) MHC class II beta sequence diversity in the deer mouse (Peromyscus maniculatus): implications for models of balancing selection. *Molecular Ecology* **10**, 2765-2773.
Rogers AR, Harpending H (1992) Population-growth makes waves in the distribution of pairwise genetic-differences. *Molecular Biology and Evolution* **9**, 552-569.
Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574.
Rousset F (1996) Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics* **142**, 1357-1362.
Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496-2497.
Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and biologist programmers. *Methods in Molecular Biology* **132**, 365-386.
Saiki RK, Gyllensten UB, Erlich HA (1988) The polymerase chain reaction. In: *Genome analysis: a practical approach.*, pp. 141-152. IRL Press, Oxford UK.
Sauermann U, Nurnberg P, Bercovitch FB, Berard JD, Trefilov A, Widdig A, Kessler M, Schmidtke J, Krawczak M (2001) Increased reproductive success of MHC class II heterozygous males among free-ranging rhesus macaques. *Human Genetics* **108**, 249-254.
Schad J, Ganzhorn JU, Sommer S (2005) Parasite burden and constitution of major histocompatibility complex in the malagasy mouse lemur, *Microcebus murinus*. *Evolution* **59**, 439-450.
Schad J, Sommer S, Ganzhorn JU (2004) MHC variability of a small lemur in the littoral forest fragments of southeastern Madagascar. *Conservation Genetics* **5**, 299-309.
Schierup MH, Vekemans X, Charlesworth D (2000) The effect of subdivision on variation at multi-allelic loci under balancing selection. *Genetical Research* **76**, 51-62.

Schwensow N, Fietz J, Dausmann KH, Sommer S (2007) Neutral versus adaptive genetic variation in parasite resistance: importance of major histocompatibility complex supertypes in a free-ranging primate. *Heredity* **99**, 265-277.
Scialdo-Krecek RC, Reinecke RK, Biggs HC (1983) Studies on the parasites of zebras .3. Nematodes of the mountain zebra from the farm Kelpie and the Namib Naukluft Park, South West Africa Namibia. *Onderstepoort Journal of Veterinary Research* **50**, 283-290.
Seddon JM, Ellegren H (2002) MHC class II genes in European wolves: a comparison with dogs. *Immunogenetics* **54**, 490-500.
Segal S, Hill AVS (2003) Genetic susceptibility to infectious disease. *Trends in Microbiology* **11**, 445-448.
Simmons MP, Ochoterena H (2000) Gaps as characters in sequence-based phylogenetic analyses. *Systematic Biology* **49**, 369-381.
Simms EL (1992) Costs of plant resistance to herbivory. In: *Plant resistance to herbivores and pathogens.* (eds. Fritz RS, Simms EL), pp. 392-425. University of Chicago Press, Chicago.
Slatkin M (1994) An exact test for neutrality based on the Ewens sampling distribution. *Genetical Research* **64**, 71-74.
Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457-462.
Slev PR, Potts WK (2002) Disease consequences of pathogen adaptation. *Current Opinion in Immunology* **14**, 609-614.
Smith S, Mang T, De Bellocq JG, Schaschl H, Zeitlhofer C, Hacklander K, Suchentrunk F (2010) Homozygosity at a class II MHC locus depresses female reproductive ability in European brown hares. *Molecular Ecology* **19**, 4131-4143.
Smuts GL (1974) Age determination in Burchells zebra Equus burchelli antiquorum from the Kruger National Park. *Journal of the Southern African Wildlife Management Association* **4**, 103-115.
Snedecor GW, Cochran WG (1989) *Statistical Methods*, 8th edn. Iowa State University Press.
Sommer S (2003) Effects of habitat fragmentation and changes of dispersal behaviour after a recent population decline on the genetic variability of noncoding and coding DNA of a monogamous Malagasy rodent. *Molecular Ecology* **12**, 2845-2851.
Sommer S (2005) The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Frontiers in Zoology* **2**, 1-18.
Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society B-Biological Sciences* **277**, 979-988.
Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**, 978-989.
Swinburne JE, Marti E, Breen M, Binns MM (1997) Characterization of twelve new horse microsatellite loci: AHT12-AHT23. *Animal Genetics* **28**, 453-453.
Swofford D (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Taberlet P, Griffin S, Goossens B, Questiau S, Manceau V, Escaravage N, Waits LP, Bouvet J (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research* **24**, 3189-3194.
Taberlet P, Waits LP, Luikart G (1999) Noninvasive genetic sampling: look before you leap. *Trends in Ecology & Evolution* **14**, 323-327.
Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437-460.
Tajima F (1989) Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595.
Takada T, Kikkawa Y, Yonekawa H, Amano T (1998) Analysis of goat MHC class II DRA and DRB genes: identification of the expressed gene and new DRB alleles. *Immunogenetics* **48**, 408-412.
Takahata N (1990) A simple genealogical structure of strongly balanced allelic lines and transspecies evolution of polymorphism. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 2419-2423.
Takahata N, Nei M (1990) Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of Major Histocompatibility Complex loci. *Genetics* **124**, 967-978.
Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24**, 1596-1599.
Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10**, 512-526.
Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In: *Miura, R. M.*, pp. 57-86.
Thoß M, Ilmonen P, Musolf K, Penn DJ (2011) Major histocompatibility complex heterozygosity enhances reproductive success. *Molecular Ecology* **20**, 1546-1557.
Trager W (1939) Acquired immunity to ticks. *Jour Parasitol* **25**, 57-81.
Traherne JA (2008) Human MHC architecture and evolution: implications for disease association studies. *International Journal of Immunogenetics* **35**, 179-192.
Trowsdale J, Groves V, Arnason A (1989) Limited MHC polymorphism in whales. *Immunogenetics* **29**, 19-24.
Turnbull PCB, Carman JA, Lindeque PM, Joubert F, Hubschle OJB, Snoeyenbos GH (1989) Further progress in understanding anthrax in the Etosha National Park, Namibia. *Madoqua* **16**, 93-104.
Turner WC (2009) *The ecology of orally ingested parasites in ungulates of Etosha National Park.* PhD, University of California at Berkeley.
Turner WC, Getz WM (2010) Seasonal and demographic factors influencing gastrointestinal parasitism in ungulates of Etosha National Park. *Journal of Wildlife Diseases* **46**, 1108-1119.
Van den Bussche RA, Hoofer SR, Lochmiller RL (1999) Characterization of MHC-DRB allelic diversity in white-tailed deer (*Odocoileus virginianus*) provides insight into MHC-DRB allelic evolution within Cervidae. *Immunogenetics* **49**, 429-437.
van Haeringen WA, van de Goor LHP, van der Hout N, Lenstra JA (1998) Characterization of 24 equine microsatellite loci. *Animal Genetics* **29**, 153-156.

van Tienderen PH, de Haan AA, van der Linden CG, Vosman B (2002) Biodiversity assessment using markers for ecologically important traits. *Trends in Ecology & Evolution* **17**, 577-582.
Van Valen L (1973) A new evolutionary law. *Evolutionary Theory* **1**, 1-30.
Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, Blocker H, Distl O, Edgar RC, Garber M, Leeb T, Mauceli E, MacLeod JN, Penedo MCT, Raison JM, Sharpe T, Vogel J, Andersson L, Antczak DF, Biagi T, Binns MM, Chowdhary BP, Coleman SJ, Della Valle G, Fryc S, Guerin G, Hasegawa T, Hill EW, Jurka J, Kiialainen A, Lindgren G, Liu J, Magnani E, Mickelson JR, Murray J, Nergadze SG, Onofrio R, Pedroni S, Piras MF, Raudsepp T, Rocchi M, Roed KH, Ryder OA, Searle S, Skow L, Swinburne JE, Syvanen AC, Tozaki T, Valberg SJ, Vaudin M, White JR, Zody MC, Lander ES, Lindblad-Toh K (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326**, 865-867.
Wagner JL, Burnett RC, Storb R (1999) Organization of the canine major histocompatibility complex: Current perspectives. *Journal of Heredity* **90**, 35-38.
Wagner JL, Derose SA, Burnett RC, Storb R (1995) Nucleotide-sequence and polymorphism analysis of canine DRA cDNA clones. *Tissue Antigens* **45**, 284-287.
Wahlund S (1928) Composition of populations and correlation appearances viewed in relation to the studies of inheritance. *Hereditas* **11**, 65-106.
Wakeland EK, Boehme S, She JX, Lu CC, McIndoe RA, Cheng I, Ye Y, Potts WK (1990) Ancestral polymorphisms of MHC class-II genes - divergent allele advantage. *Immunologic Research* **9**, 115-122.
Walker JB, Keirans JE, Horak IG (2000) *The Genus Rhipicephalus (Acari, Ixodidae): A Guide to the Brown Ticks of the World* Cambridge University Press, Cambridge.
Watterson GA (1978) Homozygosity test of neutrality. *Genetics* **88**, 405-417.
Weinberg W (1908) Über den Nachweis der Vererbung beim Menschen. *Jahreshefte des Vereins Varterländische Naturkdunde in Württemberg* **64**, 369-382.
Weir BS, Cockerham CC (1984) Estimating F-Statistics for the analysis of population structure. *Evolution* **38**, 1358-1370.
Westerdahl H, Hansson B, Bensch S, Hasselquist D (2004) Between-year variation of MHC allele frequencies in great reed warblers: selection or drift? *Journal of Evolutionary Biology* **17**, 485-492.
Wikel SK (1996) Host immunity to ticks. *Annual Review of Entomology* **41**, 1-22.
Wilson K, Grenfell BT, Pilkington JG, Boyd JM, Gulland FMD (2003) Parasites and their impacts. In: *Soay Sheep: Dynamics and Selection in An Island Population* (eds. Clutton-Brock TH, Pemberton JM), pp. 113-165. Cambridge University Press, Cambridge, MA.
Wright S (1951) The genetical structure of populations. *Annals of Eugenics* **15**, 323-354.
Yang ZH (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-1591.
Yang ZH, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution* **15**, 496-503.
Yang ZH, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431-449.

Yang ZH, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution* **22**, 1107-1118.
Yuhki N, O'Brien SJ (1990) DNA variation of the mammalian Major Histocompatibility Complex reflects genomic diversity and population history. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 836-840.
Yuhki N, O'Brien SJ (1997) Nature and origin of polymorphism in feline MHC class II DRA and DRB genes. *Journal of Immunology* **158**, 2822-2833.
Zharkikh A (1994) Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution* **39**, 315-329.
Zuur AF (2010) AED: Data files used in Mixed effects models and extensions in ecology in R (2009). *R. package version 1.0.*
Zuur AF, Ieno EN, Walker N, Saveliev AA, M. SG (2009) *Mixed effects models and extensions in ecology in R*, 1st edn. Springer Science+Business Media, LLC, New York.