# Spatial genetic characterizations of neutral and adaptive variation of Red Junglefowl (*Gallus gallus*) in South Central Vietnam

By

Hoa Nguyen-Phuc

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Animal Sciences)

at the
UNIVERSITY OF WISCONSIN-MADISON
2015

Date of final oral examination: August 3rd 2015

The dissertation is approved by the following members of the Final Oral Committee:

Mark E. Berres, Assistant Professor, Animal Sciences
M. Zachariah (Zach) Perry, Associate Professor, Forest and Wildlife Ecology
Triet Tran, Visiting Scholar, Land Tenure Center
Monica G. Turner, Professor, Zoology
Jun Zhu, Professor, Statistics

# Spatial genetic characterizations of neutral and adaptive variation of Red Junglefowl (*Gallus gallus*) in South Central Vietnam

Hoa Nguyen-Phuc

Under the supervision of Assistant Professor Mark E. Berres

At the University of Wisconsin-Madison

Genetic diversity is inherently a spatial process of stochastic and directional evolutionary forces, as well as the interactions between such forces with the underlying environment. Combined, these factors may operate simultaneously and are challenging to understand, particularly in highly-mobile species or species whose genetic properties have dependence on or influence from human activities. In this dissertation, I employed molecular technologies and spatial analyses to examine genetic diversity and structure in Red Junglefowl (*Gallus gallus*), an important agricultural species. My aim was to understand how spatial processes affect genetic diversity and how landscape patterns may influence population structure at microgeographic scales. I screened 212 wild Red Junglefowl sampled across diverse habitats in South Central Vietnam with two genomic tools. First, amplified fragment length polymorphism (AFLP) surveyed genome-wide neutral variation. Second, a single nucleotide polymorphism (SNP) panel interrogated 84 sites spanning the entire 242 Kb major histocompatibility complex (MHC) B-locus. Analyses of 289 neutral AFLP markers identified a metapopulation structure. Red

Junglefowl in all sampled localities displayed high degree of interspecific-population differentiation (overall $F_{ST}$ = 0.1028) with no evidence of contemporary long-distance genetic exchanges especially across the major barrier Annamite Mountain Range. Fine-scale spatial landscape models detected substantial intraspecific genetic subdivision to distances as low as 5 km. The magnitude of spatial neutral variation of the ground-dwelling pheasants, however, showed no causative relationship between landscape features of landcover and topography. After screening 398 chromosomes, 310 unique MHC haplotypes (77.89%) were identified. Comparison to 17 lines of domestic chickens also screened with the SNP panel indicated that wild Red Junglefowl have extraordinarily high haplotypic diversity. The vast majority of variation in MHC haplotypes (94.51%) occurred within individuals while genetic differentiation between populations was negligible (overall $F_{ST}$ = 0.0083). Likely augmented by recombination, the B-locus also exhibited a few areas of strong linkage suggesting perhaps concerted evolution against a common pathogen. Overall, the results suggest the spatial pattern of MHC is adaptive and under the influence of balancing selection. Neutral markers reflect demographic processes and movements of the Red Junglefowl. I conclude that wild populations of Red Junglefowl in Vietnam represent one of the richest resources of natural genomic variation. Both neutral and adaptive genetic diversity should be equally considered in a spatial research framework for future management of animal genetic diversity, including application to agricultural stock improvement.

*Tặng Mẹ…*

This dissertation is dedicated to my Mother for her support of all my pursuits!

# ACKNOWLEDGEMENTS

Mark Berres has been a tremendous friend and mentor during my time at the University of Wisconsin - Madison. Mark introduced me to the world of Red Junglefowl and stirred me to learn (and trained me) many new subjects: population genetics, molecular techniques, *R* modeling, just to name a few. He always impressed me with his sharp intellect, laughs, and depth of explanations (i.e. 2 chromosomes have 4 DNA strands), and the ability to grasp the essence of any problem in my landscape genetics models within just a few minutes of discussion. In all, the guidance, independence, and resources Mark granted made my PhD project possible.  Thank you, Mark!

I especially thank Triet Tran (Trần Triết) for his tireless support and friendship all the years - Triết has always been a role model and a professional big 'brother' to me. I am truly appreciative to other members of my advising committee: Zach Perry for his insights in population genetics models and comments on early chapter drafts; Monica Turner, with whom even minor questions about spatial landscape ecology ended up great academic advices; and Jun Zhu for training me so much in spatial statistics.

I acknowledge the special contributions in my study: Sean Schoville for providing invaluable inputs and critiques to Chapter 1 and Chapter 2, these manuscripts have greatly improved from his editing; Janet Fulton for her MHC genotyping expertise, sharing domestic chicken data in Chapter 3 as well as her insights for poultry science; Jeb Barzen for being amazingly supportive over the years; and Hoàng Minh Đức for incredible supports in sampling design and paperwork assistance.

Funding for this project was provided by Mark Berres, myself, Hatch Formula Funds (USDA), and Rufford Small Grants for Nature Conservation (United Kingdom).

Collecting DNA from more than 300 Red Junglefowl and Phasianid birds was not a simple task. I am greatly indebt to Nguyễn Chế, Lê Duy, Trần Thanh Tân ‘Tủm’, and Đặng Minh Trí who endured me and spent countless hours (in more than 300 days!) in the field with me. Capture efforts were also assisted and made possible by Mark Berres and Mai Khắc Trung Trực (in most field sites), Hoàng Đức Khoa, Trương Ngọc Đăng, Trương Quang Cường, Lê Quang Minh, Mr. Teng, Lê Jin, Mr. ‘Ông Già’ and his nice family (Bidoup - Núi Bà), Mr. Quẹo (Cát Tiên), Ranger Lâm (Hòn Bà), Ranger Sơn, Ranger Phong, Ranger Thuận, Ranger Ẩn, Hòa, Tùng  (Lò Gò Xa Mát), Ranger Ty (Núi Chúa), and Ranger Kiên, Minh, Kiên, Bi, Nam (Yok Đôn).

Others friends and colleagues nicely helped with sampling, fieldwork, and in the project in general. Mai Khắc Trung Trực and Trương Ngọc Đăng generously allowed me to access the captive colony of Phasianidae in Saigon Zoo and Botanical Garden; Nguyễn Trần Vỹ, Nguyễn Hoài Bão (and Nguyệt), Phùng Bá Thịnh, and Nguyễn Hào Quang were critical for locating sites, permission, and sharing field experience; Lê Khắc Quyết provided critical comments in my manuscripts, and Đặng Thị Phương Thảo let me store the valuable swap samples in her (packed) freezers and greatly helped with reference letters. I truly appreciate the logistical helps from Phùng Lê Cang, Nguyễn Trần Quốc Trung, Nguyễn Phi Ngà, Lưu Hồng Trường, Lê Thị Thùy Dương, Phan Thị Bảo Chi, Huỳnh Quang Thiện, Lê Thị Bích Trâm, and Lê Bạch Mai.

Field sampling was permitted and assisted by reserve staff; to name a few: Lê Văn Hương, Tôn Thất Minh (Bidoup - Núi Bà), Nguyễn Văn Diện, Phạm Hữu Khánh (Cát Tiên), Trần Văn Mùi, Nguyễn Hoàng Hảo (Đồng Nai), Lê Đình Xuân, Tạ Ngọc Dân (Lò Gò Sa Mát), Nguyễn Hữu Thiện, Lê Văn Thừa (Yok Đôn), and in memorial of the late Võ Thanh Liêm (Tà Kóu).

Fellow students in Berres Lab deserve special thanks - thank you to Daniel Marschalek, Matthew Hayes, Becky Kirby, Andrew Cassini, Kimberly Kelly, Jackie Edmunds, and Alissa Milonczyk for providing a great learning and working environment. Matt and Dan helped me so much to become proficient in the lab. Matt patiently reviewed my papers numerous times and greatly improved them (and introduced me to the Berres Lab). Thank you to Kim for sharing the lab work and lab duties.

Thank you to Anh Triết, Chị Linh, and their kids (Trà, Trí, Tường), the Berres family (Mark, Ann, and Max), my friend Ngô Quang Minh and his family who made my last six years in Madison enjoyable.

Finally, a deeply felt thanks to my mother and my sister, and the rest of my family who have always been my primary source of support in my academic endeavors and giving me the freedom and encouragement to take this path in life.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# INTRODUCTION TO THE THESIS

In this dissertation, I employed two molecular genetic technologies and numerous spatially explicit models to examine genetic diversity and population structure in Red Junglefowl (*Gallus gallus*), the ancestor to domesticated chickens. Drawing from a framework of landscape genetics, I specifically focused on understanding how spatial processes of genetic diversity and landscape patterns occur and interact with each other at microgeographic scales. While landscape genetics is an emerging and empirical field, emergent themes offer the best possible information about the effects of landscape patterns on genetic variation in a spatially explicit manner (Manel et al. 2003). Here, I view landscape genetics as an interdisciplinary activity where landscapes and spatial ecology are the stage on which the play of genetic diversity, evolution, as well as the current human-induced conservation issues unfolds. With equal focus on the spatial components of the intervening landscapes and the ecological genetic components of Red Junglefowl, this dissertation aims to answer practical research questions at the interface of ecology, evolution, and resource management. Moreover, because Red Junglefowl has significant agricultural importance, my research will forge new connections to poultry genetics by offering access to a previously unavailable genetic resource.

## *Spatial component of landscape genetics*

As landscapes and environments become more heterogeneous and increase in complexity, connectivity between landscape components becomes more important in influencing gene flow, which in turn affects genetic variation and viability of living organisms. Making realistic spatial interpretations for rapidly changing landscapes, as well as for population genetics, remains an obstacle in landscape genetics (Balkenhol et al. 2009a). Spatial interpretations of landscape connectivity reciprocally evolved from fields of landscape ecology and population genetics. Early models of metapopulation dynamics (Hanski and Gilpin 1997), landscape ecology (Turner et al. 2001), and geographical genetics (Epperson 2003) focused on discrete populations in a matrix composed of suitable patches intermixed among unsuitable habitat patches. Natural populations in these models were well defined in an island patch-based context and their connectivity was expressed as a simple function of geographic distances. Such models relied on mechanistic assumptions and were limited in null-hypothesis testing, such as testing for the presence of barriers to gene flow. More recently, studies have explored continuously distributed individuals or patchily distributed clusters of individuals with low densities between these clusters (Manel et al. 2003). This new spatial interpretation not only seeks to describe landscapes and genetic variation as spatially explicit gradients but also expresses their distributions in individual-based models to provide much higher resolution of the landscape-genetics relationship (Bolliger et al. 2014).

Despite its increased realistic spatial interpretation, landscape genetics is still encumbered by analytical problems (Balkenhol et al. 2009a, Guillot et al. 2009) mostly due to spatial patterns intrinsically present in both physical and biological variables of the study systems (Legendre et al. 2002). Spatial processes such as gene flow, the phenomenon of spatial autocorrelation (Legendre 1993), and environment-induced genetic variation - or spatial dependence (Wagner and Fortin 2005) - are inherent in ecological and evolutionary processes. These spatial phenomena, however, violate the assumption of independent and identical distribution (*i.i.d.*), hence perturb significant tests and inflate Type I errors (Dormann et al. 2007, Guillot et al. 2009). Multivariate analyses and distance-based geostatistical methods are considered a more appropriate way for detecting non-linear and non-independent data in landscape genetics (Wagner et al. 2005, Legendre and Fortin 2010). Essentially, researchers need to develop general and hypothesis-driven analytical framework to test for specific landscape-genetic mechanisms in their study system of interdisciplinary landscape genetics (Balkenhol et al. 2009b, Guillot et al. 2009).

### *Ecological genetics with applied focuses*

The spatial framework in landscape genetics could hold great promise in research and resource management if relevant ecological focuses, good research systems, and applied solutions are clearly delineated. Many research programs and disciplines, such as the fields of molecular ecology and conservation genetics (Diniz et al. 2008), have furthered our understanding of the spatial and geographical components of genetic diversity. As a direct consequence, biodiversity conservation is including more genetically-themed information in decision making processes. Application of ecological genetics aims to preserve species in rapidly changing environments. Advantageously, the field of landscape genetics is advancing this goal in two distinctive ways. First, instead of limiting analyses of genetic data only to summary descriptions of genetic diversity, landscape genetics advocates the use of neutral genetic markers within a spatially explicit context to determine patterns, if any, in local adaptation of populations (Schoville et al. 2012). Typically, neutral genetic population processes include migration, dispersal, and gene flow (Frankham et al. 2004). By examining the spatial distributions of alleles at different spatial hierarchies using a wide variety of organisms, landscape genetics has a better likelihood to determine the causal relationship between environmental variables and adaptive genetic variability. The second and more recent approach, is concerned with the processes of adaptive genetic variation which describes the biological function and local adaptations of entire genomic regions (Bolliger et al. 2014).

Both approaches have their own advantages. Neutral genetic markers, such as AFLPs and microsatellites, are relatively inexpensive and could be obtained in large volume to reveal ample levels of genetic variation at most loci. Alternatively, adaptive variation studies in the field of quantitative genetics could find significant heritability for most traits. However, the integration between the two approaches in landscape genetics, or in the biological sciences in general, was considered by Phillips (2005) as ‘caught between molecular knowledge in the absence of adaptive context and ecological context in the absence of molecular details’ (p. 16). Future advances in the field of landscape genetics will require reconciling adaptive and neutral variation assessments to determine how genes under selection disperse across landscape, and how gene flow counterbalances local adaptation (Manel and Holderegger 2013). An additional area of improvement may involve empirical and applied foci wherein genetic variation for ecologically or agricultural important traits (or both) is described in natural populations to the determination of the genetic basis of phenotypic and fitness differences on the landscapes (Schoville et al. 2012) and between kin species (Bernatchez and Landry 2003).

### *Dissertation organization and research questions*

This dissertation has three research chapters. The chapters are unified by two themes. First, they share a common focal species and sampling region. I sampled populations of wild Red Junglefowl (*Gallus gallus*) distributed in geographically diverse habitats in South Central Vietnam. Still considered by IUCN a non-threatened species, wild populations of Red Junglefowl have recently experienced considerable anthropogenic influences which may affect their long-term conservation (Peterson and Brisbin 1998, Fuller and Garson 2000). Red Junglefowl are considered the direct ancestor of domestic chickens (Fumihito et al. 1994). The domestication of wild fowl occurred several thousands of years ago (ca. 3000 – 6000 ybp), presumably in South and Southeast Asia and spreading globally through human dispersal and cultural development (Storey et al. 2012). Today, the global poultry industry is experiencing massive reductions of genetic diversity in commercial chicken breeds that subsequently may bring poultry (and other livestock) to a “selection wall” for growth and reproductive traits (Muir et al. 2008). Especially important are newly emerging diseases such as West Nile Virus and Avian Influenza. Not only do these diseases affect poultry and wild birds, but also exhibit considerable potential for zoonotic diseases (Berlin et al. 2008, Downing et al. 2009). Wild populations of Red Junglefowl are likely to provide the richest resource of genomic variation to poultry genetic management and could be a key genetic reservoir for maintaining a healthy poultry industry as well as minimizing the zoonotic potential of industrial-scale agriculture.

Second, I take advantage of two molecular techniques, neutral amplified fragment-length polymorphisms (AFLPs) markers and adaptive major histocompatibility complex (MHC), to characterize the ecological genetics of wild Red Junglefowl. AFLPs (Vos et al. 1995) are a marker system based on simple nucleotide substitution scoring based on allele-presence and absence. Although AFLPs cannot distinguish heterozygote genotypes, they have a proven history in population studies as it balances the necessity for more complete genomic coverage with a level of resolution capable of identifying individuals (Bonin et al. 2007). In contrast, MHC is a locus under intense selection and exhibits some of the most polymorphic genes known (Hess and Edwards 2002). Hundreds of alleles and single nucleotide polymorphisms (SNPs) have been described at certain MHC loci in laboratory strains of chickens (Fulton et al. 2006). Currently, not much is known about MHC variation in non-commercial chickens (Izadi et al. 2011) and even less in wild populations of Red Junglefowl. In the wild, selective pressures from a variety of environmental factors could generate extraordinary degrees of polymorphism that exceed either those detected in neutral loci or MHC in commercial breeds. My overall goal is to determine levels of genetic diversity and structure and to evaluate if landscape has any affect on its distribution.

In Chapter 1, the primary research question is: what are major spatial processes influencing neutral genetic diversity in Red Junglefowl across the South Central Vietnam landscape? I hypothesize that the ground-dwelling Red Junglefowl exhibit substantial inter-population variation across the landscape due to habitat fragmentation and limited fine-scale gene flow, which is mediated by strong natal philopatry. Coarse-scale analyses evaluate the magnitude of population-level differentiation while local, fine-scale analyses reflect how evolutionary forces such as spatial process and demography may have influenced the local population structure of Red Junglefowl. Typically, ecological genetics studies focus on population-based sampling and analyzing. In addition, spatial processes such as isolation-by-distance and the effects of sampling scales have not been thoroughly addressed in geographical-level genetics models. To address these issues, I employed and evaluated complementary methods of ordination, correlograms, and Bayesian clustering in determining patterns of neutral AFLP variation in the Red Junglefowl at different spatial scales. I emphasized the utility of intensive computation in clustering method in order to produce high-resolution genetic structure for wild populations. Alternatively, deterministic individual-based models such as gradient analyses can provide intuitive information about genetic variation from reasonable size dataset with strong population structures.

In Chapter 2, using the neutral diversity and population structure conceived in Chapter 1, I asked whether geographic distances or patterns of major landscape features were deciding factors responsible for the extent of such spatial variation. Currently, hypothesis testing in the individual-based framework in landscape genetics is mostly based on a null model known as the Island Model. This non-spatial framework is, therefore, confounded with spatial and non-independent data such as auto-correlated genotypes and environment-induced genetic variation. Distinguishing different spatial phenomena in both responding genetic structure and the explanatory landscapes will correctly determine the causative relationship between them. I focused on landcover and topography as two major landscape features in driving genetic variation of the widespread and locally abundant Red Junglefowl. Although the two landscape features vary within each study site, I hypothesize that Red Junglefowl are sufficiently flexible in their habitat requirements relative to the local landscape variation. Thus, I expect that elevation and land cover do not influence the birds' genetic structures. I developed a spatially explicit analytical framework wherein integrative methods of multivariate analysis, geostatistics, and GIS techniques are applicable for the research objective. Overall, this chapter and its research methods contribute to the developing discipline of landscape genetics, specifically the utility of distance-based and individual-based methods correlating genetic variation with landscape features.

Chapter 3 is an adaptive genetic variation study designed to characterize diversity at the MHC region, an essential component of immune system responses and functions. I sought to answer if the functional MHC genes in wild Red Junglefowl are, indeed, retained at high diversity and variation compared to neutrality variation described in Chapter 1 and to intensively selected chicken lines. Differences between the adaptive and neutral diversity will provide insights into the evolution history of Red Junglefowl. Balancing selection is thought to enhance MHC variation and the effects of genetic drift and gene flow has shaped the distribution of neutral variation in populations. Also, understanding how natural and artificial selection affect MHC in the wild and domestic birds (poultry) will create new avenues to understand linkage between immune system genetic variation and disease resistance. Considering the extent of the geographical sampling and the range of different habitats, I hypothesize that wild Red Junglefowl possess great diversity in their MHC genes.

To test this hypothesis, I used a novel high-density SNP detection system to genotype and assess nucleotide diversity at 84 sites distributed across approximately 240 Kp of the MHC B-locus in wild Red Junglefowl. The data set was then analyzed by extensive Bayesian inferences to reconstruct chromosomal MHC haplotypes, and from there, nucleotide diversity and population structure were estimated. The combination of high-throughput genotyping and population genetics models in the chapter represents the first comprehensive MHC haplotypic assessment in any wild bird species.

The three research chapters in the dissertation were formatted as independent manuscripts for publication in scientific journals. I made some minor formatting adjustments for the sake of consistency among chapters. I was the primary contributor for each chapter conducting data collection, analytical development, analysis, and writing. My PhD advisor and committee chair, Dr. Mark Berres, is a coauthor on all chapters, contributing to research design, idea refinement, and manuscript revisions. We collaborated with Dr. Janet Fulton to generate the MHC data in Red Junglefowl in Chapter 3.

# REFERENCES

Balkenhol, N., F. Gugerli, S. A. Cushman, L. P. Waits, A. Coulon, J. W. Arntzen, R. Holderegger, and H. H. Wagner. 2009a. Identifying future research needs in landscape genetics: where to from here? Landscape Ecology **24**(4):455-463.

Balkenhol, N., L. P. Waits, and R. J. Dezzani. 2009b. Statistical approaches in landscape genetics: an evaluation of methods for linking landscape and genetic data. Ecography **32**(5):818-830.

Berlin, S., L. J. Qu, X. Y. Li, N. Yang, and H. Ellegren. 2008. Positive diversifying selection in avian *Mx* genes. Immunogenetics **60**(11):689-697.

Bernatchez, L., and C. Landry. 2003. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? Journal of Evolutionary Biology **16**(3):363-377.

Bolliger, J., T. Lander, and N. Balkenhol. 2014. Landscape genetics since 2003: status, challenges and future directions. Landscape Ecology **29**(3):361-366.

Bonin, A., D. Ehrich, and S. Manel. 2007. Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. Molecular Ecology **16**(18):3737-3758.

Diniz, J. A. F., M. P. D. C. Telles, S. L. Bonatto, E. Eizirik, T. R. O. de Freitas, P. de Marco, F. R. Santos, A. Sole-Cava, and T. N. Soares. 2008. Mapping the evolutionary twilight zone: molecular markers, populations and geography. Journal of Biogeography **35**(5):753-763.

Dormann, C. F., J. M. McPherson, M. B. Araujo, et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography **30**(5):609-628.

Downing, T., D. J. Lynn, S. Connell, et al. 2009. Contrasting evolution of diversity at two disease-associated chicken genes. Immunogenetics **61**(4):303-314.

Epperson, B. K. 2003. Geographical genetics. Princeton University Press, Princeton, New Jersey.

Frankham, R., J. D. Ballou, and D. A. Briscoe. 2004. A primer of conservation genetics. Cambridge University Press, Cambridge, UK; New York.

Fuller, R., and P. Garson, editors. 2000. Pheasants. Status Survey and Conservation Action Plan 2000-2004. WPA/BirdLie/SSC Pheasant Specialist Group. IUCN, Gland. Switzerland and Cambrigde, UK and the World Pheasant Association, Reading, UK.

Fulton, J. E., H. R. Juul-Madsen, C. M. Ashwell, A. M. McCarron, J. A. Arthur, N. P. O'Sullivan, and R. L. Taylor. 2006. Molecular genotype identification of the *Gallus gallus* Major Histocompatibility Complex. Immunogenetics **58**(5-6):407-421.

Fumihito, A., T. Miyake, S. I. Sumi, M. Takada, S. Ohno, and N. Kondo. 1994. One subspecies of the Red Junglefowl *Gallus gallus gallus* suffices as the matriarchic ancestor of all domestic breeds. Proceedings of the National Academy of Sciences of the United States of America **91**(26):12505-12509.

Guillot, G., R. Leblois, A. Coulon, and A. C. Frantz. 2009. Statistical methods in spatial genetics. Molecular Ecology **18**(23):4734-4756.

Hanski, I., and M. E. Gilpin. 1997. Metapopulation biology: ecology, genetics, and evolution. Academic Press, San Diego, CA.

Hess, C. M., and S. V. Edwards. 2002. The evolution of the major histocompatibility complex in birds. Bioscience **52**(5):423-431.

Izadi, F., C. Ritland, and K. M. Cheng. 2011. Genetic diversity of the major histocompatibility complex region in commercial and noncommercial chicken flocks using the LEI0258 microsatellite marker. Poultry Science **90**(12):2711-2717.

Legendre, P. 1993. Spatial autocorrelation - Trouble or new paradigm. Ecology **74**(6):1659-1673.

Legendre, P., M. R. T. Dale, M. J. Fortin, J. Gurevitch, M. Hohn, and D. Myers. 2002. The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography **25**(5):601-615.

Legendre, P., and M.-J. Fortin. 2010. Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. Molecular Ecology Resources **10**(5):831-844.

Manel, S., and R. Holderegger. 2013. Ten years of landscape genetics. Trends in Ecology & Evolution **28**(10):614-621.

Manel, S., M. K. Schwartz, G. Luikart, and P. Taberlet. 2003. Landscape genetics: combining landscape ecology and population genetics. Trends in Ecology & Evolution **18**(4):189-197.

Muir, W. M., G. K. S. Wong, Y. Zhang, et al. 2008. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. Proceedings of the National Academy of Sciences of the United States of America **105**(45):17312-17317.

Peterson, A. T., and I. L. Brisbin. 1998. Genetic endangerment of wild Red Junglefowl *Gallus gallus*. Bird Conservation International **8**(4):387-394.

Phillips, P. C. 2005. Testing hypotheses regarding the genetics of adaptation. Genetica **123**(1-2):15-24.

Schoville, S. D., A. Bonin, O. Francois, S. Lobreaux, C. Melodelima, and S. Manel. 2012. Adaptive Genetic Variation on the Landscape: Methods and Cases. Annual Review of Ecology, Evolution, and Systematics **43**:23-43.

Storey, A. A., J. S. Athens, D. Bryant, et al. 2012. Investigating the global dispersal of chickens in rrehistory using ancient mitochondrial DNA signatures. PLOS One **7**(7).

Turner, M., R. Gardner, and R. O'Neill. 2001. Landscape Ecology in Theory and Practice: Pattern and Process. Springer-Verlag, New York.

Vos, P., R. Hogers, M. Bleeker, et al. 1995. AFLP - a New Technique for DNA-Fingerprinting. Nucleic Acids Research **23**(21):4407-4414.

Wagner, H. H., and M. J. Fortin. 2005. Spatial analysis of landscapes: Concepts and statistics. Ecology **86**(8):1975-1987.

Wagner, H. H., R. Holderegger, S. Werth, F. Gugerli, S. E. Hoebee, and C. Scheidegger. 2005. Variogram analysis of the spatial genetic structure of continuous populations using multilocus microsatellite data. Genetics **169**(3):1739-1752.

# CHAPTER 1

## Spatial genetic structure of the wild Red Junglefowl (*Gallus gallus*) in their core distribution range in South Central Vietnam

# ABSTRACT

Human activities have caused significant evolutionary change in species of agricultural importance through artificial selection and domestication. Threats to wild ancestors of agricultural animals can range from global eradication of the wild progenitor, to a pronounced reduction of genetic diversity, or to genetic contamination from domesticated stocks. Large-scale evaluations of genetic diversity are of central importance to the conservation of genetic resources and agriculture science, yet severely lacking for the wild ancestral populations of domestic stocks in their remaining habitats. We perform the first large-scale spatially explicit study to characterize the genetic diversity of wild populations of Red Junglefowl (*Gallus gallus)* from geographically and ecologically diverse tropical habitats in South Central Vietnam. Results of Bayesian clustering and population genetics models show a strong signature of population structure (overall $F_{ST} = 0.1028$) of Red Junglefowl in a geographical context wherein distinct population clusters comprise a metapopulation. Spatial analyses also suggest, in contrast to the birds' large geographic distribution and their ubiquitous relationship with domestic chickens, a high degree of fine-scale genetic subdivision also exist at distances as low as 5 km. Our results suggest that the current natural populations of Red Junglefowl in Vietnam are small and isolated, and are therefore susceptible to endogenous threats to their genetic diversity and perhaps may also be at risk to genetic introgression from native domestic chickens.

**Keywords**: AFLP, Bayesian cluster analysis, gene flow, metapopulation, Red Junglefowl, spatial genetic variation.

# INTRODUCTION

Chickens are among the most important domesticated animals, making tremendous contributions to human society in both economic and nutritional terms (Muir et al. 2008), but also as model systems for scientific research (Delany 2006). The domestication of wild fowl occurred several thousands of years ago, and spread globally following human demographic and cultural development (Storey et al. 2012). This process surely played a significant role in reshaping the genetic composition and phenotypic traits of wild fowl into modern chicken breeds (FAO. 2007, Groeneveld et al. 2010). Today, the global poultry industry represents a multi-billion dollar business with more than 40 billion chickens produced annually (Muir et al. 2008). The industry is predominantly governed by trans-national companies who create and market a few lines of intensively-selected chicken lines world-wide (FAO. 2007). This has resulted in massive genetic diversity reductions in commercial chickens that subsequently may bring poultry livestock to a perceived “selection wall” for growth and reproductive traits (Muir et al. 2008), as well as increase their susceptibility to zoonotic diseases (Berlin et al. 2008, Downing et al. 2009).

In addition to the lack of genetic diversity in commercially-bred chickens, indigenous or heritage chicken breeds (i.e., noncommercial lines historically created through selective breeding and locally maintained) are at a far greater risk of declining genetic diversity compared to other domesticated mammal and avian species (Fulton and Delany 2003, FAO. 2007). Maintaining and improving genetic diversity in chickens (poultry in general) are critical for long-term sustainable agriculture wherein the common approach has been only to maintain the extent of genetic variation within and among breeds, strains, and lines (Granevitze et al. 2007, Lenstra et al. 2012). Indigenous breeds of chickens have strong potential to recover genetic variation in commercial chicken lines (Muir et al. 2008). Breeds native to Southeast Asia are of particular interest since they share the most likely origin of domestication (Berthouly et al. 2010, Ngo et al. 2010). The indigenous chicken lines are considerably divergent and contain distinct genotypes compared to the inbred broilers and layers (Mekchay et al. 2014).

However, it is imperative to recognize that Red Junglefowl (*Gallus gallus*), the wild progenitors of all domesticated chickens, still exist in their native habitats (Fumihito et al. 1994, Johnsgard 1999). Red Junglefowl are medium size birds (~ 500 - 1 000 g) and mainly ground-dwelling. Morphological and demographical characters of the species such as short rounded wings, polygynous breeding, and female promiscuity indicate the birds have limited migration range and natal dispersal is the primary mode of gene flow (Johnsgard 1999). Naturally occurring Red Junglefowl are likely to provide the richest resource of genomic variation since they were not subjected to a domestication process; extant populations of Red Junglefowl likely still maintain the ancestral alleles of which very little remain in our highly derived commercial lines. Therefore, the potential for Red Junglefowl to augment poultry genetic management is considerable and should be an invaluable genetic reservoir for maintaining a healthy poultry industry. In addition, calling attention to the genetic value of wild Red Junglefowl populations by initiating early investigations into the population genetic structure and gene flow among natural Red Junglefowl populations will inform landscape management efforts and help to conserve these wild populations.

Within the Indo-Burma biodiversity hotspot, South Central Vietnam is the core range of Red Junglefowl distribution (Johnsgard 1999). Red Junglefowl are distributed in South Central Vietnam from lowlands to about 600 m in elevation, across a highly variable environmental gradient and heterogeneous landscape. The biogeography of our focal region is characterized by the Annamite Mountain Range (Trường Sơn) featuring high plateaus of approximately 200 km in length with elevations ranging up to 2 400 m. Tropical lowland forests occur in the eastern coastal areas and become more abundant in the south, while dry, deciduous forests are found mainly in the northwest. Red Junglefowl presumingly had a continuous distribution over their preferred, mostly secondary, habitats with the Annamite acting as an extended landscape barriers to the species' dispersal. The current landscape in South Central Vietnam has small and severely fragmented natural habitats that potentially influence the spatial genetic diversity of the Red Junglefowl and call into question the likelihood for continued survival.

In this study, we characterized the genetic population structure of wild Red Junglefowl in the South Central region of Vietnam. Few genetic studies of the species have been conducted (Storey et al. 2012) and genetic purity of Red Junglefowl in those studies is questionable (Brisbin et al. 2002). Abundant evidence shows that genetic exchange does occur between feral or free-ranging domestic chickens and wild Red Junglefowl (Peterson and Brisbin 1998, Berthouly et al. 2010, Mekchay et al. 2014). Our first aim is to study neutral genetic variation in wild populations of Red Junglefowl in geographically diverse habitats to determine whether landscape barriers and differentiation by means of isolation-by-distance along ecological gradients characterize the coarse- and fine-scale genetic diversity. Coarse-scale analyses will evaluate the magnitude of population-level differentiation while local, fine-scale analyses will reflect how evolutionary forces such as spatial process and demography may have influenced the population structure of Red Junglefowl. We hypothesize that the Red Junglefowl exhibit substantial interspecific population variation across the landscape due to habitat fragmentation and limited gene flow mediated by strong natal philopatry. We demonstrate the utility of intensive Bayesian computational clustering, with different clustering resolutions, and distance-based methods to investigate the relative influences of spatial processes and neutral genetic variation on population structure of this remarkable important species of pheasant.

## MATERIALS AND METHODS

### *Field sampling*

We sampled Red Junglefowl in seven protected areas in South Central Vietnam during three dry seasons in 2012, 2013, and 2014. The tropical dry season in the Annamite Mountain Range occurs from January to May and overlaps with the breeding season of Red Junglefowl. During this time, mating and territorial defense facilitate the location and sampling of birds. Outside of this time period, Red Junglefowl become very secretive and extraordinarily difficult to locate and nearly impossible to capture. Sampling sites were chosen in protected areas and included Bi Đúp Núi Bà National Park (BDP), Cát Tiên National Park and Đồng Nai Nature Reserve (hereafter CTN as the two sites are connected), Hòn Bà Nature Reserve (HBA), Lò Gò Sa Mát National Park (LGO), Núi Chúa National Park (NCA), Tà Kóu Nature Reserve (TKU), and Yok Đôn National Park (YDN) (Figure

1.1). Selection of sites was based on the presence of suitable junglefowl habitat and their relatively symmetrical distances from the putative barrier of the Annamite Mountain Range. The average area of each field sites was approximately 50 000 hectares. They mostly feature natural habitats of lowland tropical rainforest (≤ 600 m in elevation) and they are separated apart from one another (~ 180 km) by residential and non-natural habitats. Although Red Junglefowl appear able to tolerate significant habitat degradation, extensive farmland conversion and pervasive hunting in South Central Vietnam likely fractured a once continuous Red Junglefowl population into small, isolated populations. The sampling regions chosen reflect a range of isolation from human influences.

We live-captured Red Junglefowl with non-lethal walking snares (Bub 1991) adapted from local trapping customs. Decoy domestic roosters trained to produce territorial vocalizations ("baiting cocks") augmented our trapping efforts. A total of 212 birds were sampled from the seven field sites (Figure 1.1, Table 1.1). We determined age and sex of captured birds by plumage characteristics with 172 roosters, 23 hens, and 17 juvenile chicks (< 3 months old). Within sites, the average distance between capture localities was 1.12 km (max 26 km, min 0 m of birds in the same flock) depending on capture opportunities. Capture rates were estimated to be 0.92 bird per work day as the junglefowl were still highly elusive (despite being strongly territorial) in their mating season. We initially attempted to capture equal numbers of male and female junglefowl to understand sex-biased dispersal and home range. However, the use of snares alone had very low capture efficiency and required extra sampling efforts in checking and maintaining the snare lines. With the assistance of decoy roosters, we had greater capture efficiency but mainly lured and captured the territorial male junglefowl. Generally in all field sites, we mostly observed and sampled junglefowl in mature secondary-growth habitats. Bamboo forests, which have clear understory floor and often sprout after fire, provide ideal habitat and an excellent food source for the ground-dwelling junglefowl.

We marked the sampled birds with numbered aluminum bands. At all sampling localities, we recorded geographical coordinates and characterized major vegetation structure. Coordinates for each trapped Red Junglefowl were taken with a GPS receiver. For each bird, a 20 - 200 µL blood sample was obtained from the brachial vein and stored in a lysis buffer (0.1 M Tris-HCl pH 8.0, 0.01 M EDTA, 4% SDS) (Longmire et al. 2000) until DNA extraction. We used a higher SDS concentration than the recommended 2% to better lyse and preserve blood cells in high temperature field conditions.

### *Genotyping of AFLP profiles*

We generated AFLP profiles (Vos et al. 1995) for our samples with modified protocols (Marschalek and Berres 2014). Genomic DNA was extracted from blood using the Promega Wizard DNA Isolation kit (Promega Corp., Madison, WI, USA) and was assessed visually with 1% agarose gel electrophoresis for non-degraded, high molecular weight DNA. We tried various combinations of restriction enzyme pairs, pre-selective primer pairs (*x*3) and selective primer pairs (*x*24), and chose the ones that maximized the number of fragments that met a criterion of full resolution profiles (Berres 2003). We digested 200 ng of high quality DNA to completion with 20U restriction enzymes *Eco*RI (5’-G|AATTC-3’) and 20U *Ase*I (5’-AT|TAAT-3’) at 37 °C overnight followed by a 20 min heat treatment at 65 °C.

The digested fragments were ligated at 16$^o$C overnight with 400U T4 DNA ligase to double-stranded oligonucleotide adapters with overhangs complementary to the digested ends. Ligated samples were diluted 1:4 with 10mM Tris-HCl (pH 8.0) to produce DNA templates for polymerase chain reaction (PCR) amplification.

Pre-selective PCR was performed in 50 µL volumes of 10 µL diluted ligation DNA with PCR mixture (1*x* GoTaq Flexi Buffer, 1.5 mM MgCl2, 0.05 mM dNTP, 2% deionized formamide, 1.25U Taq DNA Pol I) and 15 pmoles each of primer pairs *Eco*RI+C/*Ase*I+G or *Eco*RI+G/AseI+G. Thermocycling conditions consisted of one cycle of 72 °C for 2 min, an initial denature at 94 °C for 1 min followed by 25 cycles each of 94 °C for 50 s, 56 °C anneal for 1 min, and 72 °C extension for 2 min. Pre-selective amplification products were diluted 1:19 with 10 mM Tris-HCl (pH 8.0) but those with lower amplification (determined by visual inspection on an ethidium bromide stained 1% agarose gel) were diluted 1:9.

Selective PCR amplification was performed in 25 µL volumes containing 5 µL diluted pre-selective amplification product with PCR mixture (1*x* GoTaq Flexi Buffer, 2 mM MgCl2, 0.2 mM dNTP, 2% deionized formamide, 0.625U Taq DNA Pol I) and 5 pmoles HPLC-purified primer *Eco*RI+CAT labeled with 6-carboxyfluorescein (6-FAM), and 25 pmoles *Ase*I+GA or with 5 pmoles HPLC-purified primer *Eco*RI+GG labeled with 6-carboxyfluorescein (6-FAM), and 25 pmoles *Ase*I+GC. Thermocycling conditions consisted of an initial 94 °C denature for 1 min followed by 10 cycles of a 1 min annealing touchdown (1 °C decrease each cycle) from 65 °C to 56 °C each with a 72 °C extension for 2 min. The selective amplification was completed with 18 cycles of 95 °C for 50 s, 56 °C for 1 min, and 72 °C for 2 min. Selectively amplified PCR products were purified over Sephadex G75 and stored at -80 °C. One µL of purified product was combined with 13.5 µL deionized formamide and 0.5 µL Geneflo 625 (mobility standard, CHIMERx Molecular Biology Products) for electrophoresis on an ABI 3730xl DNA Analyzer (Biotechnology Center, UW-Madison, WI).

AFLP profiles were categorized (binned) by a partially automated scoring process. First, we used *RawGeno* package (Arrigo et al. 2009) in the open-source *R* 3.1.2 environment (R Development Core Team 2011) to create bins of homologous amplicons based on their electrophoretic mobility (converted in units of base pairs). The maximum bin width was chose to be 1.2 base pairs and unconstrained for narrower bins). This procedure generated a binary matrix of presence/absence (1/0) scores for each amplicon in the AFLP fingerprint. Trace files of the individual samples were then visually inspected in *DAx* 8.0 (van Mierlo Inc., The Netherlands). Replicable PCR generation of AFLP fingerprints usually yields amplicons with a relative fluorescence intensity (RFU) between 1 000 and 15 000. Peaks with RFU of less than 50 were eliminated from the analysis. Comparing bins created by *RawGeno* and *DAx*, each of which uses a differing binning algorithm, allowed us to evaluate errors in amplicon assignment. In cases where bins did not match, visual inspection and manual bin reconstruction was performed.

### *General data analysis*

The AFLP data matrix of 212 wild-caught Red Junglefowl was analyzed with two approaches: an individual approach based on band scores and a population approach using allele frequencies. The approaches are intrinsically different in their complexities, assumptions, and computational requirements. At individual level, we calculated coefficients of similarity (Jaccard 1908) for Red Junglefowl AFLP profiles using the *R*-package *ecodist* (Goslee and Urban 2007). This software generated a square matrix of genetic (dis)similarities. We then employed three complementary steps of spatial and non-spatial unconstrained ordination and created correlograms to determine patterns of genetic differentiation at different spatial scales, i.e. clusters, clines, and isolation-by-distance (IBD) (Wright 1943). At the population level, we combined Bayesian inferences and population genetic summary statistics to estimate allele frequencies, identify populations, and calculate pairwise $F_{ST}$-values. The population-based methods convert the observed bi-allelic AFLP data into frequencies and assume linkage and Hardy-Weinberg equilibrium (HWE), implicit assumptions that are often violated in naturally occurring populations. Performance of individual approach and population approach, particularly how spatial scales are incorporated and described in the models, is discussed in detail in this study.

### *Individual-based analyses*

We employed the centered and scaled Principal Component Analysis PCA (Jongman et al. 1995) in the *R*-package *ade4* (Dray and Dufour 2007) with the *R*-package *pca3d* display (Weiner 2013) for clustering ordination, in order to identify groupings based on overall genetic variation in Red Junglefowl. We then tested the PCA's principal components using Moran's *I* (Moran 1950) with Monte Carlo randomization method (1 000 permutations) in the *R*-package *spdep* (Bivand et al. 2008) for spatial autocorrelation. We used spatial PCA (sPCA), an exploratory procedure in the *R*-package *adegenet* (Jombart et al. 2008) to investigate the correlation between geographic distance and genetic variation. The procedure may be useful to identify fine-scale spatial patterns of genetic variability when mapping the spatial components of the principal components' scores. As suggested by Jombart et al. (2008), we modified the default symmetrical Gabriel connectivity graph setting in sPCA to provide spatially explicit locations and estimate spatial ranges. The sPCA's first principal component's scores were then regressed to their localities by a linear least-squared method and the residuals were visually interpolated (by inverse distance weighting) across the sampling region using open-sourced QGIS 2.4 Chugiak (QGIS Development Team 2014).

In order to determine the presence and, more importantly, spatial range of relatedness, we constructed correlograms among all Red Junglefowl samples. Correlograms (Sokal 1986) differ from ordinations as they are not global statistic procedure *per se*. Moreover, visual evaluation better describes the spatial range of influence between samples (IBD). An IBD pattern would be expected if dispersal is restricted only by distance, with increasing genetic differentiation occurring over greater geographical distances. If a landscape feature restricts dispersal, individuals captured in close proximity may be very different genetically and an IBD pattern may not occur. We employed the *R*-package *ncf* (Bjornstad 2013) to construct Mantel multivariate (cross) correlogram with 1 000 permutations (Mantel 1967, Bjornstad et al. 1999) and equal discrete distance classes in 10 km.

### *Population-based analyses*

We first examined patterns of population genetic variation using summary statistics. First, a Bayesian method with uniform prior distributions between samples (Zhivotovsky 1999) was performed in *AFLPsurv* (Vekemans et al. 2002). For each observed population, we calculated the expected heterozygosity ($H_E$), proportion of polymorphic sites, and pairwise genetic differentiation ($F_{ST}$) tested with a randomization procedure consisting of 1 000 permutations. We used the rarefaction function in the *R*-package *vegan* (Oksanen et al. 2013) to create thresholds for expected private allelic richness by considering equally-sized random subsamples from each population (Kalinowski 2004) and then recorded the numbers of observed private alleles that were greater than the threshold of 5% of the total subsample size.

To test if distinct populations existed in our sample, we used the Bayesian clustering method in the *R*-package *Geneland* 4.0.4 (Guillot et al. 2005) with admixture and correlated allele frequencies for dominant AFLP markers. Similar to the band-based ordination methods, the clustering models were both spatially explicit (using sampling geographic coordinates and genotypes) and non-spatial (genotypes only). Both models were subjected to 2 000 000 Markov chain Monte Carlo (MCMC) iterations, thinning by a factor of 100 times. A posterior burn-in of 2 000 iterations was allowed (i.e. 200 000 burn-in out of the total 2 000 0000 MCMC iterations). We ran each model 1 000 times on UW-Madison’s HTCondor high-throughput computer system and computed the average posterior distributions of modal $\Delta K$ populations.

Apart from the stochastic nature of the MCMC sampling and the high dimensionality of the dataset, biological phenomena will influence how individuals are apportioned into specific population. First is the relative relationship between our sampling scale and the spatial range of the Red Junglefowl, which we assumed was regulated by home-range demography and possibly the intrinsic quality of the sampling habitats. Increased habitat connectivity may increase dispersal, thereby preventing genetic differentiation, and affect the potential resolution of inferred genetic clusters. Second, since our sample design was opportunistic within a site, some Red Junglefowl may be sampled from cryptic linages or populations that are quite different to the well-represented junglefowl in our sampling pool. If individuals meeting these criteria are sampled at low density, assignment into specific populations will be more difficult to achieve, particularly at the global scale of our sampling design.

We addressed the global sampling issue by running additional clustering inferences for each site individually to evaluate differentiation in population membership. The additional *Geneland* models, as well as sPCA ordinations and correlogram correlations, for the local sampling sites (local models) were parameterized similarly to the models with all samples included (global models). The local correlograms had finer discrete distance class at 1 km to reflect maximum distance of about 20 km within a local sampling site. With clustering models at both global and local scales, aside from the important mean log posterior density distribution of $\Delta K$ reported by *Geneland* and our estimation of modal $K$ values out of 1 000 replicates, we also constructed dendrograms of average linkage of unweighted pair group method with arithmetic mean (UPGMA) among our samples (Figure

1.7). The UPGMA topology does not depict phylogenetic relationships or structure *per se*, but rather the mean posterior probability of common cluster memberships. A probability of 0.00 indicates that two Red Junglefowl were always placed in the same population cluster in 1 000 runs, whereas a probability of 1.0 indicated that the two Red Junglefowl were never grouped together in any of the replicates. Junglefowl that consistently changed their memberships in both global and local clustering models may represent sampled from rare, under-represented, or cryptic populations.

## RESULTS

The two selective primer pairs yielded 389 replicable polymorphic AFLP loci ranging from 50 to 616 base pairs. The monomorphic sites were 42, or ~ 90% of the total profiles was polymorphic. Plotting pairwise genetic dissimilarities of the Red Junglefowl versus their localities (Figure 1.2A) and global Mantel correlogram (Figure 1.2B) both indicated an absence of spatial autocorrelation ($r = 0.11$, $p < 0.001$). This demonstrated that a confounding IBD effect generally does not exist and our spatial analyses are unbiased by our sampling strategy.

### *Patterns of genetic differentiation*

The first three PCA principal components (PC) identified clear genetic differentiation among the seven sampling sites (Figure 1.3). The first PC (explaining 10.43% of total variance, Moran's $I = 0.49$, $p < 0.001$) separated not only geographical clusters of the Red Junglefowl but also reveal considerable genetic structure within the two largest and least disturbed sites (lowland tropical forest CTN and highland dry forest YDN). The second PC (5.57% of variance, $I$ = 0.42, $p$ < 0.001) and third PC (4.35% of variance, $I$ = 0.42, $p$< 0.001) reinforced partitions observed in CTN and YDN, again both had significant within-site genetic differentiation. The HBA site, which is in the northeast foothills of the Annamite Mountain Range, had a central position in the PCA plot and showed overlaps with the clusters of CTN and YDN which are in the further west of the Annamite Mountain Range. Although this observation may suggest low levels of genetic exchanges across this range, but it is also possible that it results from a lack of resolution in the PCA analysis.

The global sPCA in all 212 Red Junglefowl samples resulted in high positive eigenvalues and uniformly low negative eigenvalues (Figure 1.4A - insert). This, together with the overall well-defined sPCA's regressed gradient variances (Figure 1.4A), illustrated monotonic clines of genetic similarities along the east and the west sides of the Annamite Mountain Range landscape. Specifically, we found evidence of local structure within individual sPCA models in the four major sampling sites (CTN, LGO, HBA, and YDN) (Figure 1.4B, Figure 1.5 a). Both the eigenvalues and residual values of the local sPCA scores indicated that Red Junglefowl in lowland habitats south of the Annamite (CTN and LGO) had less intraspecific genetic structure (fewer groups represented in the sites) and thus exhibited a high degree of genetic variance within each group. The Red Junglefowl in the northern highlands of the Annamite (HBA and YDN) had the opposite spatial pattern, with more groups identified (high positive eigenvalues) and lower within-group variance (dense contours of the sPCA scores).

The spatial autocorrelation analyses of cumulative distance classes indicated that the Red Junglefowl in South Central Vietnam exhibited genetic correlation at fine scales and were very site specific (Figure 1.5 b). The HBA site (and YDN to a lesser degree) showed high degrees of genetic relatedness between the neighboring Red Junglefowl. Then, spatial autocorrelation steadily declined to distances of approximately 5 km, where it had negative correlation as distance increased. The two lowland sites of CTN and LGO, again, had high variance among neighboring samples and showed no autocorrelation. Overall, the four correlograms revealed a transition to negative values at roughly ~ 5 - 6 km distance.

### *Population structure*

Genetic diversity characterized by the previous summary statistics showed a high degree of polymorphism and evidence of private alleles in each sampling site (Table 1.1). The overall $F_{ST}$ for genetic differentiation is 0.1028 (95% resampling confidence interval - 0.0106 to 0.0111). Statistically significant genetic differentiation also existed among the four major sampling sites (Table 1.1), the lowland sites and highland sites were different in their genetic differentiation. Pairwise $F_{ST}$ values ranged from 0.0267 (of HBA/YDN is the most well connected pair) to greater than 0.1500 for long-distance pairs across the Annamite (Figure 1.6).

Bayesian clustering in our non-spatial global model (based on genotypes of all 212 samples) converged on an estimate of eight clusters from the seven studied sites (UPGMA - similar to Figure 1.7, not shown). Each of three sites in the west of the Annamite (CTN, LGO, and YDN) formed a distinct genetic cluster. CTN and YDN are the two larger sites and were further divided into two individual clusters each, as previously observed in the PCA diagram (Figure 1.2). The HBA also formed a unique genetic cluster and then within it, four additional clusters representing possibly admixed populations in the far-east coastal region (HBA-NCA-TKU), northern highlands (HBA-BDP and HBA-YDN) and along the east side of the Annamite (HBA-CTN). Upon incorporating geographic localities (spatial model), *Geneland* also estimated eight geographic/genetic clusters (Figure 1.7). The inferred cluster memberships were almost identical between the spatial and non-spatial model (regression $R^2 = 0.788$, $p < 0.001$). In both the spatial and non-spatial models, the modal number of populations was $\Delta K = 9$ (43% of 1 000 runs). Red Junglefowl from each of these geographic sites tended to cluster with birds from the same region. There are, however, some exceptions as these clusters included some junglefowl from other geographic areas that could represent migrants or perhaps an artifact caused by under-represented sampling of individuals in genetically distinct populations.

Additional resolution was identified when focused at local scale (Figure 1.5 c-d). The Red Junglefowl in CTN were generally assigned to three distinct clusters (61%, or 610 out of 1 000 runs had $\Delta K = 3$) instead of two as being inferred by the global model. Biologically, this may be reflective of two relatively disturbed bamboo forests in the middle of the reserve and a large well protected region connected by northern corridor (represented as two regions at the two sides in the map). HBA, the most complex topographical area among the four major sites, also had one additional cluster (from 4 to 5 in 42% of replicates). Here, the more disturbed central region was separated into two different clusters. The two topographically flat sites, LGO and YDN, yielded more population clusters under local models (from 1 to 4 in 74% of replicates in LGO, and from 2 to 4 in 55% in YDN). There was no clear pattern of cluster assignment based on geography in these two sites; the model mostly assigned junglefowl with proximal geographic localities to the same clusters. This result is realistic as the absence of topography may promote greater connectivity.

We observed that $F_{ST}$ likely underestimated genetic differentiation among our sampled populations. The estimation of $F_{ST}$ is based on observed heterozygosities and would be downward biased when interspecific heterozygosities vary among the sampled individual populations (Hedrick 2005). As such, when applied $F_{ST}$ estimation in the inferred genetic clusters (e.g. the modal $\Delta K$ =9 for the global clustering or the added up 18 from individual local models), genetic differentiation in our data sharply increased to 0.1468 for the whole sampling region and up to 0.3169 pairwise between individual sites with the majority of them were greater than 0.1500 (not reported here). This confirmed the strong population structure of junglefowl in South Central Vietnam as already observed by the ordination and Bayesian methods.

## DISCUSSION

### *Spatial pattern and population structure*

We found strong population genetic structure at coarse geographic scales and evidence of fine-scale genetic subdivision at distances as low as 5 km. The average sampling distance of 1.12 km in our study appears to be appropriate for detecting genetically divergent groups of junglefowl, especially as we did not detect IBD among the birds. Here, we discuss the importance of genetic structure and the scale of spatial ranges in the studied Red Junglefowl, and the implications of these results for genetic management and conservation.

Across our sampling region, the Annamite Mountain Range is likely an impassable barrier for ground-dwelling pheasants, including Red Junglefowl, particularly between the southern lowland and eastern coastal sites. Although Red Junglefowl can occupy habitats up to 1 800 m in elevation (Johnsgard 1999), we rarely observed Red Junglefowl over 600 m. The birds sampled at high elevations were all from the northern highland BDP site, and they exhibited substantial genetic differentiation from lowland populations. Moving further to the YDN northwest highlands, we observed even greater genetic differentiation between their highland sub-populations and the lowland populations (not reported here, mentioned in the biased $F_{ST}$ section). Evolutionary and ecological evidence places South Central Vietnam at the convergence of two biogeographic regions, mostly due to the Annamite Mountain Range: the Annamese Mountains region consists of subtropical drier monsoon habitats in the northwest uplands, and the Cochinchina region of moist tropical lowlands in the south and an acrid microclimatic region in the eastern coast (MacKinnon 1997, Sterling and Hurley 2005). The inferred genetic clusters of Red Junglefowl in our study were geographically concordant to these regions and confirm the importance of the Annamite to the broad-scale population genetic diversity of this species.

The eastern coastal sampling sites contained fewer private alleles compared to those in the western regions of the Annamite. The coastal birds were also consistently grouped into an overall population regardless of the geographic distances between sites. This contrasts to the highland Red Junglefowl, which separate into distinct population clusters even across short spatial distances. Natural dispersal and movement of Red Junglefowl in the coastal region may be limited by the inherent lack of landscape connectivity, as human population densities are very high in this region. However, movement could presumably be enhanced by human capture, transport, and release of wild Red Junglefowl (undoubtedly having occurred over thousands of years) or by genetic introgression from domestic fowl (Storey et al. 2012). Under artificial selection and substantial inbreeding, genetic differentiation of domesticated animals are driven by genetic drift (low effective population sizes) causing rapid fixation of allele frequencies, whereas allele frequency changes will be more gradual in most wild populations (Lenstra et al. 2012). Understanding the mechanism of reduced genetic variation in the coastal Red Junglefowl cluster is not possible with our current dataset, but has important implications for future research of how genetic diversity is being lost in wild Red Junglefowl populations. It is also useful to note that if introgressive hybridization actually occurs, the human transport of female Red Junglefowl may occur largely to the exclusion of males (Storey et al. 2012). Although known to be polygynous, the mating system of Red Junglefowl has not been studied in detail. Nevertheless, territorially dominant males presumably drive only young males to disperse. Thus, the rate of introgression could be limited by male dispersal distance and reproductive success of these introgressed males.

Importantly, our fine-scale spatial analyses suggested distinct characters of metapopulation structure in our sampling sites. Local genetic structure within isolated population clusters typically arises when gene flow and the dispersal range of the organisms are limited. Short-distance dispersal and home range ecology are perhaps the least studied topic in birds (Paradis et al. 1998). It is commonly believed that birds have few behavioral limitations to dispersal due to their substantial vagility (With et al. 1997). However, some recent studies have shown a strong effect of limited dispersal on genetic structure, especially due to sex-biased dispersal (Pierson et al. 2010). In our study, Red Junglefowl had well-defined geographic distribution and some evidence of admixture between sites in close proximity (between two southern lowland sites CTN-LGO, or between three northern highlands YDN-BDP-HBA, or between the panmictic coastal sites HBA-NCA-TKU). Long-distance genetic similarity was rarely observed in the region. The classical stepping-stone model for this genetic pattern can be ruled out as there was no correlation between geographic distances and genetic dissimilarity. On the other hand, a metapopulation structure resulting from the fragmentation of a formerly continuous population or model of completely subdivided populations is not possible based on our data but could be addressed through simulation studies. Under both of these models, our observed pattern of population genetic differentiation implied quite limited gene flow.

### *Spatial sampling design and model performance*

The Monte Carlo inference in Bayesian clustering methods employs resampling randomizations of the observed data as a basis for inference. The presence of IBD - i.e. regular increase or decrease in genetic variability with geographic distance due to non-random mating or limited dispersal - generally leads to many false positives in Bayesian clustering inferences (Meirmans 2012). It is also possible that the model will fail to explain spatially explicit genetic variation (Frantz et al. 2009, Schwartz and McKelvey 2009). Stratified sampling, such as we employed in this study, is particularly effective for gradient analysis in landscape genetics (Storfer et al. 2007) and helps to determine whether or not IBD patterns exist. The lack of significant IBD in our study suggests that our Bayesian clustering inferences are robust. Balkenhol (2009) provided insightful perspectives about model settings, necessary requirements, as well as the importance of combining different methods in Bayesian clustering procedures. In this present study, we emphasized the utility and application of intensive computation when performing the Bayesian MCMC clustering method. With support of a high-throughput computing system, we were able to base our inferences on 1 000 runs when typically 5 or 10 runs are performed for genetic population studies. This enhanced capability allowed us to construct average posterior densities of individual population membership and served to add an additional interpretive dimension. As a statistical procedure, this allowed us more confidence that stochastic processes inherent to high-dimensional MCMC inferences did not bias the underlying genetic characteristics in our dataset. Deterministic individual-based methods, such as PCA and sPCA models employed in this study, do provide intuitive information about genetic variation from reasonably sized datasets with strong population structure. They also helped to provide a framework in which to guide more sophisticated analyses.

The caveat of applying spatially explicit Bayesian clustering in relation to a spatial sampling design, however, is that sampling randomness does not hold at very fine scales, as most living organism are genetically correlated at these scales (Guillot et al. 2009). This represents trade-offs not only in spatial analysis procedures (Fortin and Dale 2005) but also in sampling schemes for spatial and landscape genetics (Guillot et al. 2009). In our global spatial models, when all data points were included, the requirement of complete spatial randomness (CSR) is generally met. These models, however, provide low resolution in identifying transition areas between local population clusters. On the other hand, our analyses focusing on the local spatial scale identified unique structure in transition areas between populations inhabiting pristine forest and disturbed sites. This interesting information suggests that Red Junglefowl may utilize human modified landscapes, though these areas are likely recolonized secondarily from more pristine forests. As a ground-dwelling species, Red Junglefowl have specialized feeding and territory requirements of open canopy and clear understory floor to feed on leaf-litter invertebrates (Beebe 1926). We suggest that secondary or marginal habitat preference in Red Junglefowl allows the birds to quickly cross open landscape structure but this also makes them particularly prone to hybridizing with domesticated chickens inhabiting the area.

### *Conservation and management implications*

The ordination and Bayesian analyses used in this study found a very complex population structure in Red Junglefowl. Our models supported a metapopulation with divergent clusters of Red Junglefowl occurring over large-scale landscapes. Whether these populations have been separated by unsuitable habitats or due to the birds' spatial range remains unclear. Given the current level of human activities in at least some of the regions, it is unlikely that Red Junglefowl in the existing protected areas are or will be connected by natural dispersal. A metapopulation with limited gene flow will quickly accrue genetic divergences and is significant to diversity of the species (Keyghobadi 2007). Curiously, if there is a desire to conserve alleles of Red Junglefowl, separating a once large and contiguous population into many small sub-populations is an ideal means to achieve this goal (Denniston 1978).

Using Red Junglefowl as a model species, and the results of our current analyses, we expect that metapopulation patterns also occur in other Phasianids occurring in or endemic to the same region. In contrast to Red Junglefowl, the distribution and habitat requirements are much narrower, and in many cases, remain unstudied. The relevance of our research with wild Red Junglefowl in their core distribution ranges has important implications for genetic resource management and conservation of this species of agricultural importance.

From a broad-scale perspective, the characterization of geographically structured Red Junglefowl populations suggests site-specific conservation strategies for the species, e.g. the designation of a distinct management unit (Moritz 1994). Given our observed patterns of genetic diversity, we emphasize the importance of populations in the western portions of the Annamite Mountain Range. In this area, large Red Junglefowl populations still occur in well-protected natural habitats and exhibit high intraspecific genetic variation. Historically, the Annamite Mountain Range likely acted as a topographical barrier to Red Junglefowl. The eastern coastal areas currently have less genetic variation, likely due to substantial human impacts; much of the historical variation may already be lost.

Whether the observed divergences in the wild Red Junglefowl in this present study will lead to conservation of the species and applications in future poultry genetic management depends on additional initiatives to further identify ancestral lineages, resiliency to current environmental changes, and management programs that sustain the core distribution ranges of the species in South and Southeast Asia.  Further efforts to estimate adaptive variation, such as immunological diversity at the well-known major histocompatibility complex (MHC) (Hess and Edwards 2002), may strengthen the appeal of preserving wild Red Junglefowl and related species.  Research on functional genetic diversity in Red Junglefowl could highlight their importance in understanding the significant of variation as considerable genetic diversity has been lost at both academic and industry locations over the past four decades (Fulton and Delany 2003).

# REFERENCES

Arrigo, N., J. W. Tuszynski, D. Ehrich, T. Gerdes, and N. Alvarez. 2009. Evaluating the impact of scoring parameters on the structure of intra-specific genetic variation using *RawGeno*, an *R* package for automating AFLP scoring. BMC Bioinformatics **10**(33).

Balkenhol, N. 2009. Evaluating and improving analytical approaches in landscape genetics through simulations and wildlife case studies. Ph.D Dissertation. University of Idaho, Ann Arbor.

Beebe, W. 1926. A monograph of the Pheasants. Dover Publications, New York.

Berlin, S., L. J. Qu, X. Y. Li, N. Yang, and H. Ellegren. 2008. Positive diversifying selection in avian *Mx* genes. Immunogenetics **60**(11):689-697.

Berres, M. E. 2003. The Roles of Mating System and Landscape Topography in Shaping the Genetic Population Structure of the White-Bearded Manakin (*Manacus manacus*) in Trinidad, West Indies,. PhD Thesis. University of Wisconsin-Madison, Madison, Wisconsin.

Berthouly, C., X. Rognon, T. N. Van, et al. 2010. Vietnamese chickens: a gate towards Asian genetic diversity. BMC Genetics **11**(53).

Bivand, R., E. J. Pebesma, and V. Gómez-Rubio. 2008. Applied spatial data analysis with R. Springer, New York.

Bjornstad, O. 2013. *ncf*: spatial nonparametric covariance functions. *R* package version 1.1-5. http://CRAN.R-project.org/package=ncf.

Bjornstad, O. N., R. A. Ims, and X. Lambin. 1999. Spatial population dynamics: analyzing patterns and processes of population synchrony. Trends in Ecology & Evolution **14**(11):427-432.

Brisbin, I. L., Jr., A. T. Peterson, R. Okimoto, and G. Amato. 2002. Characterization of the genetic status of populations of Red Junglefowl. Journal of the Bombay Natural History Society **99**(2):217-223.

Bub, H. 1991. Bird trapping and bird banding - A handbook for trapping methods all over the world. Cornell University Press, Ithaca, New York.

Delany, M. E. 2006. Avian genetic stocks: The high and low points from an academia researcher. Poultry Science **85**(2):223-226.

Denniston, C. 1978. Small population size and genetic diversity. Implication for endangered species. Pages 281-289 *in* S. A. Temple, editor. Endangered birds. Management techniques for preserving threatened species. University of Wisconsin Press, Madison, WI.

Downing, T., D. J. Lynn, S. Connell, et al. 2009. Contrasting evolution of diversity at two disease-associated chicken genes. Immunogenetics **61**(4):303-314.

Dray, S., and A. B. Dufour. 2007. The *ade4* package: Implementing the duality diagram for ecologists. Journal of Statistical Software **22**(4):1-20.

FAO. 2007. The State of the World's Animal Genetic Resources for Food and Agriculture. Food and Agriculture Organisation of the United Nations, Rome.

Fortin, J. M., and M. Dale. 2005. Spatial Analysis: A Guide for Ecologists. Cambridge University Press, Cambridge.

Frantz, A. C., S. Cellina, A. Krier, L. Schley, and T. Burke. 2009. Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? Journal of Applied Ecology **46**(2):493-505.

Fulton, J. E., and M. E. Delany. 2003. Poultry genetic resources - Operation rescue needed. Science **300**(5626):1667-1668.

Fumihito, A., T. Miyake, S. I. Sumi, M. Takada, S. Ohno, and N. Kondo. 1994. One subspecies of the Red Junglefowl *Gallus gallus gallus* suffices as the matriarchic ancestor of all domestic breeds. Proceedings of the National Academy of Sciences of the United States of America **91**(26):12505-12509.

Goslee, S. C., and D. L. Urban. 2007. The *ecodist* package for dissimilarity-based analysis of ecological data. Journal of Statistical Software **22**(7):1-19.

Granevitze, Z., J. Hillel, G. H. Chen, N. T. K. Cuc, M. Feldman, H. Eding, and S. Weigend. 2007. Genetic diversity within chicken populations from different continents and management histories. Animal Genetics **38**(6):576-583.

Groeneveld, L. F., J. A. Lenstra, H. Eding, et al. 2010. Genetic diversity in farm animals - a review. Animal Genetics **41**:6-31.

Guillot, G., R. Leblois, A. Coulon, and A. C. Frantz. 2009. Statistical methods in spatial genetics. Molecular Ecology **18**(23):4734-4756.

Guillot, G., F. Mortier, and A. Estoup. 2005. *GENELAND*: a computer package for landscape genetics. Molecular Ecology Notes **5**(3):712-715.

Hedrick, P. W. 2005. A standardized genetic differentiation measure. Evolution **59**(8):1633-1638.

Hess, C. M., and S. V. Edwards. 2002. The evolution of the major histocompatibility complex in birds. Bioscience **52**(5):423-431.

Jaccard, P. 1908. Nouvelles recherches sur la distribution florale. Bulletin de la Société Vaudoise Des Sciences Naturelles **44**:223-270.

Johnsgard, P. A. 1999. The Pheasants of the world: biology and natural history. Smithsonian Institution Press, Washington, D.C.

Jombart, T., S. Devillard, A. B. Dufour, and D. Pontier. 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity **101**(1):92-103.

Jongman, R., C. Braak, and O. van Tongeren. 1995. Data Analysis in Community and Landscape Ecology. Cambridge University Press.

Kalinowski, S. T. 2004. Counting alleles with rarefaction: Private alleles and hierarchical sampling designs. Conservation Genetics **5**(4):539-543.

Keyghobadi, N. 2007. The genetic implications of habitat fragmentation for animals. Canadian Journal of Zoology-Revue Canadienne De Zoologie **85**(10):1049-1064.

Lenstra, J. A., L. F. Groeneveld, H. Eding, et al. 2012. Molecular tools and analytical approaches for the characterization of farm animal genetic diversity. Animal Genetics **43**(5):483-502.

Longmire, J. L., M. Maltbie, and R. Baker. 2000. Use of "Lysis Buffer" in DNA isolation and its implication for museum collections. Texas Tech University Museum.

MacKinnon, J. R. 1997. Protected areas systems review of the Indo-Malayan realm. Asian Bureau for Conservation, Canterbury, England.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Journal of Cancer Research **27**:209-220.

Marschalek, D. A., and M. E. Berres. 2014. Genetic Population Structure of the Blister Beetle: Core and Peripheral Populations. Journal of Heredity **105**(6):784-792.

Meirmans, P. G. 2012. The trouble with isolation by distance. Molecular Ecology **21**(12):2839-2846.

Mekchay, S., P. Supakankul, A. Assawamakin, A. Wilantho, W. Chareanchim, and S. Tongsima. 2014. Population structure of four Thai indigenous chicken breeds. BMC Genetics **15**(40).

Moran, P. A. 1950. A test for the serial independence of residuals. Biometrika **37**:178-181.

Moritz, C. 1994. Defining Evolutionarily-Significant-Units for Conservation. Trends in Ecology & Evolution **9**(10):373-375.

Muir, W. M., G. K. S. Wong, Y. Zhang, et al. 2008. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. Proceedings of the National Academy of Sciences of the United States of America **105**(45):17312-17317.

Ngo, T. K. C., H. Simianer, H. Eding, H. V. Tieu, V. C. Cuong, C. B. A. Wollny, L. F. Groeneveld, and S. Weigend. 2010. Assessing genetic diversity of Vietnamese local chicken breeds using microsatellites. Animal Genetics **41**(5):545-547.

Oksanen, J., F. Blanchet, Kindt R, et al. 2013. *vegan*: Community Ecology Package. *R* package version 2.0. http://CRAN.R-project.org/package=vegan.

Paradis, E., S. R. Baillie, W. J. Sutherland, and R. D. Gregory. 1998. Patterns of natal and breeding dispersal in birds. Journal of Animal Ecology **67**(4):518-536.

Peterson, A. T., and I. L. Brisbin. 1998. Genetic endangerment of wild Red Junglefowl *Gallus gallus*. Bird Conservation International **8**(4):387-394.

Pierson, J. C., F. W. Allendorf, V. Saab, P. Drapeau, and M. K. Schwartz. 2010. Do male and female black-backed woodpeckers respond differently to gaps in habitat? Evolutionary Applications **3(**3):263-278.

QGIS Development Team. 2014. QGIS Geographic Information System. http://qgis.osgeo.org. Open Source Geospatial Foundation Project.

R Development Core Team. 2011. ***R***: A Language and Environment for Statistical Computing. http://www.R-project.org. R Foundation for Statistical Computing.

Schwartz, M. K., and K. S. McKelvey. 2009. Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. Conservation Genetics **10**(2):441-452.

Sokal, R. 1986. Spatial data analysis and historical processes. *In* Fourth Intl Symposium of Data Analysis and Informatics. North Holland, Versailles, France.

Sterling, E., and M. M. Hurley. 2005. Conserving Biodiversity in Vietnam: applying Biogeography to conservation research. Proceeding of the California Academy of Sciences **56**(I9):98-118.

Storey, A. A., J. S. Athens, D. Bryant, et al. 2012. Investigating the global dispersal of chickens in rrehistory using ancient mitochondrial DNA signatures. PLOS One **7**(7).

Storfer, A., M. A. Murphy, J. S. Evans, et al. 2007. Putting the 'landscape' in landscape genetics. Heredity **98**(3):128-142.

Vekemans, X., T. Beauwens, M. Lemaire, and I. Roldan-Ruiz. 2002. Data from AFLP markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. Molecular Ecology **11**(1):139-151.

Vos, P., R. Hogers, M. Bleeker, et al. 1995. AFLP - a New Technique for DNA-Fingerprinting. Nucleic Acids Research **23**(21):4407-4414.

Weiner, J. 2013. *pca3d*: Three dimensional PCA plots. *R* package version 0.2. http://CRAN.R-project.org/package=pca3d.

With, K. A., R. H. Gardner, and M. G. Turner. 1997. Landscape connectivity and population distributions in heterogeneous environments. Oikos **78**(1):151-169.

Wright, S. 1943. Isolation by Distance. Genetics **28**:114-138.

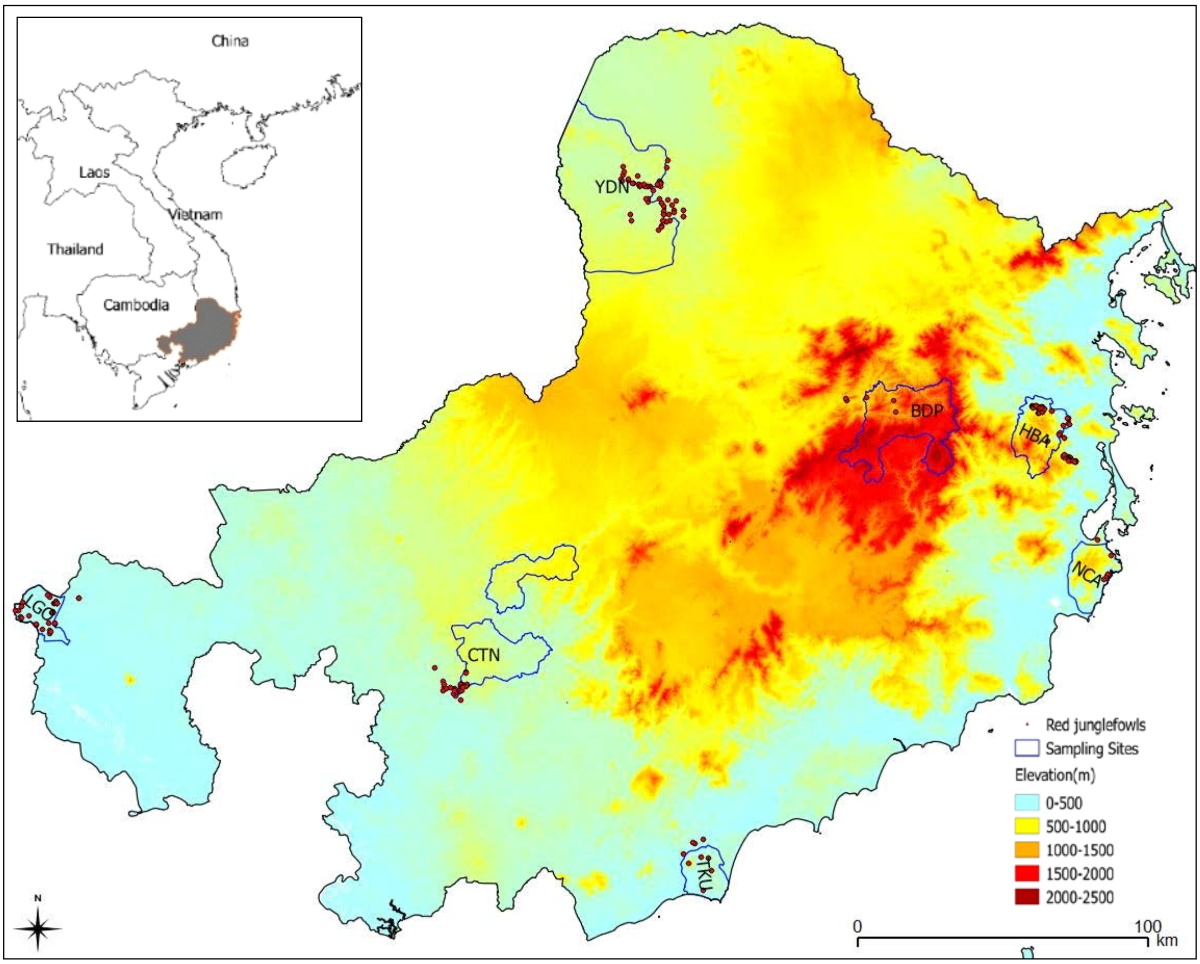
Zhivotovsky, L. A. 1999. Estimating population structure in diploids with multilocus dominant DNA markers. Molecular Ecology **8**(6):907-913.

Table 1.1: Genetic diversity of Red Junglefowl in seven field sites.

*Abbreviations: sample sizes (N); proportion of polymorphic markers (PLP); expected heterozygosity ($H_E$) with standard error (s.e.); private alleles (PA); genetic differentiation ($F_{ST}$ with 95% resampling confidence intervals) (for the four major sampling sites with > 30 samples). Bi Doup - Nui Ba National Park (BDP), Cat Tien National Park and Dong Nai Nature Reserve (CTN), Hon Ba Nature Reserve (HBA), Lo Go Sa Mat National Park (LGO), Nui Chua National Park (NCA), Ta Kou Nature Reserve (TKU), and Yok Don National Park (YDN).*

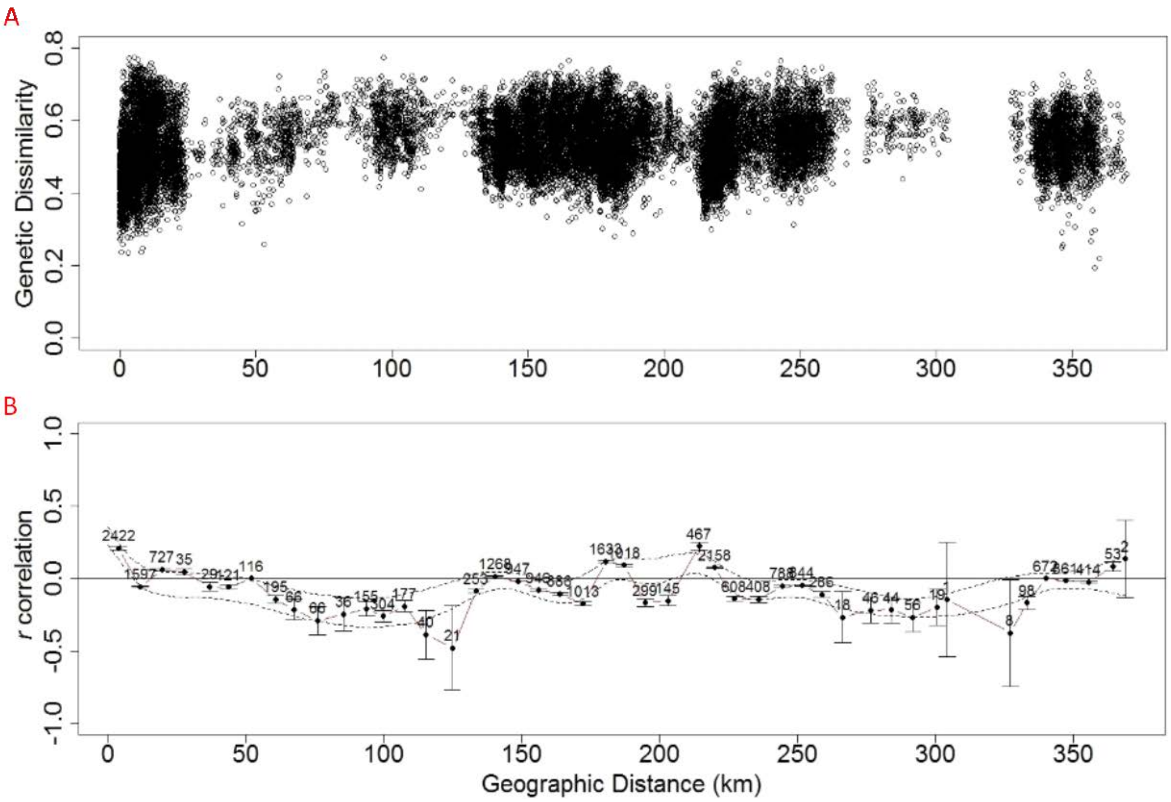
| **Sampling sites** | ***N*** | ***PLP*** | ***He (± s.e)*** | ***PA*** | ***$F_{ST}$***<br>**(95% resampling confidence)** |
|---|---|---|---|---|---|
| **BDP** | 5 | 0.296 | 0.1242 ± 0.0091 | 1 | -- |
| **CTN** | 44 | 0.445 | 0.1533 ± 0.0086 | 16 | 0.0713<br>(-0.0132, 0.0129) |
| **HBA** | 56 | 0.427 | 0.1492 ± 0.0086 | 8 | 0.1392<br>(-0.0098, 0.0123) |
| **LGO** | 34 | 0.368 | 0.1243 ± 0.0083 | 9 | 0.0625<br>(-0.0124, 0.0160) |
| **NCA** | 6 | 0.386 | 0.1380 ± 0.0089 | 3 | -- |
| **TKU** | 9 | 0.432 | 0.1432 ± 0.0091 | 9 | -- |
| **YDN** | 58 | 0.458 | 0.1916 ± 0.0089 | 33 | 0.1559<br>(-0.0165, 0.0216) |

Figure 1.1: Sampling sites with the Annamite topography.



*Bi Doup - Nui Ba National Park (BDP), Cat Tien National Park and Dong Nai Nature Reserve (CTN), Hon Ba Nature Reserve (HBA), Lo Go Sa Mat National Park (LGO), Nui Chua National Park (NCA), Ta Kou Nature Reserve (TKU), and Yok Don National Park (YDN).*

Figure 1.2: Overall spatial genetic patterns for all sampling sites.



*(A) Pairwise genetic over geographic distances and (B) Correlogram of genetic correlation with numbers indicate pair-wise dissimilarities within the distance classes.*

Figure 1.3: Principal Component Analysis of genetic variation.

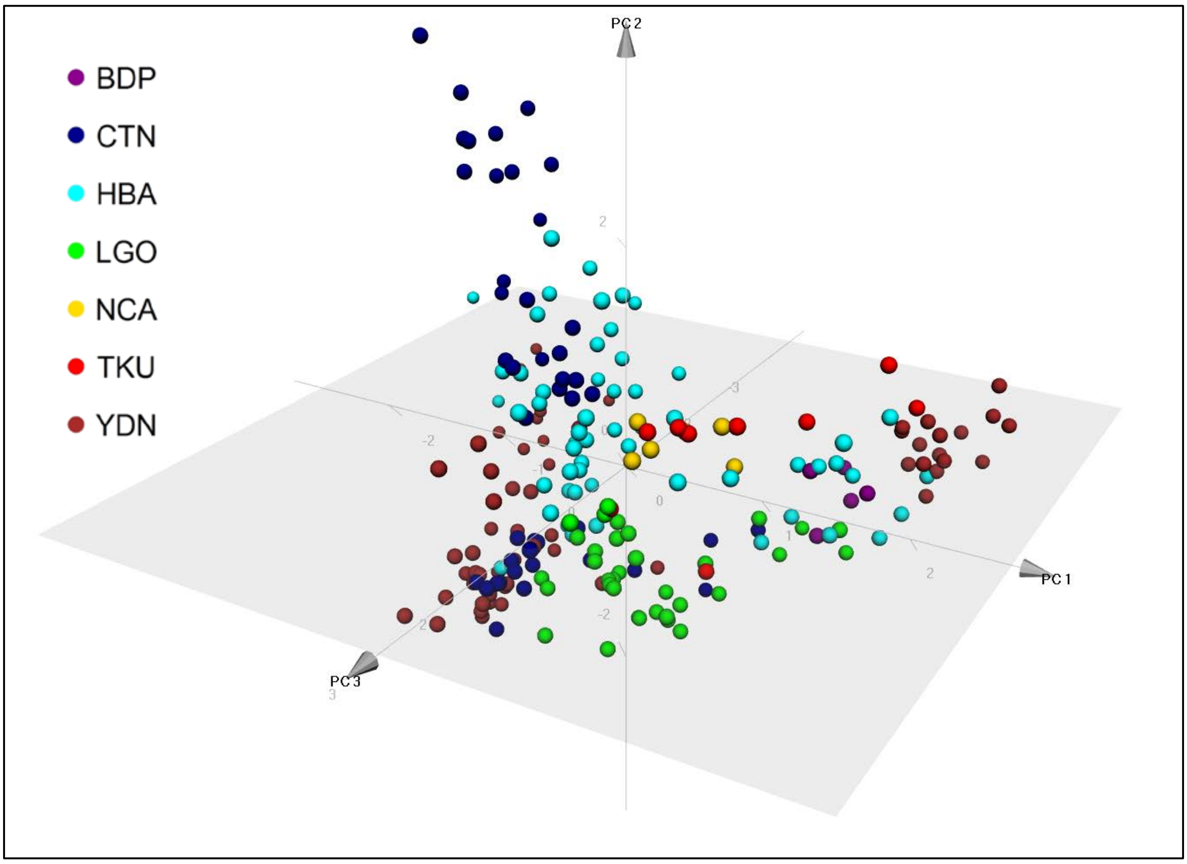
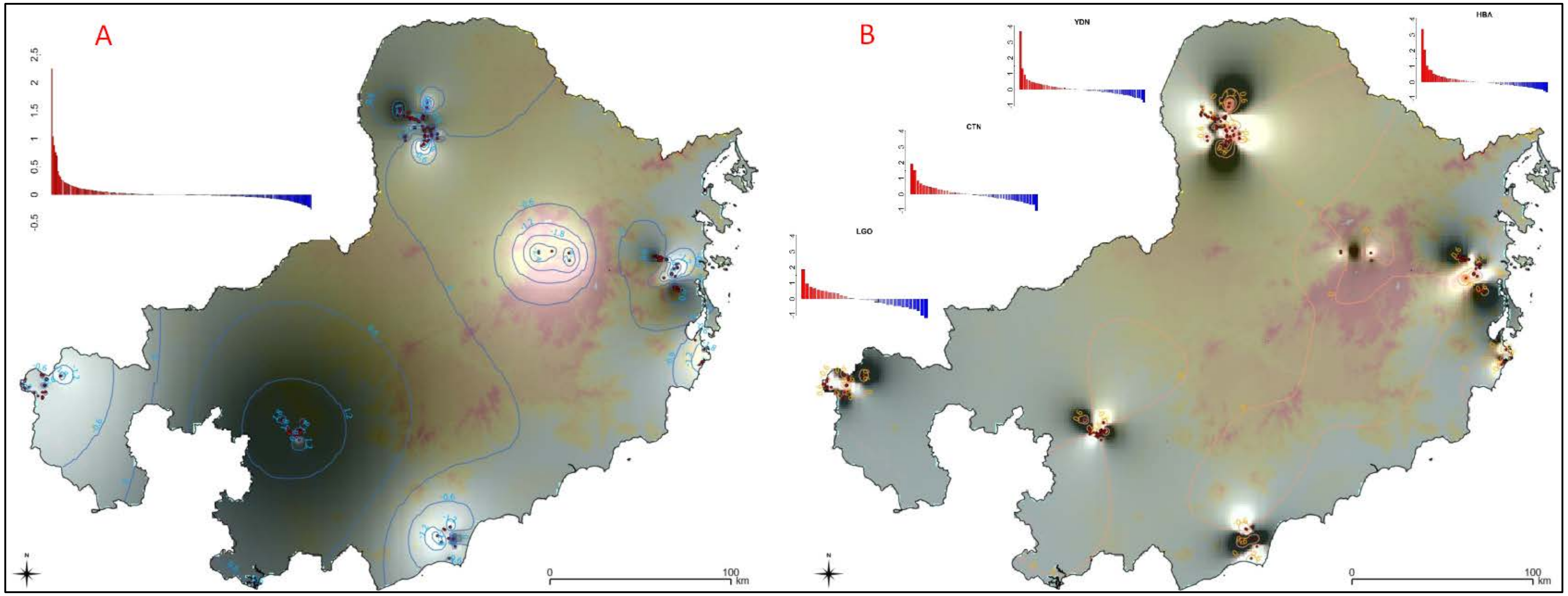
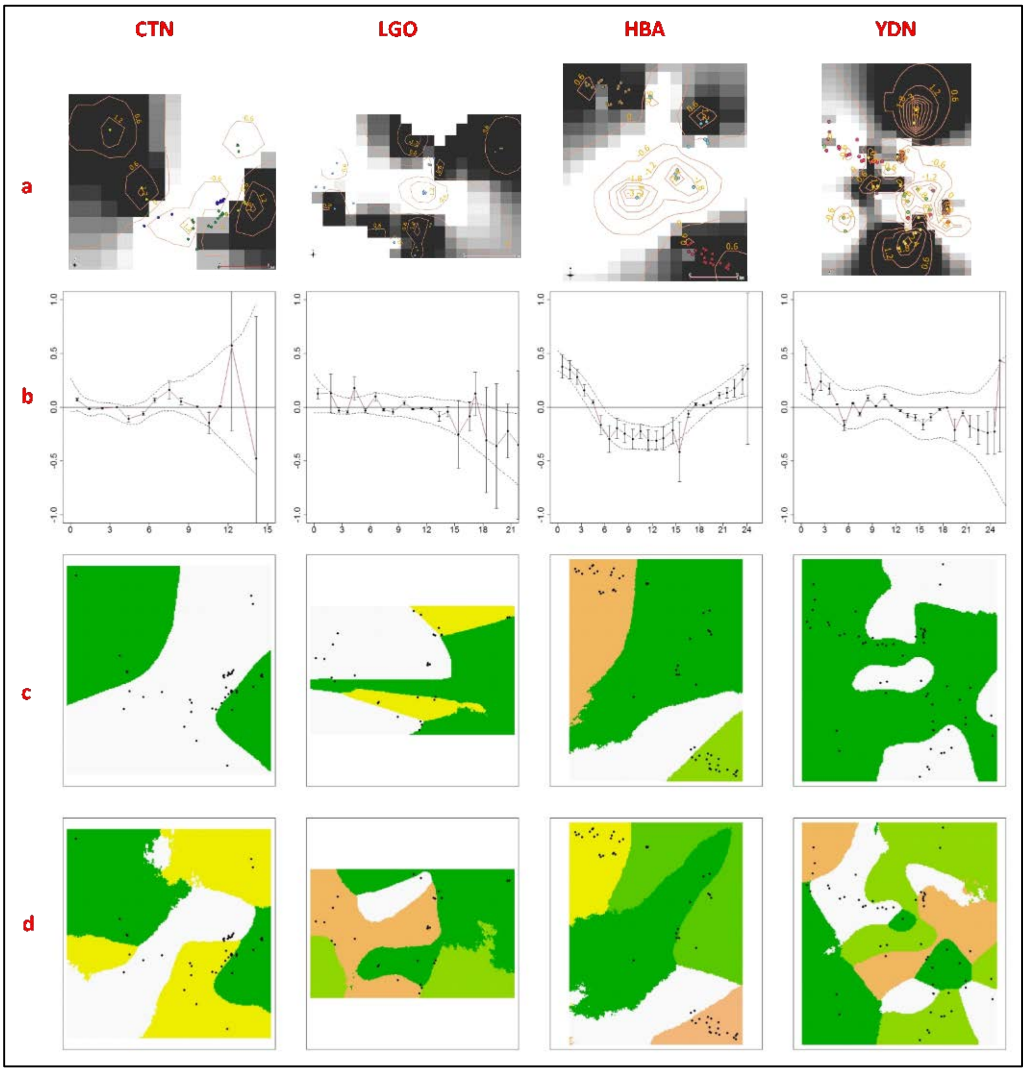


Figure 1.4: Spatial Principal Component Analysis (sPCA).



*(A) Global scale. (B) Local scale (B). Dots are samples, contours define similarities in the component scores across the landscape. Inserts are the respective eigenvalues of the sPCAs.*

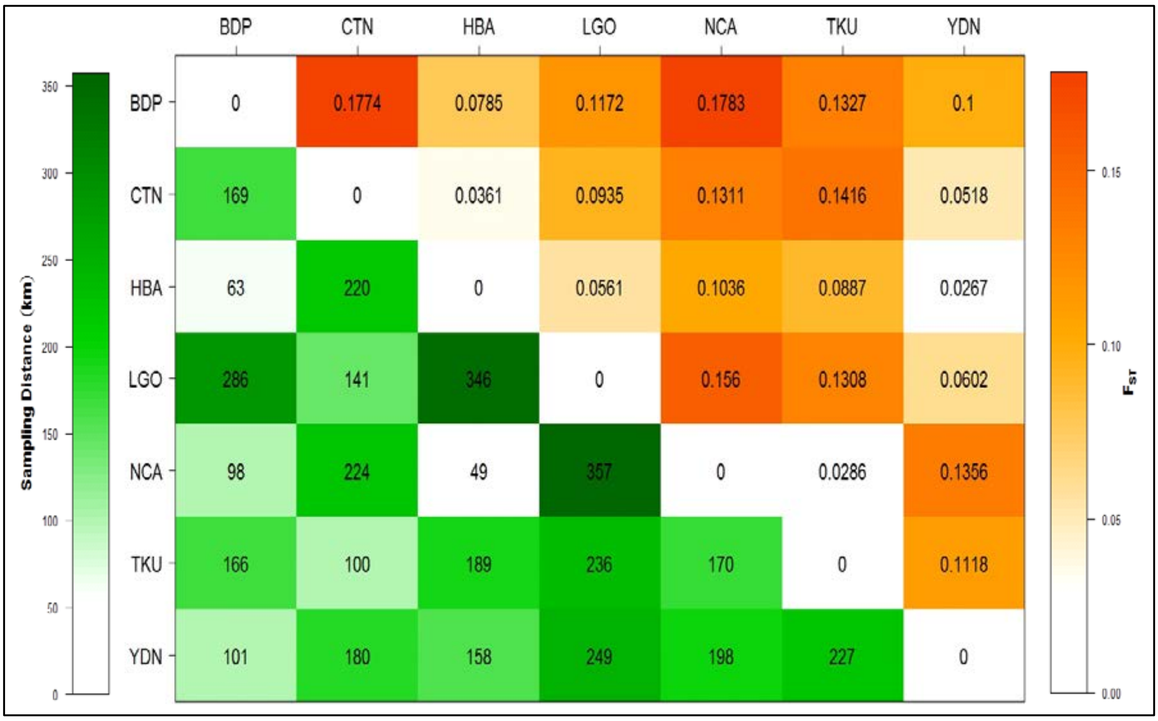
Figure 1.5: Spatial population structure.



*(a) sPCA, (b) correlogram, (c) Bayesian global clustering, (d) Bayesian local clustering.*

*Legends: (a) dots: samples, contours: component scores for similarity; (b) ordinate: correlations; abscissa: cumulative distance classes; error bars: 95% confidence bootstrapped; dashed lines: confidence intervals of 1 000 permutations around the null hypothesis of a random distribution; (c) & (d) dots: samples, color regions are the posterior probability spatial clusters; plots represent only one run (out of 1 000) that has highest log posterior density.*

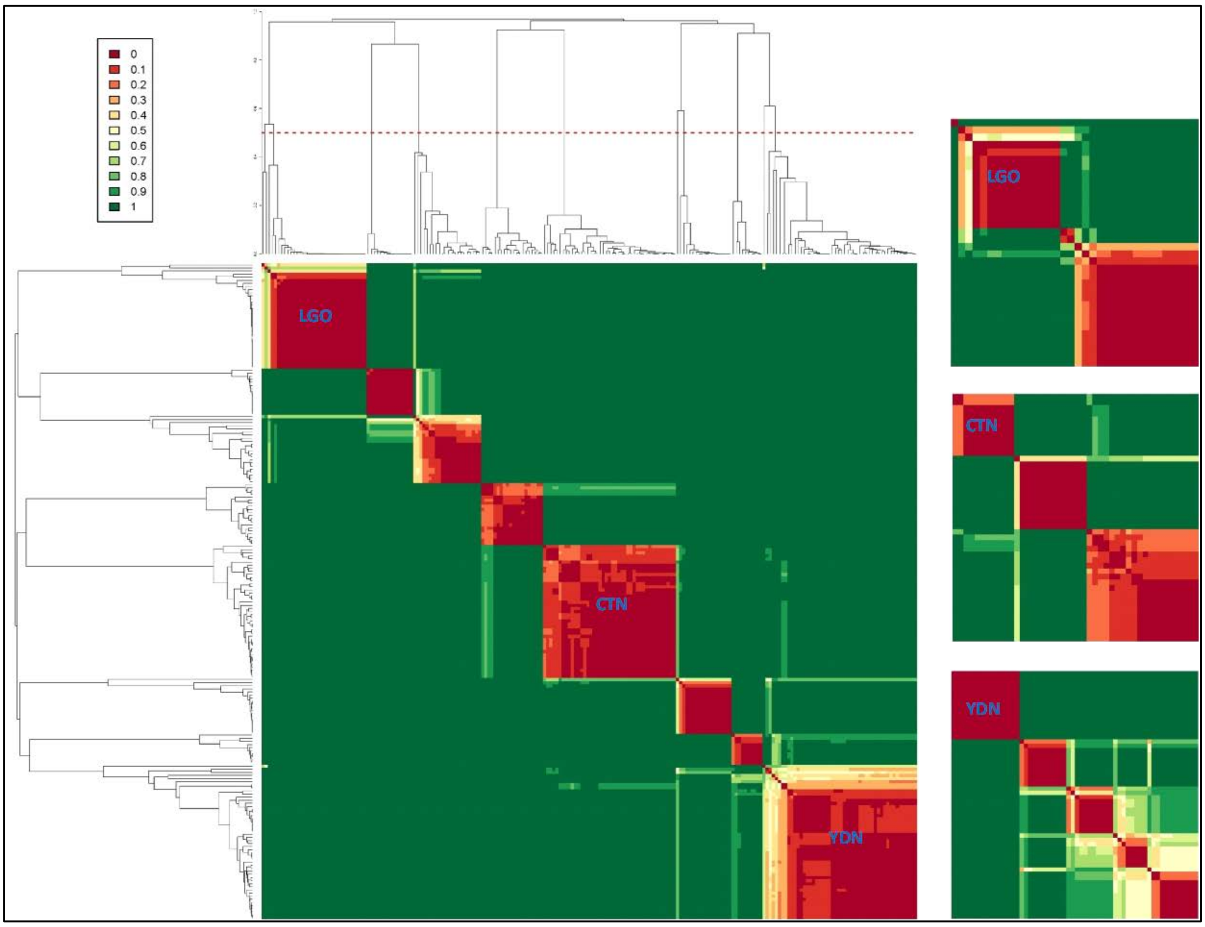
Figure 1.6: Pairwise genetic differentiation $F_{ST}$ and geographic distances.



| | BDP | CTN | HBA | LGO | NCA | TKU | YDN |
|---|---|---|---|---|---|---|---|
| BDP | 0 | 0.1774 | 0.0785 | 0.1172 | 0.1783 | 0.1327 | 0.1 |
| CTN | 169 | 0 | 0.0361 | 0.0935 | 0.1311 | 0.1416 | 0.0518 |
| HBA | 63 | 220 | 0 | 0.0561 | 0.1036 | 0.0887 | 0.0267 |
| LGO | 286 | 141 | 346 | 0 | 0.156 | 0.1308 | 0.0602 |
| NCA | 98 | 224 | 49 | 357 | 0 | 0.0286 | 0.1356 |
| TKU | 166 | 100 | 189 | 236 | 170 | 0 | 0.1118 |
| YDN | 101 | 180 | 158 | 249 | 198 | 227 | 0 |

*$F_{ST}$ is above the diagonal and d geographic distances in kilometers is below the diagonal.*

Figure 1.7: *Geneland*'s UPGMA dendrogram.



*From 1 000 iterations of global Geneland's Bayesian clustering (spatial model) with three inserts representing local clustering (LGO, CTN, and YDN).*

# CHAPTER 2

## Spatial dependence models and correlation of neutral genetic variation in Red Junglefowl (*Gallus gallus*)

## ABSTRACT

**Context**:  Hypothesis tests in landscape genetic frameworks are typically based on a non-spatial null model, the Island Model, the assumptions of which are confounded when using spatially organized and biologically non-independent data.  Spatial genetic processes such as restricted gene flow, autocorrelated allele frequencies, and environment-induced genetic variation create many unrealistic assumptions within the Island Model; real (natural) populations are very likely to violate these assumptions and create substantial analytical bias.

**Objectives**:  We present a spatially explicit framework that integrates methods of multivariate analysis, geostatistics, and GIS techniques, all of which are applicable for analysis of spatial data.  To evaluate if genetic variation has a spatial dependence on landscape features, we examined if geographic distances or patterns of major landscape features were the primary factors responsible for the extent of genetic structure in a terrestrial pheasant species, Red Junglefowl (*Gallus gallus*).

**Methods**:  We used a dataset containing 386 neutral genetic markers in 192 Red Junglefowl sampled from four diverse landscapes in South Central Vietnam.  Using observed and simulated genetic data, we evaluated the relative influences of landscape features, sampling localities, and genetic population structure to determine if landscape features modulated patterns of observed allele frequencies.

**Results**:  Allele frequencies were weakly autocorrelated in Red Junglefowl that were trapped within 6 km of each other in the two lowland populations and there was no genetic autocorrelation at any sampling extent in the two highland populations.  We found no evidence of spatial dependence in genetic variation to the two landscape features of landcover and topography. This suggests that the spatial genetic variation in the Red Junglefowl is more related to demography or specific movement characteristics (or both) rather than any dependence on landscape or sampling arrangements.

**Conclusions**: A spatially-explicit framework was able to exclude any correlative strength of genetic variation and landscape features in natural populations of Red Junglefowl. The application of landscape genetics promotes aims of improving conservation planning for Red Junglefowl and other related Phasianids, many of which are critically endangered. In this case study, spatial genetic variation of this agriculturally important predecessor suggests site-specific conservation plans, particularly in areas with increased human activity.

**Keywords**: autocorrelation, multivariate analysis, geostatistics, landscape genetics, Red Junglefowl, spatial dependence.

# INTRODUCTION

One primary goal of population genetics, particularly when applied to naturally occurring populations of organisms, is to detect and locate genetic discontinuities. If spatial delineations are observed within a population, it is often concluded that each population is a distinct, evolutionary operational unit, perhaps connected to other nearby populations by some estimated level of gene flow. Analyses of this type are important as they can be used to guide species management decisions, population reintroduction, and the monitoring of threatened species. However, identifying the reason(s) why these discontinuities occur is often only possible in limited cases, generally as an effect of physical landscape features or demographic characteristics of the study organism. Interpretation of the source of the genetic pattern becomes even more difficult if population differentiation occurs, but does so in a putatively “uniform” environment. A new approach, landscape genetics, has emerged as an empirical discipline designed to integrate information about landscape features and their effects on genetic variation in a spatially explicit manner (Manel et al. 2003, Anderson et al. 2010).

Like classical population genetics and phylogeography, landscape genetics also seeks to detect and locate genetic discontinuities in natural populations (Diniz et al. 2009). Importantly, it also seeks to assess the correlation of these discontinuities with landscape and environmental features. Thus, while both disciplines are concerned with microevolutionary processes, landscape genetics is more concerned with how genetic structure is created across space explicitly. The prevailing framework in landscape genetics is to first collect genetic samples of organisms, preferably at fine scales across a landscape. Second, molecular technologies are applied to detect genetic variation and the presence or absence of genetic structure. This is best accomplished with clustering methods that do not predefine a population (e.g. collection site). Third, an attempt is made to correlate patterns of allelic diversity with landscape features using spatially explicit models.

Hypothesis tests in this framework are typically based on a non-spatial null model, the Island Model (Wright 1943). However, particularly in natural populations, assumptions implicit under this model are often violated when using spatial, categorical, and strongly multivariate information such as genetic markers (Wagner and Fortin 2005). Naturally occurring spatially dependent processes such as gene flow may be autocorrelated , or isolation-by-distance (IBD) (Wright 1943) - the equivalent term 'spatial autocorrelation' (Legendre 1993) - and environmentally induced genetic variation creates many unrealistic assumptions that violate the Island Model framework.

Spatial autocorrelation is a phenomenon where values of some variables sampled at nearby locations are not independent from each other (Legendre 1993). For example, in continuously distributed populations with genetic autocorrelation, individual dispersal is expected to be patterned such that levels of gene flow gradually decrease with increasing geographic distances. This scenario will create a gradient of genetic differentiation among individuals, the patterns of which depend on not only demography but also the dispersal process itself (e.g. one- or two-dimensional movement patterns). Essentially, the dispersal process is also dependent on the landscape or some environment features, themselves likewise have spatial autocorrelation in their patterns. This may act as confounding factor and create additional spatial structuring or 'spatial dependence' in genetic variation

(Jongman et al. 1995, Wagner and Fortin 2005, Dormann et al. 2007). Although being indispensable parts in ecological and evolutionary processes, the two spatial phenomena of autocorrelation and dependence violate the assumption of independent and identical distribution (i.i.d.) in statistical tests and hence perturb significant tests and inflate Type I errors (Dormann et al. 2007, Guillot et al. 2009).

Specifically, hypothesis tests in landscape genetics studies that attempt to correlate genetic variation and environmental features may be confounded in at least two ways. First, models that estimate spatial genetic variation and population structure, including Bayesian clustering methods, are biased by the presence of genetic IBD (Frantz et al. 2009, Schwartz and McKelvey 2009, Meirmans 2012). Here, the Bayesian inference algorithms may erroneously create local genetic clusters - equivalent to populations - from samples whose genetic differentiation comprise a gradient due to spatially correlated allele frequencies. Second, tests for landscape correlation and hierarchical population structure, such as Mantel's permutation (Mantel 1967, Sokal 1986), are also biased by spatial dependence and non-independent data. These biases underestimate sampling variances and introduce unpredictable and spurious effects (Balkenhol et al. 2009, Guillot et al. 2009, Meirmans 2012). The development, testing, and application of spatially explicit models combined with assessment of causative relationships between genetic variation and landscape features will make the application of landscape genetics to natural populations more accurate and precise.

The present study aims to determine if spatial autocorrelation exists in wild populations of Red Junglefowl (*Gallus gallus*) and whether or not their allelic distribution is correlated with landscape features of landcover and topography. The Red Junglefowl, a medium-sized ground-dwelling pheasant, is widely considered the direct ancestor of domestic chickens. Before spreading globally through human-mediated dispersal (Storey et al. 2012), the domestication of wild junglefowl was thought to initiate in Asia, likely from Red and Green Junglefowl and perhaps other related lineages (Fumihito et al. 1994, Eriksson et al. 2008). Red Junglefowl occur in lowland tropical rainforests up to 600 m in South and Southeast Asia. Natural habitats in these areas have been greatly modified in recent decades (Fuller and Garson 2000). However, many protected areas still contain Red Junglefowl and the species is locally abundant in wooded habitats. Some authors even consider the a species a presumptive forest indicator (Beebe 1926, Johnsgard 1999).

In a recent study, we documented a strong global and fine-scale genetic structure of Red Junglefowl populations in the South Central region of Vietnam (Chapter 1). Populations were assigned to different genetic clusters along two sides of the Annamite Mountain Range, a major dispersal barrier to Red Junglefowl in the region. Our analysis suggested a ‘classical’ metapopulation structure where, contradictory to the overall range and natural abundance of the species, intra-population genetic differentiation was significant and long-distance dispersal either no longer occurred or was maintained at very low levels.

Given the expansive distribution of Red Junglefowl and seemingly large, yet apparently genetically isolated populations, we predict geographic distance and the breeding structure *per se* to be the most important factors that explain local genetic variation in Red Junglefowl in the study landscapes. Although the two concerned landscape features in the study of landscape and topography vary within each landscape, we hypothesize that Red Junglefowl are sufficiently flexible in their habitat requirements relative to the local landscape variation, and thus we expect that elevation and land cover - or spatial dependence - will not influence the birds' genetic structure.

## METHODS

### *Study area and sampling*

We sampled wild Red Junglefowl ($N$ = 192) in four protected areas in South Central Vietnam: Cát Tiên National Park and Đồng Nai Nature Reserve (hereafter CTN as the two sites are connected, $n$ = 44), Hòn Bà Nature Reserve (HBA, $n$ = 56), Lò Gò Sa Mát National Park (LGO, $n$ = 34), and Yok Đôn National Park (YDN, $n$ = 58) (Figure 2.1). The sites are separated from one another (~ 230 km) by residential and non-natural habitats. They were divided into two groups based on altitudinal ranges and habitat types. The average elevation of lowland sampling localities in CTN and LGO was 80 m (range: 5 - 120 m) above sea level and predominantly featured lowland tropical rainforest. The average elevation of highland sites (HBA and YDN) was 250 m (range: 60 - 450 m) above sea level and was composed of mixed tropical and deciduous forests. Sites CTN and YDN were the two larger reserves (mean: 99 000 hectares) compared to HBA and LGO (mean: 22 000 hectares).

Red Junglefowl were live-captured by the walking-snare method (Chapter 1) in three dry seasons in 2012, 2013, and 2014. We stratified sampling within the field sites to account for the presence of junglefowl and capture opportunities. We computed the point pattern $\hat{K}$ function (Ripley 1981) to quantify the influence of stratified sampling in relation to genetic variation and landscape features. The significance of $\hat{K}$ in each field site was estimated with

1 000 Monte Carlo simulations using the *spatstat* package (Baddeley and Turner 2005) in the open-source *R* 3.1.2 environment (R Development Core Team 2011).

### *Genetic analyses*

Blood samples from live-captured junglefowl were used as a source of DNA for genetic analyses. We generated AFLP fingerprints with two primer pairs *Eco*RI+CAT/ *Ase*I+GA and *Eco*RI+GG/ *Ase*I+GC from our DNA samples. Once all alleles were binned, a total of 386 loci were available for this study. Genetic variation and population structure of these AFLP profiles were examined using both summary statistics and Bayesian clustering method. For each sampling site, the software *AFLPsurv* (Vekemans 2002), using a uniform prior distribution (Zhivotovsky 1999) of allele frequencies, estimated moderate to strong average population structure ($F_{ST}$) (Table 2.1) among the four sampling sites. The Bayesian clustering method also allocated individuals into populations reflecting the four sites. Additionally, evidence of local population sub-structure, ranging from 3 to 5 clusters, was identified. (Table 2.1). Local *K* populations were estimated by the *R*-package *Geneland* 4.0.4 (Guillot et al. 2005) with 2 000 000 Monte Carlo Markov Chain and 1 000 replications. Details of the field sampling, AFLP genotyping, and genetic variation analyses were previously presented (Chapter 1).

### *General analytical approach*

Our current method is based on multi-scale ordination procedure (Wagner 2004), integrating Canonical Correspondence Analysis (CCA) of direct gradient analysis (ter Braak

1995) and geostatistical variogram analysis (Ripley 1981) to determine whether landscape patterns influenced genetic variation. Multiscale ordination was originally presented by Noy-Meir and Anderson (1971) for block size variance analysis (lattice data), then was further developed by ver Hoef and Glenn-Lewin (1989) for assessing species associations over different geographic ranges and scales. We adapted Wagner's integration of geostatistics (2004) by combining the *R*-package *vegan* (Oksanen et al. 2013) into our analytical approach. Our study extended the application of these integrative procedures with landscape genetics using individual-based grid-based remote sensing landscape determinants and molecular variances. In brief, the analytical approach (Figure 2.2) began with estimating resemblance matrices of pair-wise dissimilarity coefficients for genetic relatedness (**G)** and cost distances of landscape features (**L)** for each field site. We then followed with the constrained ordination CCA for describing the deviance of **G** from expectations under the influence of the concerned **L** landscape pattern. We also performed unconstrained Correspondence Analysis (CA) separately on **G** and **L** to evaluate their spatial patterns. The CCA's standardized outcome matrix (**Q**) was then regressed onto **L** to decompose the total variances into fitted variance matrix ($\mathbf{Q}_{\text{fit}}$) and residual matrix ($\mathbf{Q}_{\text{res}}$). To describe the spatial dependence in the covariances, two separated variogram matrices $\mathbf{Q}^2{}_{\text{fit}}(h)$ and $\mathbf{Q}^2{}_{\text{res}}(h)$ were calculated for the decomposed values, and then their variograms were constructed.

The formation of the response **G** and predictive **L** matrices of dissimilarity coefficients was an essential step in our analytical procedures. There are various methods and terms that are used to describe individual-based pairwise dissimilarity coefficients of genetic variation and landscape features. Bonin et al. (2007) provided a detailed review of measures for genetic dissimilarities in AFLP profiles where the Jaccard index (Jaccard 1908) is arguably most applicable. The AFLP genetic technology only describes a binary, two allele system, presence/absence (1/0). Heterozygotes of so-called dominant genetic markers, such as those produce by AFLP, cannot routinely be identified and thus were subsumed into the presence category. For landscape features, the properties of dissimilarity coefficients between two observation points can be described as landscape connectivity or its inverse landscape resistance (Turner and Gardner 1990, Spear et al. 2010), effective or functional distance (Ferreras 2001), least-cost distance/path (Adriaensen et al. 2003), resistance distance (McRae 2006), or conductance matrix (van Etten and Hijmans 2010). To avoid confusion, in this context we used the terms ‘dissimilarity coefficients’ for the pairwise AFLP genetic relatedness in our Red Junglefowl samples and in the simulated genetic data (see next section) and ‘cost distances’ for the structural translation of the landscape feature connectivity between observations. Both types essentially have semi-metric properties and are particularly applicable to non-Euclidean measurements (Legendre and Anderson 1999) such as those observed in landscape genetics.

### *Dissimilarity coefficients of genetic relatedness*

We calculated the genetic dissimilarity coefficient matrix under the assumption that the AFLP loci are independent, under Hardy-Weinberg equilibrium and IBD does not exist. To evaluate the response, we explored the observed data together with three simulated scenarios. These three simulated sets of spatially organized allele frequencies had similar sampling sizes and sampling arrangements as in the observed data. They were different in their spatial allele frequency distributions: one set had spatially organized allele frequencies, one with random spatially organized local *K* populations (as with the observed data), and one exhibiting a panmictic population. In the first case, the population memberships for individual samples were also inferred from the observed data.

We followed the simulation algorithms of Guillot and Santos (2009) using *Geneland* where all three scenarios had correlated allele frequencies and high genetic differentiation ($F_{ST} = 0.30$) between their local populations. We simulated local populations such that their allele frequencies were sampled from independent Dirichlet distributions in which they exchange genes from a unique and common migrant pool (Excoffier et al. 2009, Guillot and Santos 2009). We calculated $\mathbf{G_{obs}}$ and $\mathbf{G_{sim}}$ matrices of genetic dissimilarities, respectively, from the observed and simulated allele frequencies using the *R*-package *ecodist* (Goslee and Urban 2007), then performed unconstrained ordination CA, regression, and constructed their distance-based variograms (see Multiscale ordination and geostatistics).

### *Cost distances of landscape resistance*

The **L** matrix of pairwise cost distances are the cumulative costs of connectivity between two observations (Adriaensen et al. 2003, Spear et al. 2010). This involved the establishing of friction maps (grid) of transition surfaces and a least-cost algorithm describing the pairwise cumulative cost distances of Red Junglefowl on these surfaces.

We constructed two transition surfaces from two GIS source layers: a radar topography image (90 m, Farr et al. 2007) and an enhanced thematic landcover image (15 m, USGS Landsat 7 ETP Plus). The friction layers were re-rasterized at a pixel resolution of 90 m and re-classified between 1 and 100 in each layer using the *R*-package *raster* (Hijmans 2014). Grid cell values in the transition layers were defined as a continuous gradient. Their source values, ranging from 0 to 2 345 m for elevations and from 1 to 255 for spectral signatures of landcover in South Central Vietnam, were reclassified (no supervision) to numbers between 1 and 100 allowing each layer to have equal importance relative to connectivity (Cushman et al. 2006). The transition surfaces imposed objective assumptions of mobility with 1 for optimum conditions and 100 for most unsuitable habitat in each surface.

We calculated landscape cost distances based on a symmetrical eight-direction movement and least-cost algorithm (Adriaensen et al. 2003) using the *R*-package *gdistance* (van Etten and Hijmans 2010). The eight-direction neighbor allows for movements along the diagonal of the cells and the least-cost calculations create matrices of $\mathbf{L}_{\mathrm{ele}}$ and $\mathbf{L}_{\mathrm{cov}}$, respectively, for elevation and landcover cost distances.

### *Multiscale ordination and geostatistics*

Multiscale ordination CCA of **G** to $\mathbf{L}_{\mathrm{ele}}$ or $\mathbf{L}_{\mathrm{cov}}$ produced the re-scaled **Q** matrix (standardization) that describes the deviances of the genotypic **G** from those expected after partialling out the effects of exogenous **L** landscape factors. This is equivalent to an assumption that the studied genotypes are independent with their landscapes (Borcard et al. 2011, Oksanen et al. 2013). To evaluate this assumption, we performed weighted linear regression of the standardized **Q** to the explanatory **L** in order to partition the total variances in **Q** into explained and residual variances. The explained variances were presented in a matrix of fitted values, $\mathbf{Q}_{\mathrm{fit}}$, describing the influence of landscape features' cost distances on genetic relatedness (or spatial dependence), and the residual variances were in a matrix of residuals, $\mathbf{Q}_{\mathrm{res}}$, reflecting the intrinsic genetic variation (or spatial autocorrelation).

We calculated variogram matrices of $\mathbf{Q}^2(h)$, $\mathbf{Q}^2{}_{\mathrm{fit}}(h)$, and $\mathbf{Q}^2{}_{\mathrm{res}}(h)$ to spatially describe covariances in the total **Q**, the fitted values $\mathbf{Q}_{\mathrm{fit}}$, and the residuals $\mathbf{Q}_{\mathrm{res}}$ as a function of distance $h$. The variograms had equal discrete distance classes of $h$ = 1 000 m reflecting the observed defensed territory sizes of the rooster junglefowl. We also performed significance tests for spatial autocorrelation for the residual values and the total variances for each distance class using the *R*-package *vegan* (Oksanen et al. 2013) with 1 000 Mantel-permutations and $\alpha = 0.05$ (Wagner 2004, Borcard et al. 2011). Following Wagner (2004), we simulated a point-wise 95 percent confident interval envelope for the total variances $\mathbf{Q}^2(h)$ matrix to test the statistical significance of deviation from the null hypothesis.

## RESULTS

### *Sampling arrangements and genetic structure*

We sampled 192 Red Junglefowl in the four study sites in South Central Vietnam, mostly in mature secondary growth forests. Within sites, the average distance between capture localities was 1.12 km (max 26 km, min 0 m of birds in the same flock) depending on capture opportunities. Capture of Red Junglefowl was highest in bamboo-dominated forests where forest floors were clear of understory and seasonal fires provide sprouting food sources and predator-escape clearance. Spatial arrangements of Red Junglefowl had *ad hoc* positive $\hat{K}$ values describing moderate to strong clustered spatial patterns in the four study sites (Figure 2.3 a). The site HBA had particularly strong aggregated sampling due to habitat patchiness and steep topography in the area. The observed overall genetic differentiation $F_{ST}$ among all four study sites was 0.14. Our Bayesian clustering models also detected between 3 and 5 local population clusters in the study sites (Figure 2.3 b and Table 2.1). Within-site genetic differentiation ($F_{ST}$) differed between the lowland and highland sites: 0.06 to 0.07 respectively in the two lowland sites (CTN and LGO) and 0.14 and 0.16 respectively in the two highland (and topographically more heterogeneous) sites (HBA and YDN) (Table 2.1).

### *Spatial autocorrelation by CA models*

The observed differences in genetic differentiation between the lowland and highland sites also revealed evidence of their genetic spatial patterns (Figure 2.3 c). The homogenous lowland sites with weak within-site genetic differentiation also had weak spatial autocorrelation in their proximity distance classes (~ 2 - 3 km) whereas the two highland sites displayed autocorrelated genotypes at close distances and then accrued strong increases in their variances up to 6 km. The highland sites HBA and YDN featured some degree of landscape or sampling heterogeneity: sampling localities in HBA strongly aggregated in three clusters in the foothill of the Annamite and the large YDN site had mosaic habitats of bamboos, deciduous forests, and seasonal forest fires. Beyond 6 km and particularly 8 km, genetic variance appeared to be spatially independent in the all four field sites.

Comparison of the above data set (Figure 2.3 b & c) with the simulation scenario with spatial structure (Figure 2.4 a & d) revealed that genetic differentiation (the only parameter differing between the two sets), in fact, has influences on spatial genetic pattern. Increased genetic differentiation reduced sampling errors and increased autocorrelated variances in the simulated scenario, a situation similar to the previous comparison between spatial genetic patterns in the lowland and highland sites. We observed that total inertia in the simulated scenario did not substantially increase from increasing genetic differentiation (Table 2.1), suggesting that the geostatistical variograms are more sensitive in detecting spatial trends than the unconstrained CA procedure.

Next, comparison of the three simulated patterns showed that the two hypothetical spatial genetic patterns (the random and panmictic scenarios) have greater influences on the spatial variance trends than those from sampling aggregations or genetic differentiation. Spatial variance trends were negligible in the absence of spatial genetic structure in these two scenarios (Figure 2.4 b-c and e-f). Strong aggregated sampling efforts, such as those in HBA, but with either random or panmictic genetic structure, may result in similar no-trend spatial variances. While the three simulated scenarios all had strong large genetic differentiation ($F_{ST}$ = 0.30), this parameter was not a deciding factor for spatial genetic trends as in the above comparison case with the observed spatial genetic patterns.

From three settings above, given our spatially organized genetic data, we concluded that the integrative methods of ordination CA and geostatistical variograms can effectively detect and describe distance-based variances of the Red Junglefowl in the four study sites.

### *Spatial dependence by landscape patterns in CCA models*

Trends of spatial genetic variance in the sampled Red Junglefowl were explained by landscape patterns in two ways. First, the observed variance of the CA-based spatial genetic patterns (previous section) and of the CA-based landscape variations were operating differently within the study spatial ranges. The two landscape patterns of elevation and landcover had variances that were strongly correlated at close distances with no observed sampling errors (i.e., no nugget effect) (Figure 2.5 a-b). They also had relatively large amount of their total inertias compared to those of genetic variances (Table 2.1).

However, the strongly correlated landscape patterns did not explicitly account for the spatial patterns of genetic variances in the four field sites. Our CCA-based variogram models had very low sill values (less than 0.05) describing subtle degrees of spatial dependence (Figure 2.5 c) where the spatial phenomenon seemed to be site-specific. Such instances should be interpreted with caution. Again, the lowland and highland sites were different in their spatial dependence of genetic variation compared to the local landscape patterns. Both CA-based and CCA-based variograms of the CTN and LGO lowlands had similar spatial genetic patterns where we identified weak spatial autocorrelation of genetic variation in the CA-based models (Figure 2.3 c) and weak decreases of residual variances in the CCA-based variograms (Figure 2.5 c) for the first two or three distance classes (~ 2 - 3 km). When removed, these correlated effects in the lowland sites would result in flat variance trends in their spatial dependence analyses. This suggested spatial independence of genetic variation to landscape features in the two lowland sites.

CA-based and CCA-based variograms of the two highland sites were opposite in how their spatial genetic variation was organized. There were correlated genotypes detected in these two sites by our CA models (Figure 2.3 c) but virtually no CCA-based residuals observed (Figure 2.5 c). We ruled out the contributions of genetic differentiation and spatial arrangements to the spatial variance trends here, as these factors were already examined in the three simulated data sets. Of the total inertias recorded in the CCA models, the two landscape features of elevation and landscover patterns accounted for less than 2% of genetic variances in the two highland sites (Table 2.1). Therefore, we concluded that the highland sites' spatial genetic trends appeared to be explained mostly by the underlying genetic structure (local *K* populations) rather than dependence on any landscape feature we analyzed.

In general, the CCA-based spatial dependence analysis here observed that beyond 3 km the genetic residuals in the two lowland parks appeared to be spatially independent as showed in HBA and YDN thorough their full spatial ranges. Weak spatial dependence implied an insignificant role of both elevation and landcover to the extent of observed local genetic variation in all the four study sites. The total variances (of both explained and residual variances) did not exceed the point-wise envelop of the variograms at any distance class and not in any study site (Figure 2.5 c). This suggested the genetic-landscape correlation does not depend on scale and that a regionalized (local) analysis is not necessary.

# DISCUSSION

Many empirical landscape genetic studies use individual-based approaches to quantify autocorrelation in genetic variation and then correlate them to the landscape determinants (e.g. Coulon et al. 2004, Broquet et al. 2006, Cushman et al. 2006). Spatial dependence between landscape and responding variables has been mainly investigated in the fields of landscape ecology and community ecology (e.g. Wagner 2003, Wagner 2004). Our study is the first to address the spatial phenomenon of dependence and autocorrelated residuals in the responding genetic variables hypothesized to be created by landscape features. The integrated analyses of ordination, regression, and geostatistics in our study gave the important finding that elevation and landcover are not responsible for the extent of genetic variation in the ground-dwelling Red Junglefowl.

There are two analytical advantages in our method for spatially non-independent data. First, multiple regressions in CCA and in the subsequent step in our models are powerful techniques in dealing with non-stationary processes (Legendre 1993). The means and variances in **G** and particularly in **L** matrices are dependent on locations due to spatial autocorrelation. Regressions help to de-trend the data by removing the means but not affecting the variances (while often related to the means) and may still depend on locations (Haining 2003). These re-trended and partitioned data, which is achieved by the subsequent regression step, then can be used in any conventional statistical methods where independence assumption may be appropriate. Second, distance-based geostatistical variograms directly deal with both spatial autocorrelation and dependence and are useful in describing the concerned genetic variation as a function of geographic distances. We demonstrated here that variograms are more sensitive and informative in detecting and describing spatial genetic patterns.

### *Analytical sales vs Model structures*

One of the major challenges in landscape genetics research is to characterize spatial heterogeneity in a manner and a scale meaningful to the phenomenon under consideration (McGarigal et al. 2009). Expressing spatial landscape heterogeneity and dependence in genetic variation can be especially problematic as the two processes of ecology and evolution are generally operating at different temporal and spatial scales (Anderson et al. 2010). We applied two different analytical scale and model structure procedures to evaluate our spatially explicit data.

In correcting the effects of spatial autocorrelation in genetic variation, population genetic studies may perform resampling adjustment to finer scales (Storfer et al. 2007) or separating local and regional geographic extents in their models (Frichot et al. 2013). Oftentimes, the assumption of random mating is considered reasonable on local scales and extents. This assumption becomes more questionable as the spatial scale of the study domain increases and IBD will likely occur. Whereas sampling scales and capture opportunities normally cannot be influenced in molecular ecology studies, our three simulated scenarios suggested spatial sampling, at least to the extent of sampling arrangement, does not attribute to the distribution of genetic variation and correlated genotypes. Regarding the separate local extents, in this study we considered each study site separately with 1.12 km average distance between samples. Distance-based variogram analyses confirmed that spatial autocorrelation merely existed in the four local models relative to the ecological extent of the studied Red Junglefowl.

In modeling spatial dependency, consider a regression relationship between the response genetic data ***y*** and the landscape determinants ***x***: $y = \beta_0 + \beta_1 \times x_1 + ... + \beta_n \times x_n + \varepsilon$ where $\varepsilon$ represents the error term. There are two general approaches in dealing with $\varepsilon$ : to quantify it in model residuals (‘soaking-up’ errors) or to incorporate spatial autocorrelation into the model structure (‘filtering’ errors) (Diniz et al. 2009). ‘Soaking up’ spatial dependence or quantifying and separating the correlated residuals by ordination and regression techniques is the main approach in this present study.

Alternatively, we used the ‘filtering’ approach with Bayesian clustering for redefining population membership in the response variables (Chapter 1) (Figure 2.3 d). We ran 1 000 Markov chain Monte Carlo replications of a spatially explicit genetic model to define an average posterior probability of any two individuals occurring in the same cluster. This procedure generated dendrograms of average linkage of unweighted pair group method with arithmetic mean (UPGMA) for the Red Junglefowl. We treated the UPGMA relationships inferred $\mathbf{G_{inf}}$ genetic dissimilarity matrix in depicting the mean posterior probability of common cluster memberships. A probability at 0.00 in UPGMA indicates that two individuals always were placed in the same cluster in the 1 000 replicates. This is equivalent to a genetic dissimilarity coefficient equal to 0.00 in $\mathbf{G_{inf}}$ indicating two junglefowl are identical in their AFLP profiles. Whereas UPGMA and $\mathbf{G_{inf}}$ genetic dissimilarity were equal 1.00 when two junglefowl were never grouped together by the clustering algorithm. This effectively removed the nugget effect in the variogram of $\mathbf{G_{inf}}$ (Figure 2.3 d) implying no sampling variances existed. The total inertia in CA model for $\mathbf{G_{inf}}$ also significantly increased (Table 2.1). In a sense, the procedure 'filtered' spatial autocorrelation in the responding variables and made them more equivalent in correlating to the highly patchy landscape determinants (not reported here).

### *Classification of grid-based transition layers*

The CA-based variograms of the two landscape features did not have sill values when their variance curves increased with distance throughout the sampling extents and there was apparent absence of the upper constant variance values (Figure 2.5 a-b). This indicated highly heterogeneous landscape configurations on continuous surfaces in the four study sites. The determinant landscape variables may intrinsically change rapidly in space (with topography data) or being highly autocorrelated because of recent fragmented impacts (with landscover data). Such spatial heterogeneities are not uncommon in natural systems where patterns of landscape features generally change more rapidly in space than genetic variation. Additionally, we observed that the choice of landscape distance measures, which include the grid-based landscape transition layers and the least-cost distances that are the spatially explicit trajectories drawn on such layers, also influences the proportion of variation in the cost matrices in our study. This did not necessarily affect the significance of the relationship between the responding genetic variables and the cost distance matrix entries. However, thorough evaluations in the topic may be useful (Zapala and Schork 2006). Between the two entities in calculating landscape distance measures in this study, least-cost modeling is well-established (e.g. Adriaensen et al. 2003) in the fields of landscape ecology and landscape genetics. We stressed the application of least-cost method in estimating ‘distance’ and characterizing pairwise cumulative costs or distances of connectivity between two observations. This differed from the method’s original measure in ‘path’ modeling which has more focuses in pattern analysis or population-specific connectivity indices.

Because the least-cost model and its **L** cost distance matrix were estimated from grid-based representations of the landscape, the quality of the source GIS rasters and the classification and customization of them in producing the transition layers, as well as the cumulative movement costs, were required for the reliability and robustness of the method. The underlying movement and the assigned cost values to grid cells are not known on most study species (including Red Junglefowl) in landscape genetic research (Epps et al. 2007). Direct GIS applications in landscape models, such as customization in landscape connectivity model and classification of remote sensing rasterized images, are normally unavailable to most researchers (Etherington 2011). In our study, the cumulative landscape distance costs were inferred from unsupervised classifications of elevation values and spectral signatures of landscover. We determined them based on the observed abundance data on junglefowl in lowland wooded habitats. As such, those values may reflect habitat use of the Red Junglefowl but not necessarily their movement costs. With GIS models, high resolution imageries (e.g. at least 5 m or higher) can be used to improve habitat and landcover classifications (Lo and Choi 2004). This mechanistic interpretation, however, has not always resulted in realistic transition layers for the study species as little is known on the relative permeability of the species in relation to its habitat requirements and movements.

For example, research conducted by Broquet et al. (2006) on the American Marten, a locally common species in wooded habitats similar to Red Junglefowl, found that the spatial resolution of 75 m best represent the balance between least-cost model accuracy and the ecology of the species.

We examined the influence of landscape classification methods on the significance of landscape distance measures and spatial dependence analysis, then, evaluated the performance of our landscape models. The unsupervised classification schemes employed in this study were modified in two different scenarios whereas kept the spatial resolution of the two source GIS rasters fixed at 90 m (Figure 2.6). This generated new transition layers for elevation and for landcover and implicitly converted the landscape configurations from grid-based GIS rasters (fine grains) to pattern-liked landscape metrics (coarse grains). The two modified landscape classification schemes may help to offset the shortfall in our method for not addrressing the classical pattern analyses of landscape configuration (e.g. Turner et al. 2001). The implications of using population-specific versus pair-wise measurements are topics needed for further elaborated in the field of landscape genetics (Balkenhol et al. 2009). Here, the modified landscape transition layers provided different **L** cost matrices that had similar variograms and they did not significantly affect the outcomes of spatial dependency analysis (e.g. Figure 2.5 c). Therefore, we believed that, given our genotypic data and their spatial arrangements, landscape metrics and pattern analyses may not be useful in describing genetic variation and movement of the locally common junglefowl in our field sites.

In conclusion, population genetic structure of Red Junglefowl was not influenced by spatial dependence to the two landscape features of elevation and landcover. Although primarily terrestrial, Red Junglefowl are highly mobile and have the capacity to move over a wide varieties of habitats and landscape terrain. This mobility may be all that is required to create local population structure with negligible presence of IBD effects. Whether the spatial genetic structure of the Red Junglefowl observed in the study was created only by demography and/or breeding system remains an open question. It may be addressed by evaluating the population genetic patterns in other related junglefowl and other pheasant species. Here, this study contributed to the developing discipline of landscape genetics, specifically the utility of distance-based and individual-based methods correlating genetic variation with landscape features. The integrative and spatially explicit methods could be valuable in disentangling the relative effects of the two spatial phenomena of autocorrelation and dependence to provide reliable conclusions about spatial extents of the study species in respect to the influences of population structure and spatial sampling arrangements.

# REFERENCES

Adriaensen, F., J. P. Chardon, G. De Blust, E. Swinnen, S. Villalba, H. Gulinck, and E. Matthysen. 2003. The application of 'least-cost' modelling as a functional landscape model. Landscape and Urban Planning **64**(4):233-247.

Anderson, C. D., B. K. Epperson, M. J. Fortin, R. Holderegger, P. M. A. James, M. S. Rosenberg, K. T. Scribner, and S. Spear. 2010. Considering spatial and temporal scale in landscape-genetic studies of gene flow. Molecular Ecology **19**(17):3565-3575.

Baddeley, A., and R. Turner. 2005. Spatstat: an R package for analyzing spatial point patterns. . Journal of Statistical Software **12**(6):1-42.

Balkenhol, N., L. P. Waits, and R. J. Dezzani. 2009. Statistical approaches in landscape genetics: an evaluation of methods for linking landscape and genetic data. Ecography **32**(5):818-830.

Beebe, W. 1926. A monograph of the Pheasants. Dover Publications, New York.

Bonin, A., D. Ehrich, and S. Manel. 2007. Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. Molecular Ecology **16**(18):3737-3758.

Borcard, D., F. Gillet, and P. Legendre. 2011. Numerical ecology with R. Springer, New York.

Broquet, T., N. Ray, E. Petit, J. M. Fryxell, and F. Burel. 2006. Genetic isolation by distance and landscape connectivity in the American marten (Martes americana). Landscape Ecology **21**(6):877-889.

Coulon, A., J. F. Cosson, J. M. Angibault, B. Cargnelutti, M. Galan, N. Morellet, E. Petit, S. Aulagnier, and A. J. M. Hewison. 2004. Landscape connectivity influences gene flow in a roe deer population inhabiting a fragmented landscape: an individual-based approach. Molecular Ecology **13**(9):2841-2850.

Cushman, S. A., K. S. McKelvey, J. Hayden, and M. K. Schwartz. 2006. Gene flow in complex landscapes: Testing multiple hypotheses with causal modeling. American Naturalist **168**(4):486-499.

Diniz, J. A. F., J. C. Nabout, M. P. D. Telles, T. N. Soares, and T. F. L. V. B. Rangel. 2009. A review of techniques for spatial modeling in geographical, conservation and landscape genetics. Genetics and Molecular Biology **32**(2):203-211.

Dormann, C. F., J. M. McPherson, M. B. Araujo, et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography **30**(5):609-628.

Epps, C. W., J. D. Wehausen, V. C. Bleich, S. G. Torres, and J. S. Brashares. 2007. Optimizing dispersal and corridor models using landscape genetics. Journal of Applied Ecology **44**(4):714-724.

Eriksson, J., G. Larson, U. Gunnarsson, et al. 2008. Identification of the Yellow skin gene reveals a hybrid origin of the domestic chicken. PLOS Genetics **4**(2).

Etherington, T. R. 2011. Python based GIS tools for landscape genetics: visualising genetic relatedness and measuring landscape connectivity. Methods in Ecology and Evolution **2**(1):52-55.

Excoffier, L., T. Hofer, and M. Foll. 2009. Detecting loci under selection in a hierarchically structured population. Heredity **103**(4):285-298.

Farr, T. G., P. A. Rosen, E. Caro, et al. 2007. The shuttle radar topography mission. Reviews of Geophysics **45**(2).

Ferreras, P. 2001. Landscape structure and asymmetrical inter-patch connectivity in a metapopulation of the endangered Iberian lynx. Biological Conservation **100**(1):125-136.

Frantz, A. C., S. Cellina, A. Krier, L. Schley, and T. Burke. 2009. Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? Journal of Applied Ecology **46**(2):493-505.

Frichot, E., S. D. Schoville, G. Bouchard, and O. Francois. 2013. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. Molecular Biology and Evolution **30**(7):1687-1699.

Fuller, R., and P. Garson, editors. 2000. Pheasants. Status Survey and Conservation Action Plan 2000-2004. WPA/BirdLie/SSC Pheasant Specialist Group. IUCN, Gland. Switzerland and Cambrigde, UK and the World Pheasant Association, Reading, UK.

Fumihito, A., T. Miyake, S. I. Sumi, M. Takada, S. Ohno, and N. Kondo. 1994. One subspecies of the Red Junglefowl *Gallus gallus gallus* suffices as the matriarchic ancestor of all domestic breeds. Proceedings of the National Academy of Sciences of the United States of America **91**(26):12505-12509.

Goslee, S. C., and D. L. Urban. 2007. The *ecodist* package for dissimilarity-based analysis of ecological data. Journal of Statistical Software **22**(7):1-19.

Guillot, G., R. Leblois, A. Coulon, and A. C. Frantz. 2009. Statistical methods in spatial genetics. Molecular Ecology **18**(23):4734-4756.

Guillot, G., F. Mortier, and A. Estoup. 2005. *GENELAND*: a computer package for landscape genetics. Molecular Ecology Notes **5**(3):712-715.

Guillot, G., and F. Santos. 2009. A computer program to simulate multilocus genotype data with spatially autocorrelated allele frequencies. Molecular Ecology Resources **9**(4):1112-1120.

Haining, R. P. 2003. Spatial data analysis : theory and practice. Cambridge University Press, Cambridge, UK ; New York.

Hijmans, R. 2014. *raster*: Geographic data analysis and modeling. R package version 2.3-12. . http://CRAN.R-project.org/package=raster.

Jaccard, P. 1908. Nouvelles recherches sur la distribution florale. Bulletin de la Société Vaudoise Des Sciences Naturelles **44**:223-270.

Johnsgard, P. A. 1999. The Pheasants of the world: biology and natural history. Smithsonian Institution Press, Washington, D.C.

Jongman, R., C. Braak, and O. van Tongeren. 1995. Data Analysis in Community and Landscape Ecology. Cambridge University Press.

Legendre, P. 1993. Spatial autocorrelation - Trouble or new paradigm. Ecology **74**(6):1659-1673.

Legendre, P., and M. J. Anderson. 1999. Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. Ecological Monographs **69**(1):1-24.

Lo, C. P., and J. Choi. 2004. A hybrid approach to urban land use/cover mapping using Landsat 7 Enhanced Thematic Mapper Plus (ETM+) images. International Journal of Remote Sensing **25**(14):2687-2700.

Manel, S., M. K. Schwartz, G. Luikart, and P. Taberlet. 2003. Landscape genetics: combining landscape ecology and population genetics. Trends in Ecology & Evolution **18**(4):189-197.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Journal of Cancer Research **27**:209-220.

McGarigal, K., S. Tagil, and S. A. Cushman. 2009. Surface metrics: an alternative to patch metrics for the quantification of landscape structure. Landscape Ecology **24**(3):433-450.

McRae, B. H. 2006. Isolation by resistance. Evolution **60**(8):1551-1561.

Meirmans, P. G. 2012. The trouble with isolation by distance. Molecular Ecology **21**(12):2839-2846.

Noy-Meir, I., and D. J. Anderson. 1971. Multiple pattern analysis, or multiscale ordination: towards a vegetation hologram. Pages 207-232 *in* G. P. Patil, E. C. Pielou, and W. E. Waters, editors. Statistical ecology Volume 3 - Many species populations, ecosystems, and system analysis. Pennsylvania State University Press, University Park.

Oksanen, J., F. Blanchet, Kindt R, et al. 2013. *vegan*: Community Ecology Package. *R* package version 2.0. http://CRAN.R-project.org/package=vegan.

R Development Core Team. 2011. ***R***: A Language and Environment for Statistical Computing. http://www.R-project.org. R Foundation for Statistical Computing.

Ripley, B. 1981. Spatial statistics. Wiley, New York.

Schwartz, M. K., and K. S. McKelvey. 2009. Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. Conservation Genetics **10**(2):441-452.

Sokal, R. 1986. Spatial data analysis and historical processes. *In* Fourth Intl Symposium of Data Analysis and Informatics. North Holland, Versailles, France.

Spear, S. F., N. Balkenhol, M. J. Fortin, B. H. McRae, and K. Scribner. 2010. Use of resistance surfaces for landscape genetic studies: considerations for parameterization and analysis. Molecular Ecology **19**(17):3576-3591.

Storey, A. A., J. S. Athens, D. Bryant, et al. 2012. Investigating the global dispersal of chickens in rrehistory using ancient mitochondrial DNA signatures. PLOS One **7**(7).

Storfer, A., M. A. Murphy, J. S. Evans, et al. 2007. Putting the 'landscape' in landscape genetics. Heredity **98**(3):128-142.

ter Braak, C. 1995. Ordination. Pages 91-163 *in* R. Jongman, C. ter Braak, and O. van Tongeren, editors. Data Analysis in Community and Landscape Ecology. Cambridge University Press.

Turner, M., R. Gardner, and R. O'Neill. 2001. Landscape Ecology in Theory and Practice: Pattern and Process. Springer-Verlag, New York.

Turner, M. G., and R. H. Gardner, editors. 1990. Quantitative methods in landscape ecology the analysis and interpretation of landscape heterogeneity. Springer-Verlag, New York.

van Etten, J., and R. J. Hijmans. 2010. A Geospatial Modelling Approach Integrating Archaeobotany and Genetics to Trace the Origin and Dispersal of Domesticated Plants. PLOS One **5**(8).

Vekemans, X. 2002. AFLP-SURV. Laboratoire de Génétique et Ecologie, Végétale, Université Libre de Bruxelles, Belgium.

ver Hoef, J. M., and D. C. Glenn-Lewin. 1989. Multiscale Ordination: A Method for Detecting Pattern at Several Scales. Vegetatio **82**(1):59-67.

Wagner, H. H. 2003. Spatial covariance in plant communities: Integrating ordination, geostatistics, and variance testing. Ecology **84**(4):1045-1057.

Wagner, H. H. 2004. Direct multi-scale ordination with canonical correspondence analysis. Ecology **85**(2):342-351.

Wagner, H. H., and M. J. Fortin. 2005. Spatial analysis of landscapes: Concepts and statistics. Ecology **86**(8):1975-1987.

Wright, S. 1943. Isolation by Distance. Genetics **28**:114-138.

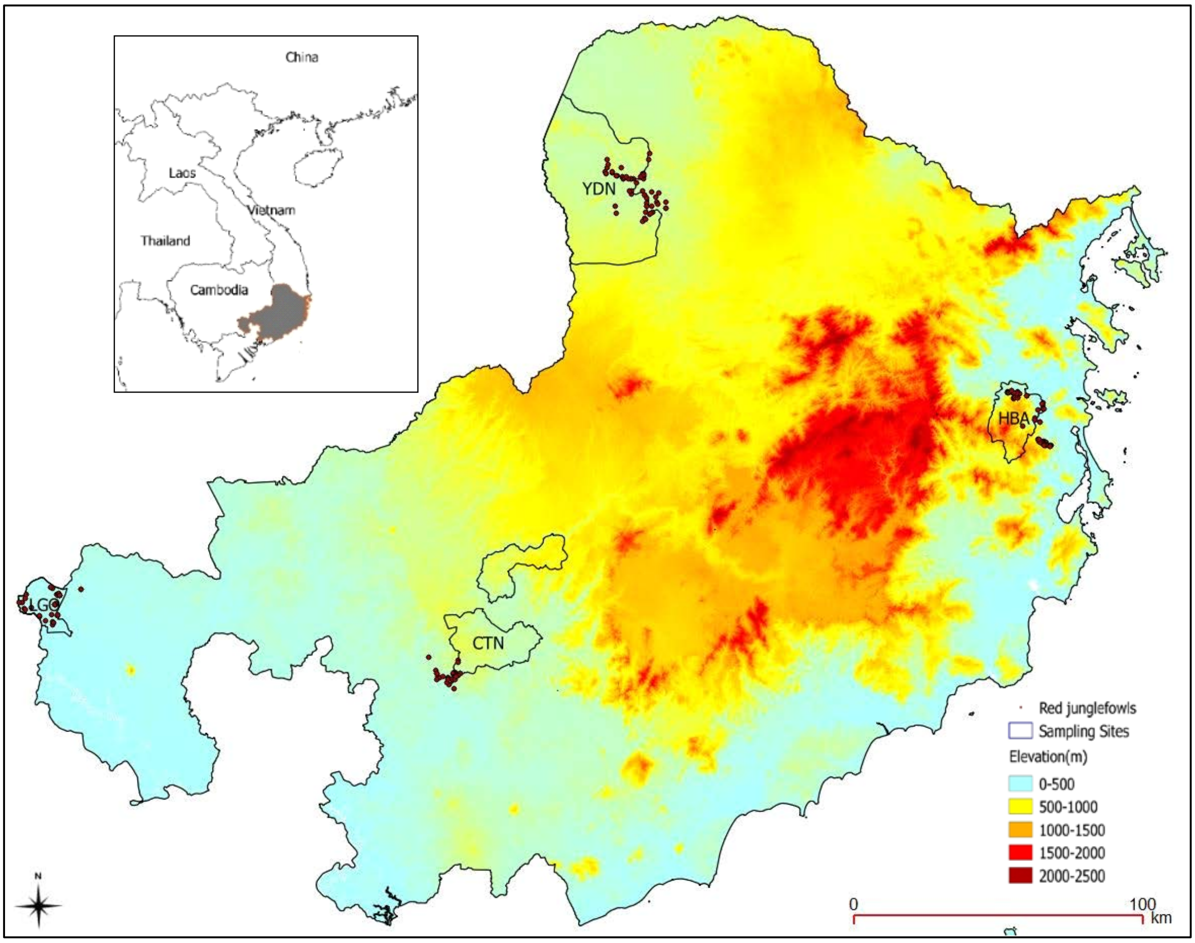
Zapala, M. A., and N. J. Schork. 2006. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proceedings of the National Academy of Sciences of the United States of America **103**(51):19430-19435.

Table 2.1: Local population genetic structure and their variances.

*n - sampling sizes, K - local populations, CA - correspondence analysis, CCA - canonical correspondence analysis.* only for spatially structure scenario.*

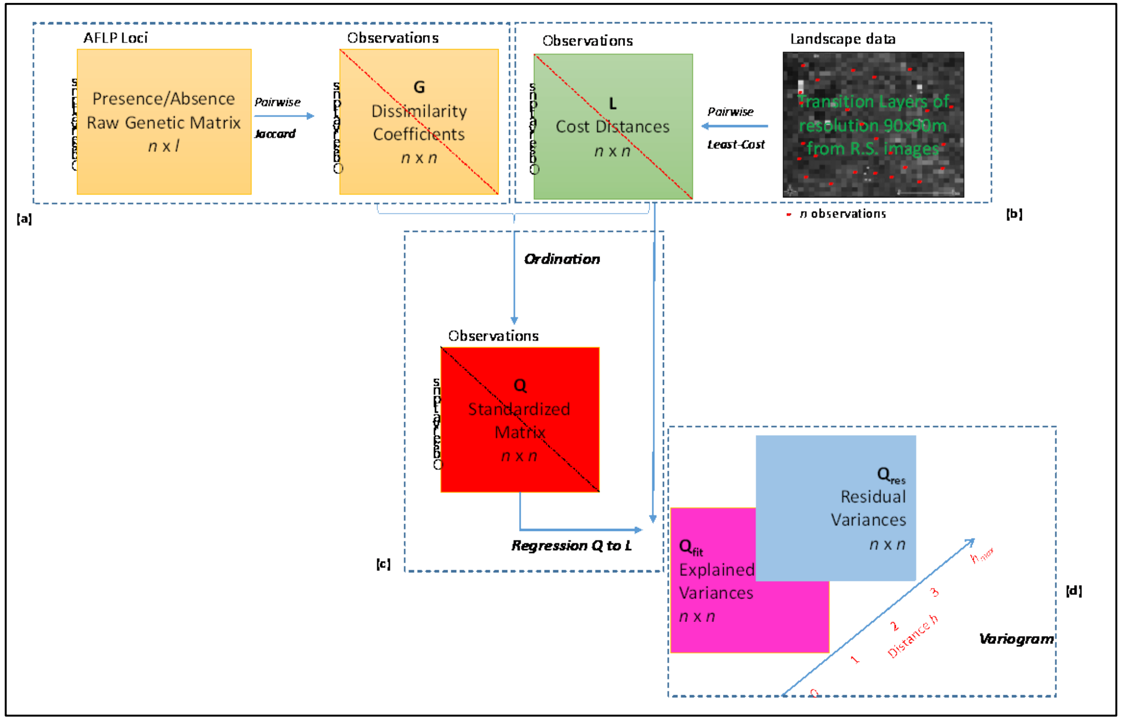
| **Site** | **CTN** | **HBA** | **LGO** | **YDN** |
|---|---|---|---|---|
| *n* | 44 | 56 | 34 | 58 |
| $F_{ST}$ | 0.071 | 0.139 | 0.063 | 0.156 |
| *K* | 3 | 5 | 4 | 4 |
| Variances of CA of observed genetic data | 3.70% | 3.10% | 4.10% | 4.30% |
| Variances of CA of simulated* genetic data | 5.20% | 4.70% | 5.50% | 4.40% |
| Variances of CA of ‘imposed clustering’ genetic data | 37.70% | 28.10% | 53.20% | 24.20% |
| Variances of CA of elevation | 20.40% | 21.50% | 23.60% | 21.70% |
| Variances of CA of landcover | 22.10% | 37.70% | 25.90% | 18.90% |
| Variances of CCA with elevation | 6.84% | 1.10% | 13.60% | 1.38% |
| Variances of CCA with landcover | 5.67% | 1.06% | 15.20% | 1.38% |

Figure 2.1: Sampling sites in South Central Vietnam.



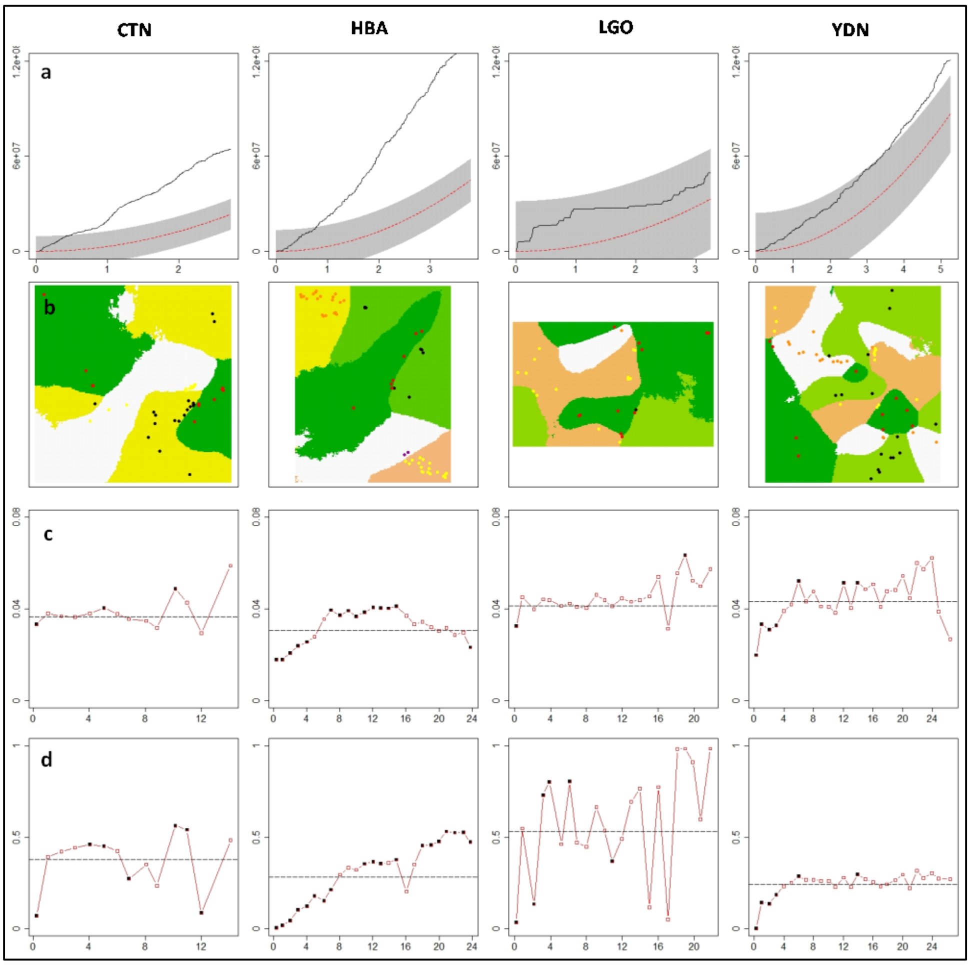
*With the putative Annamite landscape barriers. Sampling sites: Cat Tien National Park and Dong Nai Nature Reserve (here after CTN, as the two sites are connected), Hon Ba Nature Reserve (HBA), Lo Go Sa Mat National Park (LGO), and Yok Don National Park (YDN).*

Figure 2.2: Schematic diagram of Multiscale Ordination framework.



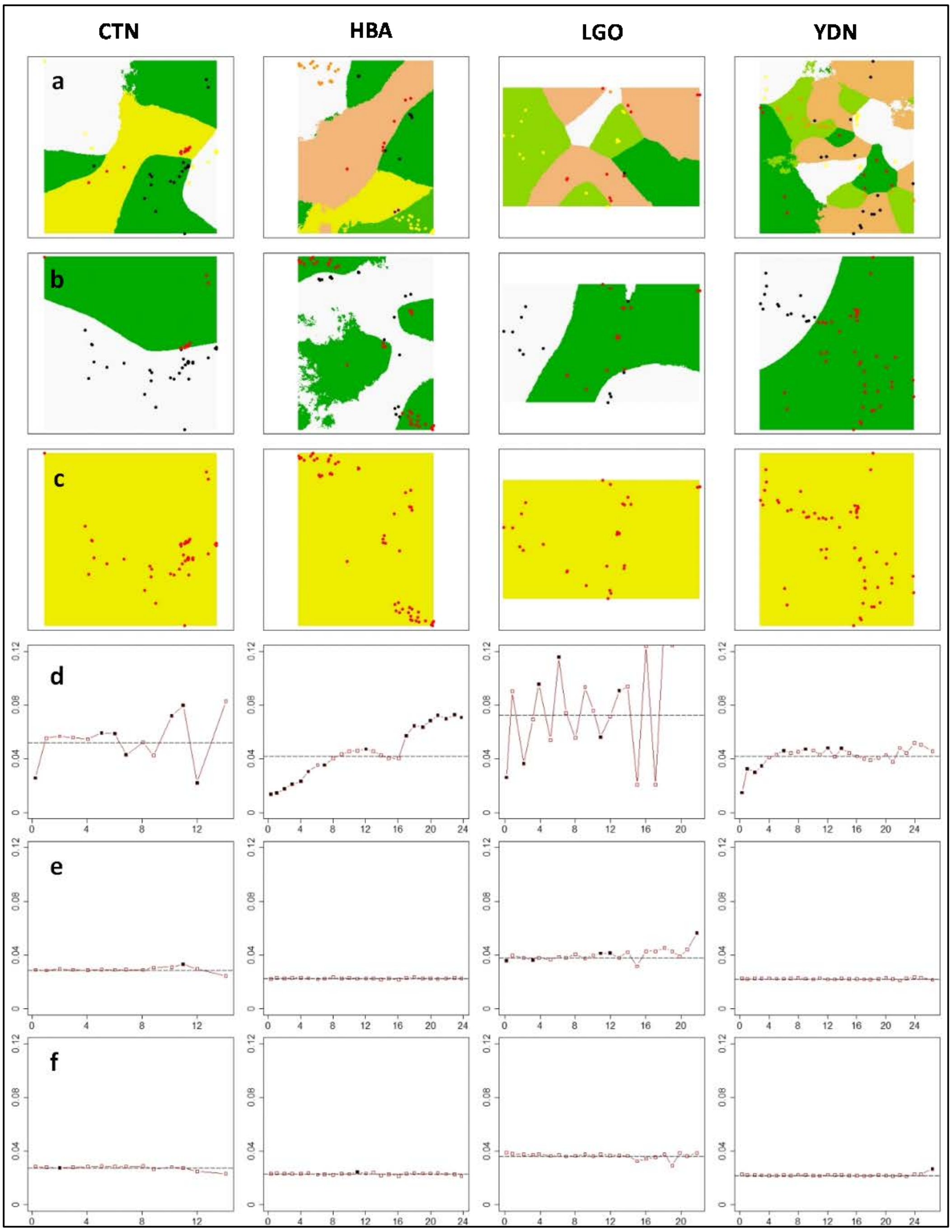
*(a) a dissimilarity coefficient matrix from a genotypic data, (b) a cost distance matrix from least-cost modeling on landscape transition surfaces, (c) ordination methods of CA correspondence analysis and CCA canonical correspondence analysis, and (d) regression and geostatistical variograms.*

Figure 2.3: Spatial structure in the four field sites.



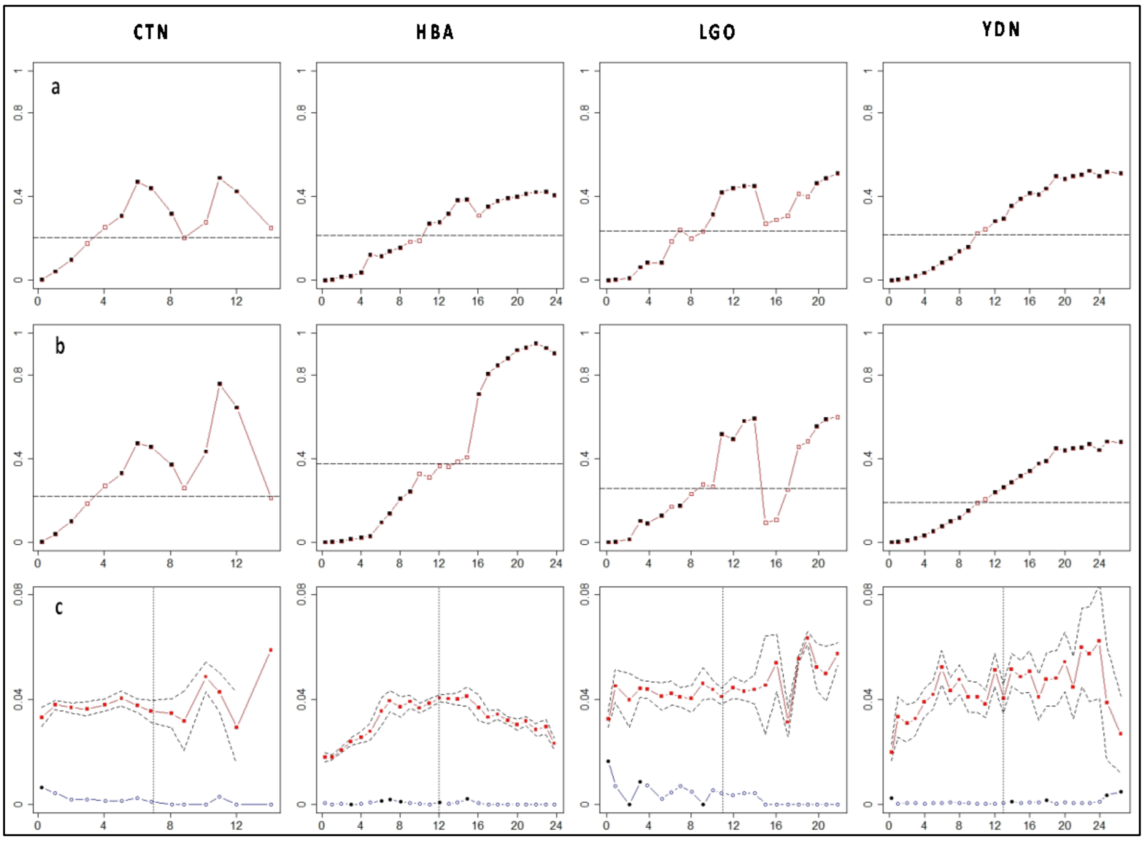
*(a) Observed Ripley's $\hat{K}$ point pattern analysis of sampling arrangement (blue) in relation to a theoretical complete spatial randomness K (red) and its Monte Carlo simulation (shading). $\hat{K} > K$ and falls outside the shading areas indicates a clustered pattern. The K entities were plotted as distance of argument (abscissa, km) vs theoretical Poisson distribution (ordinate). (b) Observed Bayesian posterior genetic cluster memberships. Dots are the study's georeferenced samples. Color regions and dot colors are the posterior probability spatial clusters. Plots represent only one run (out of 1 000) that has highest log posterior density for better visual communication. (c) Observed variograms of CA-based genetic variance trends to distance. (d) Inferred (with 'imposed' spatial autocorrelation) variograms of CA-based genetic variance trends for nugget removing effect. Filled points are significant autocorrelated. Mantel-permutations for the respective distance classes. Distance is in km.*

Figure 2.4: Population structure and variograms of three simulated data sets.



*(a-c) Bayesian posterior genetic cluster memberships for three simulated scenarios of spatially structured (a), random (b), and panmictic (c). (d-f) variograms of CA-based genetic variance trends for three simulated scenarios of spatially structured (d), random (e), and panmictic (f) to distance. Legends are as of Figure 2.3.*

Figure 2.5: Variograms of spatial autocorrelation and dependence.



*(a) CA-based landscape variances of Elevation. (b) CA-based landscape variances of Landscover. (c) Spatial partitioning of CCA-based genetic variances to Elevation (similar results with Landscover, not reported here). Red is total variances (explained and residual variances). Blue is residual variances. Dashed lines are point-wise 95 percent confident interval envelope for the total variances.*

Figure 2.6: Classification schemes for raster images.

A

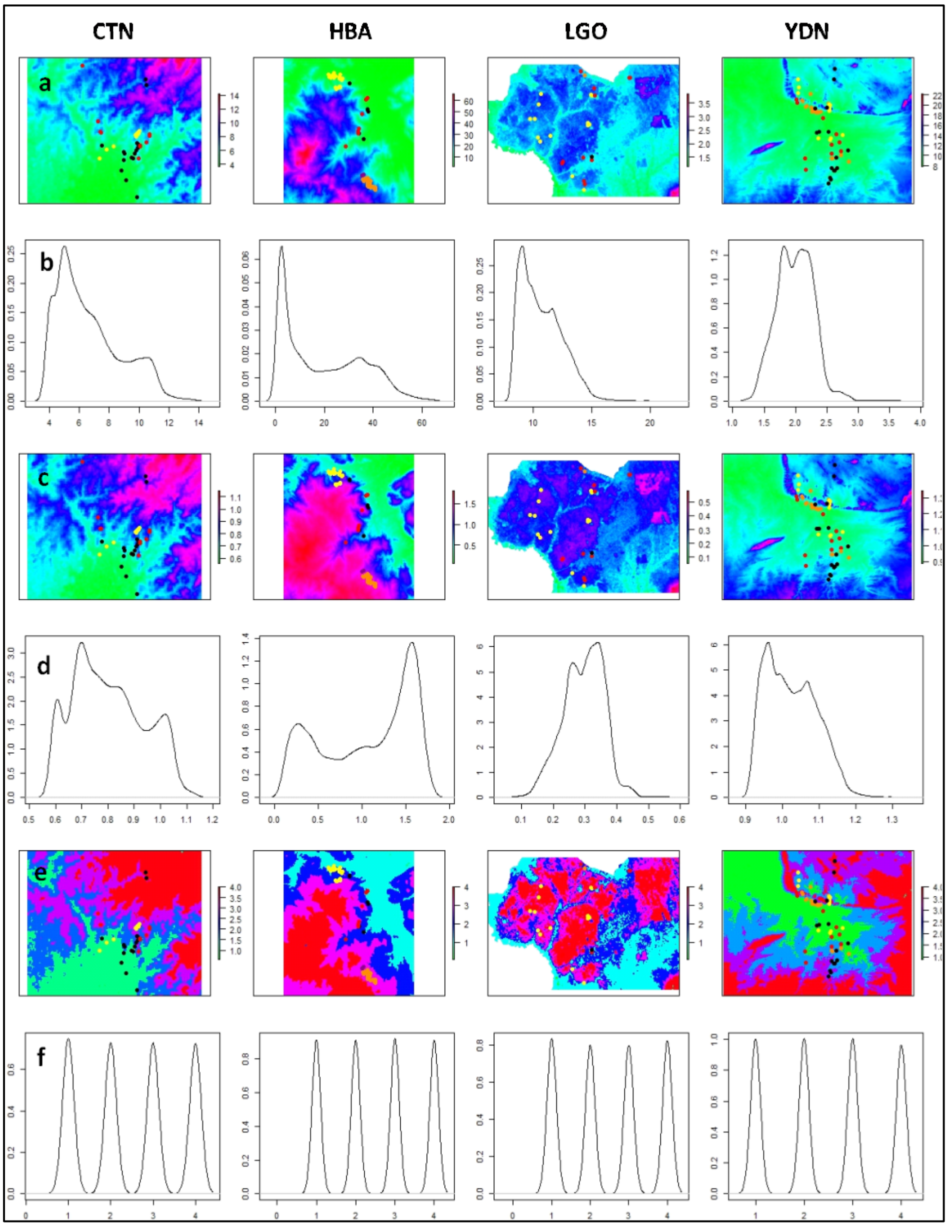
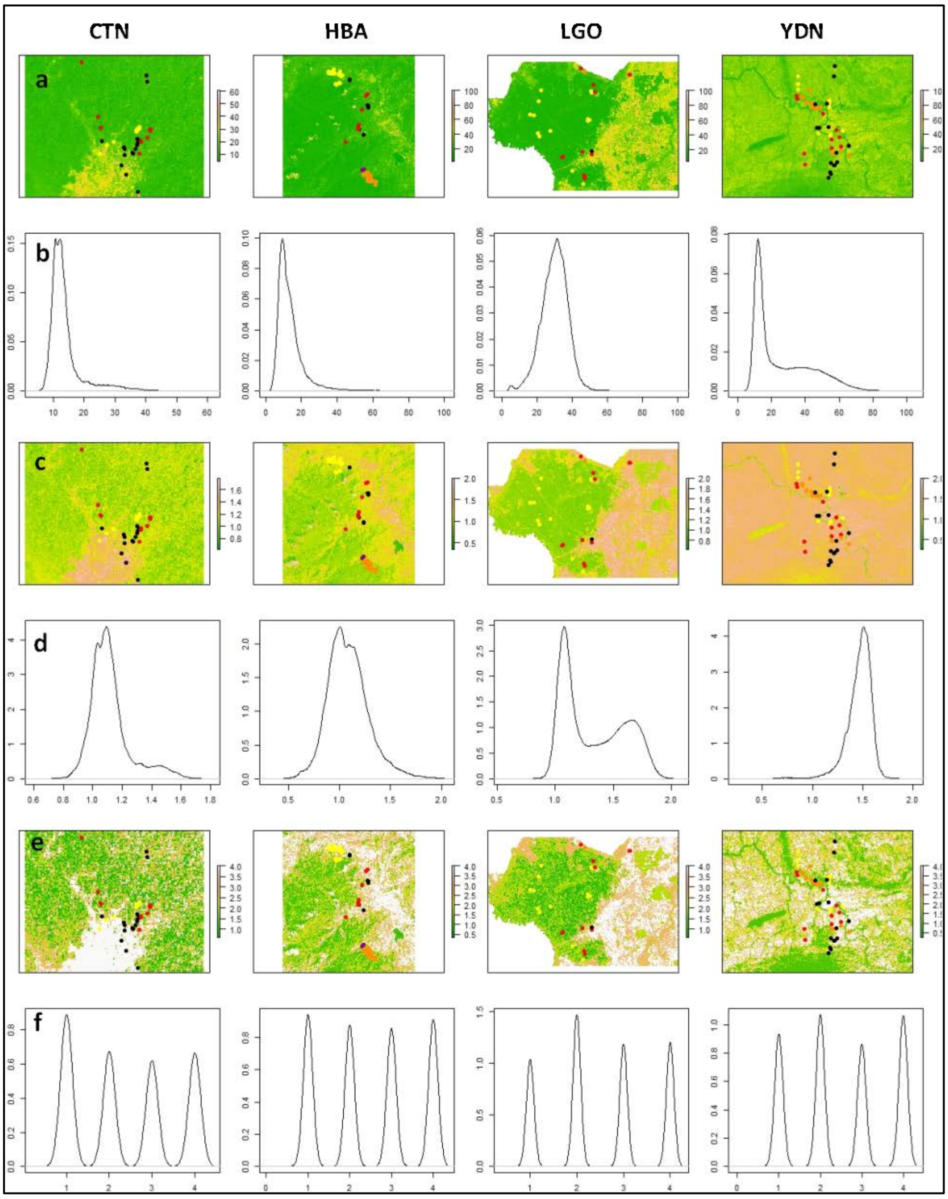


Figure 2.6 (cont.)

B



*(A) Elevation and (B) Landscover raster images. (a) Unsupervised classification scheme 1-100. (b) Cell density plots of (a). (c) Log-transformation of the 1-100 scale (now from 0 to 2). (d)  Cell density plots of (c). (e) Quantile reclassification (now with only 4 categorical values). (f) Cell density plots of (e).*

# CHAPTER 3

## Genetic variation of the Major Histocompatibility Complex (MHC) in wild Red Junglefowl (*Gallus gallus*)

# ABSTRACT

The major histocompatibility complex (MHC) is a multi-family cluster of genes that encodes proteins that modulate immuno-responsiveness. While studies of MHC are relatively common in domesticated poultry, almost nothing is known about this highly polymorphic locus from wild Red Junglefowl (*Gallus Gallus*), the progenitor of domestic chickens. We investigated the diversity of MHC within and among four wild Red Junglefowl populations across diversified natural habitats in South Central Vietnam. Based on an 84 SNP panel spanning nearly 242 Kb of the MHC B-locus, we identified 313 unique haplotypes in 398 chromosomes. None of these haplotypes have been described before and we did not observed any domestic haplotype variants in the wild populations of Red Junglefowl. Analysis of molecular variance (AMOVA) revealed that 94.51% of observed variation was accounted for by within individual diversity. Little genetic variance was apportioned within and among populations, the latter only accounting for 0.83%. We found evidence of recombination, including hotspots, and limited linkage disequilibrium among loci. Compared to domestic chickens, our results suggest extraordinarily high haplotype diversity remains in wild Red Junglefowl and is consistent with a pattern of balancing selection. Wild Red Junglefowl populations in Vietnam, therefore, represent one of the richest resources of natural genomic variation that could directly help to improve agricultural diversity.

**Keywords**: adaptive variation, balancing selection, Major Histocompatibility Complex, MHC class B, Red Junglefowl.

# INTRODUCTION

Our current livestock diversity originated from wild ancestors by altering the genome of these animals through domestication process (i.e. artificial selection) over many thousands of years of human influence. Intensified agricultural activities have recently driven agricultural genetic diversity to a potential crisis: the extent of genetic variation within and among livestock breeds, strains, and lines - which are cornerstones of agricultural diversity - have been precipitously eroded in the past few decades (FAO 2007, Groeneveld et al. 2010). In poultry, considerable genetic diversity has been lost by promoting intensively-selected inbred commercial chicken lines (Muir et al. 2008), replacing indigenous and non-commercial heritage breeds (FAO 2007), and eliminating specialized research breeds in academia and industry (Fulton and Delany 2003, Delany 2006).

Essentially, the current genetic composition in derived domestic livestock contains only a small fraction of the genetic diversity present their wild ancestors, most of which are now either extinct or highly endangered (FAO 2007). This situation is very different from species of crop plants whose ancestors still remain in the wild, often at the centers of their origin, and represent an invaluable source of genetic variation accessible for current and future breeding initiatives (FAO 2007). The capacity of wild progenitors of domesticated animals to maintain genetic diversity and to adapt to changing environments, therefore, remains vital not only for the health of their natural populations but also presents an untapped source of genetic information to maintain and improve current and future agricultural diversity in livestock breeds and lines.

In this respect, Red Junglefowl (*Gallus gallus*) is an interesting species, representing one of the few remaining identifiable ancestors to a domesticated animal line (FAO 2007). Moreover, wild populations of Red Junglefowl still occur naturally in their native ranges (Johnsgard 1999, Brisbin et al. 2002), a sizeable region extending from Southeast to Central Asia. Before spreading globally through human-mediated dispersal (Storey et al. 2012), chicken domestication is believed to have initiated historically in South and Southeast Asia from wild Red Junglefowl, perhaps with inclusion of some other junglefowl lineages (Fumihito et al. 1994, Eriksson et al. 2008). The habitat preference of Red Junglefowl is lowland tropical rainforests in the Asian continent where natural habitats have been greatly modified in recent decades (Fuller and Garson 2000).

For the purposes of this study, it is imperative to emphasize the importance of documenting the sampling origin of Red Junglefowl, particularly in previous studies from which samples were taken for DNA analysis from ‘wild’ breeds. Japp & Hollander (1954) first proposed that the Red Junglefowl be the standard wildtype in chicken genetics. However, in most, if not all, cases of previous genetic studies, the ‘wild’ Southeast Asia junglefowl samples were from zoo birds or other populations of unknown geographic localities (e.g. Fumihito et al. 1994, Granevitze et al. 2007, Berlin et al. 2008, Berthouly et al. 2010, Ngo et al. 2010, Worley et al. 2010, Mekchay et al. 2014). Even the Red Junglefowl female that formed the *Gallus gallus* reference sequence is known to be considerably introgressed with White Leghorn alleles. Essentially, the wild behavior of Red Junglefowl is extreme comparing to other large and threatened species in the same Phasianidae family; it does not tolerate captivity (pers. obs., Collias and Collias 1996, Brisbin et al. 2002, Codon 2012). Thus, any Red Junglefowl obtained from captive populations must have been crossed with domestic lines in order to maintain them. Indeed, only after a 3 or 4 generations of crossing wild male Red Junglefowl to domestic heritage females (e.g. ‘gà tre’’) will offspring survive well and tolerate humans (pers. obs.). Efforts for in-captivity breeding of pure wildtype Red Junglefowl have been attempted, but apparently always end in failure (*pers. obs.; pers. comm*.).

In the current study, we are the first to sample wild Red Junglefowl in geographically diverse habitats in South Central Vietnam. Analyses of the neutral amplified fragment length polymorphisms (AFLP) from the samples showed significant spatial variability wherein distinct population clusters of Red Junglefowl comprise a metapopulation structure (Chapter 1). Fine-scale analysis of genetic information and landscape models suggested that the magnitude of intraspecific population differentiation was considerable but indicated no causative relationship from landcover and elevation but rather demography and movements of the birds (Chapter 2). These observed spatial genetic patterns in the wild Red Junglefowl, however, reflect analyses conditioned on neutral genetic markers to study population structure and an ecological extension of spatially explicit theory (Holderegger et al. 2006).

Neutral variation, by definition, is not subject to selective processes. Nevertheless, genetic diversity is an inherently dynamic process involving not only stochastic but also directional forces where living organisms may experience selective pressures from a variety of exogenous factors and local adaptations (Frankham et al. 2004). Availability of the wild progenitor species of domestic livestock lines enable us to understanding evolutionary genetic processes in wild populations, but also will be essential to evaluate how domestication has altered the ‘progenitor’ genome. In the context of changing climates and emerging infectious diseases, understanding how processes linked to natural and artificial selection affect genes and gene families will create new avenues to understand linkage between immune systems and genetic variation, and the evolution and ecology of the organisms themselves.

The current study examines adaptive genetic diversity in the Major Histocompatibility Complex (MHC) B-locus region in wild Red Junglefowl. The avian MHC B-locus occurs on chromosome 16 and has significantly different content and organization than mammalian MHC but at least aspects of specific genes are functionally equivalent (Guillemot et al. 1989, Kaufman et al. 1999a, 1999b, Afanassieff et al. 2001, Rogers et al. 2003, Hunt et al. 2006). Because of extensive genetic variation that is subject to intense balancing selection, the MHC offers a paradigm of adaptive evolution study at molecular level (Edwards and Hedrick 1998). The most convincing associations between specific MHC haplotypes and pathogen resistance and susceptibility are known from domestic chickens. In chickens, the MHC B haplotypes have strong associations with disease resistance, such as to Marek’s disease (Bacon 1987), and the gene complex has been studied extensively in commercial flocks (e.g. Fulton et al. 2006, Guangxin et al. 2014), as well as in non-commercial breeds (e.g. Izadi et al. 2011), but not in their wild ancestors, Red Junglefowl.

Using an extensive single nucleotide polymorphisms (SNP) panel, this study aims to quantify levels of diversity and variation in MHC B in four natural Red Junglefowl populations in South Central Vietnam. SNPs are one of the most common types of genetic variation in eukaryotic genomes that are increasingly employed in conservation genetics and analysis of population evolutionary history. Considering the extent of the geographical sampling and the range of different habitats, we hypothesize that wild Red Junglefowl possess great diversity in their MHC B haplotypes. Our hypothesis is further conditioned on the fact that commercial selection has undoubtedly reduced allelic diversity even within domesticated chicken lines by 50% (e.g. Fulton et al. 2006, Muir et al. 2008, unpubl. data). In concert with immunological studies, understanding how natural and artificial selection affect MHC will create new avenues to understand linkage between immune system genetic variation, disease resistance, and the evolution and ecology of the organisms themselves. If substantially different MHC variation exists, populations of wild Red Junglefowl will serve as genetic resources in future breeding and conservation programs.

## MATERIALS AND METHODS

### *Field sampling*

Red Junglefowl (*N* = 199) were live-captured by the walking-snare method (Chapter 1) from four protected areas in South Central Vietnam in three dry seasons in 2012, 2013, and 2014. The sampling sites are Cát Tiên National Park and Đồng Nai Reserve (hereafter CTN as the two sites are connected, *n* = 46), Hòn Bà Nature Reserve (HBA, *n* = 56), Lò Gò Sa Mát National Park (LGO, *n* = 39), and Yok Đôn National Park (YDN, *n* = 58) (Figure 1.1). Selection of sites was based on the presence of suitable habitats. The study sites mostly feature natural habitats of lowland tropical rainforest (≤ 600 m in elevation) and they are separated apart from one another (~ 180 km) by residential and non-natural habitats.

Red Junglefowl are medium-sized pheasants (~ 500 - 1 000 g) and mainly ground-dwelling. We sampled 163 roosters, 19 hens, and 17 juvenile chicks (< 3 month old, with rearing female junglefowl). The birds showed key phenotypic and behavioral characteristics of pure, wildtype Red Junglefowl as previously described in their native range in Southeast Asia (Beebe 1926, Delacour 1977). They have slender blackish tarsi in both sexes, longer spur-lengths in male roosters, and complete absence of a comb in adult female hens. Morphologically, the males undergo a summer moult of the neck hackles to an overall dark 'eclipse' plumage following the breeding season (generally June-September, which is also the rainy season) (Brisbin et al. 2002). In their natural habitats, Red Junglefowl are extremely timid and do not tolerate the presence of humans. High densities of Red Junglefowl occur in bamboo-dominated forests where clear understory forest floors and seasonal natural fires provide sprouting food sources and predator-escape clearance for the ground-dwelling birds. The species utilizes a polygynous breeding characterized by female promiscuity. Natal dispersal is the primary mode of gene flow (Johnsgard 1999). Young non-territorial male roosters have a fairly large home range, and often move up to a few kilometers per day (unpubl. data).

### *Collection of genomic DNA and genotyping*

For each bird, 20 - 200 µL of blood was obtained from the brachial vein and stored in a lysis buffer (0.1 M Tris-HCl pH 8.0, 0.01 M EDTA, 4% SDS) (Longmire et al. 2000) until DNA extraction. We used higher SDS concentration to better lyse and preserve blood cells in high temperature field conditions. Genomic DNA was extracted from blood using the

Promega Wizard DNA Isolation kit (Promega Corp., Madison, WI, USA) and was assessed visually with 1% agarose gel electrophoresis to confirm non-degraded, high molecular weight DNA.

We focused on SNPs located in a 241 833 base pair (bp) region of the MHC B-locus (GenBank accession number AB268588). This region was originally sequenced by Shiina et al. (2007) (Figure 3.2) and established the framework for comparative MHC genomics in avian species (e.g. Hosomichi et al. 2006). The targeted SNPs were then adapted to the KASP (Kompetitive Allele Specific PCR) high-throughput genotyping platform (Semagn et al. 2014), to create a panel of 84 SNPs specifically for commercial lines of layers, broilers, and heritage chicken breeds (e.g. Fulton et al. 2014). We applied this SNP panel to 199 Red Junglefowl obtained in Central Vietnam.

The KASP genotyping relies on competitive allele-specific PCR that accurately scores bi-allelic SNPs and InDels (insertions and deletions) at specific loci. For each SNP tested, we chose KASP primers using the following criteria: (1) a SNP must be flanked by at least 50 bp on either side and exhibit sequence characteristics amenable to primer design; (2) the frequency difference between the two genotypes must be ≥ 5; and (3) the read depth must be ≥ 5. For each SNP, two allele-specific forward primers and one common reverse primer were designed. We performed KASP assays in a final reaction volume of 5 µL containing 1*x* KASP reaction mix, 0.07 µL of assay mix (12 µM each allele-specific forward primer and 30 µM reverse primer) and 10 - 20 ng of genomic DNA. Amplifications were performed in thermal-cyclers with cycles of 15 min at 94 °C, 10 touchdown cycles of 20 s at 94 °C and 60 s at 65 - 57 °C (the annealing temperature was reduced 0.8 °C per cycle), 26 - 35 cycles of 20 s at 94 °C, and 60 s at 57 °C. Fluorescence detection of the reactions was performed using an Omega Fluorostar scanner (BMG LABTECH GmbH, Offenburg, Germany). Data were analyzed using the KlusterCaller 1.1 software (KBioscience).

### *Haplotype analyses*

Since chickens are diploid organisms, the determination of a haplotype from a set of genotypic SNPs is not immediately possible. For example, consider two SNPs occurring on the same chromosome, both with alleles A and G. If both SNPs are observed as heterozygotes, it is unclear whether one chromosome contains allele A at both loci and the other chromosome contains allele G in both loci, or whether one chromosome contains allele A at the first locus and allele G at the second locus and the other chromosome contains alleles G and A, respectively. Therefore, in the absence of extended pedigrees (which are rarely available for wild populations) construction of haplotypes from genotypic SNP information requires statistical inference (Browning and Browning 2011).

We used *PHASE* 2.1.1 (Stephens et al. 2001), a program that yields Bayesian estimates of haplotypes and their frequencies from genotypic data under the assumption of random mating, to estimate haplotypes from genotypic SNP data (simultaneously using all the 84 SNPs). During the haplotype reconstruction process, each allele in a SNP genotype is assigned to one or the other parental chromosome, assuming the presence of recombination. Also, it is worth mentioning that while we collected Red Junglefowl samples in other field sites across the South Central region of Vietnam (Chapter 1). However, we restricted MHC haplotype reconstruction analyses to include only the four largest populations due to concerns that the haplotype phasing accuracy decreases markedly with smaller sample sizes (Browning and Browning 2009, 2011).

Also using *PHASE*, we estimated the recombination rate $\rho$ between SNPs across the study MHC region. Here $\rho$ is the factor by which the recombination rate between any two loci exceeds the background recombination parameter $\hat{\rho} = 4Nc$ (Posada 2002, Rokas et al. 2003) itself derived from the MHC SNP data set. Advantageously, *PHASE* generates a posterior probability distribution of the recombination parameter and can be checked for convergence. When estimating recombination rates, we employed the MR model in *PHASE*, which makes explicit allowance for intragenic recombination. Runs consisted of 1 000 iterations as a burn-in, 1 000 secondary iterations, and a thinning interval of 1. The commonly cited value of $r$ = 0.0004 per site (equivalent to 1 recombination event per million bp per generation) was used as the initial starting point. Each dataset was run 10 times with a different starting seed, and checked for convergence by checking consistency among haplotype frequency estimates and the goodness-of-fit measure for each of the 10 runs. The final haplotype assignments were taken from the replicate with the best average goodness-of-fit.

As recommended by Posada & Crandall (2001), we used another estimation procedure to estimate the same population recombination parameter from the MHC sequence data. The presence and significance of inter-site recombination was evaluated with *SequenceLDhot* (Fearnhead 2006). This method used an approximate marginal likelihood method of Fearnhead and Donnelly (2002) to detect recombination hotspots (sites exhibiting levels of recombination much larger than the background rate) from population genetic data. A likelihood ratio (LR) statistic is calculated for each locus to test if a hotspot is present. If the LR exceeds a chosen recombination rate, the SNP or SNPs associated are indicative of a recombination hotspot. A value of 10 corresponds to a false-positive rate of < 1 hotspot in 1.2 Mb. As our coverage region is much smaller (approximately 1/5), we chose a smaller value because the test as originally formulated may be too conservative. A simple plot of the LR statistics can be visualized to assess differences in recombination at different positions along the MHC B-locus.

For measures of polymorphism and neutrality, we employed the program *DNAsp* 5.10.1 (Librado and Rozas 2009) to calculate basic sequence statistics including nucleotide diversity ($\pi$) and Tajima's *D*. The *D* statistic (Tajima 1989) is commonly used to distinguish between a neutrally evolving sequence from one evolving under a non-random process, e.g. directional or balancing selection. The number of pairwise differences between haplotypes (Rohlf 1973) was also computed, and based on a parsimony distance criterion, used to create a minimum spanning tree to depict genetic distances among the haplotypes found in each population (Prim 1957). Hierarchical topologies for the sampled MHC haplotypes were created using the online service "Interactive Tree of Life" (Letunic and Bork 2011).

To capture the strength of linkage disequilibrium (LD) between pairs of MHC SNP markers, we first computed for each SNP locus an exact test of Hardy Weinberg equilibrium (HWE) (Wigginton et al. 2005). Next, we estimated LD between each consecutive SNPs with the pairwise disequilibrium coefficient $D'$ (Lewontin 1964) using *Haploview* 4.2 (Barrett et al. 2005). Although LD is preferably estimated using high-frequency polymorphisms (Reich et al. 2001) we accepted the default parameters to include SNPs showing a minor allele frequency (MAF) of at least 0.05. To define a set of consecutive sites between which there is little or no evidence of recombination - a haplotype block - we used the *D′*-based criteria of (Gabriel et al. 2002), as implemented in *Haploview*, for each Red Junglefowl population separately.

Finally, we estimated how genetic variance was partitioned within and among the constructed MHC haplotypes in the four Red Junglefowl populations using an analysis of variance framework (Weir 1996) for molecular data - AMOVA (Excoffier et al. 1992) implemented in *Arlequin* 3.5 (Excoffier and Lischer 2010) . This technique treats haplotype distances as deviations from an estimate of the group mean, and uses the squared deviations as variances. Significance of the covariance components associated with the three levels of genetic structure was tested in our data (haplotypes within individuals, haplotypes within populations, and haplotypes among populations). *Alerquin* also estimated genetic covariance - a parameter equivalent to genetic differentiation $F_{ST}$ - among the four sampled populations. Statistical significance was tested by permuting individual genotypes among populations 10 000 times.

## RESULTS

### *MHC haplotype variation*

We successfully genotyped 199 individual Red Junglefowl (398 chromosomes) using the KASP 84-SNP platform. The 84 loci were distributed across 241 833 bp region of the MHC B-locus, an average of 1 site per 2.9 Kb (Table 3.1). Of the 84 loci examined, 79 (94%) were polymorphic. Despite this high level of nucleotide variation, none of the loci exhibited more than two alleles (Table 3.1). Haplotypes inferred by *PHASE*, assuming the presence of recombination, met or exceeded 80% for nearly all SNPs.

Stratified by each population, a total of 313 haplotypes were identified (Table 3.2). Three of these (0.96%) were shared between populations: *h*180 was shared between the sites HBA and LGO, *h*178 between HBA and YDN, and *h*241 between LGO and YDN. CTN did not have shared haplotypes with the other field sites. Thus, 310 unique haplotypes distributed among 398 chromosomes (78%) were identified, indicating extraordinarily high haplotype diversity in wild Red Junglefowl. All of the 310 haplotypes found in Red Junglefowl, to the best of our knowledge, have not been reported in domestic chickens, either commercial lines or heritage breeds.

Within a population, haplotype diversity was also considerable. All (100%) of the haplotypes in CTN were unique. Haplotype duplication was minimal in the remaining three population samples, generally with no more than two occurring. Nevertheless, haplotype diversity (Hd) approached 100% in HBA (99%), LGO (98%), and YDN (99%) (Table 3.2). A few haplotypes occurred at higher frequencies in HBA, LGO, and YDN. In these population samples, 8, 6, and 5 occurrences of a specific haplotype were observed, respectively. The highest haplotype frequency recorded occurred in HBA where haplotype *h*147 was found in 8 of 112 chromosomes (7.14%).

Nucleotide diversity $\pi$ in the study MHC region was substantial and consistent, averaging approximately 28% across each population. The same was true for estimates of Tajima's *D* (average = 2.1, $p < 0.05$) showing strong evidence that the MHC B-locus in Red Junglefowl departed from expectations of neutrality (Table 3.2).

### *Estimates of MHC recombination*

The overall background recombination parameter $\hat{\rho}$ in the MHC B-locus in the four sampling populations were relatively low, but consistent across each population: CTN: 0.0065; HBA: 0.0020; LGO: 0.0013; YDN: 0.0024 recombination events per bp per generation (Table 3.2). The average recombination rate between SNPs, however, ranged between 1.13 and 1.34. Moreover, there was evidence that several recombination 'hotspots' ($\rho$ significantly higher than the average and background rates) (Figure 3.3A, 3.4A, 3.5A, 3.6A). In general, the region spanning approximately 180-235 Kb exhibited the lowest estimates of recombination in all four populations. There was strong evidence for variable recombination across the entire MHC B-locus; the recombination rate between SNP pairs was neither consistent across the MHC region nor was it similar between populations.

Individuals sampled from CTN exhibited more hotspots of recombination than the other three sampling sites. The least amount of recombination occurred in the LGO population likely due to the fact that the number of haplotypes identified were nearly half that of the other three populations. This was not a consequence of haplotype diversity since it was similar in magnitude. Rather, the sample size (specifically the number of chromosomes sampled) was reduced considerably because of genotyping failures. Efforts are underway to redo these samples.

Analysis of the MHC SNP dataset with *SequenceLDhat* yielded similar magnitudes and patterns of recombination as estimated by PHASE (not shown) many of which met the requirements of statistical significance for hotspots (Figure 3.7).

### *Estimates of MHC linkage disequilibrium*

The population recombination rate affects the extent of LD (Hill and Robertson 1968) and is an important facet in the evolutionary history of a population. Prior to measuring evidence for pairwise LD between each SNP site across the study MHC B-locus, SNP frequencies at each of the 84 sites were evaluated in terms of HWE expectations and minimum allele frequency (MAF). In both cases, failure to remove sites exhibiting these characteristics will bias significantly the results of any LD analysis. Three instances of deviation from HW expectations were found ($p < 0.001$), two (SNPs 14 and 79) in HBA and one (SNP-65) in YDN (Table 3.1). A limited but variable number of monomorphic loci occurred in each population and along with the three sites exhibiting extreme allele frequencies, removed prior to the LD analysis.

Overall, very little evidence of LD across the 242 Kb MHC region is present in any of the four populations and consistent with the elevated inter-site recombination and hotspots we found by the *PHASE* inferences. However, a number of LD blocks were identified (Figure 3.3B, 3.4B, 3.5B, 3.6B). One LD block comprising SNPs 81 and 82, separated by 1 794 bp, was common in each analysis and exhibited a *D'*= 1 (complete linkage) and logarithm (base 10) of odds of linkage (LOD) > 2 (large LOD scores favor the presence of linkage) in each case. A second LD block linked SNPs 61 and 62 across 2 030 bp, but occurred only in individuals sampled from YDN (Figure 3.6B). In HBA and to a lesser extent in CTN, high values of *D'* persisted but were not strong enough to meet significance requirements because of reduced LOD scores (small LOD scores indicate that linkage is less likely) (Figure 3.3B, 3.4B). The remaining LD blocks included SNPs 21, 22, and 23 in populations CTN, HBA, LGO, and YDN and SNPs 50 and 51 in LGO and YDN.

### *MHC population structure*

Analysis of the 310 MHC haplotypes suggested a diverse distribution where the Red Junglefowl haplotypes appeared to evenly distribute over the four sampling sites (Figure 3.8). The most striking feature of the MHC haplotypes is a lack of spatial organization. The hierarchically arranged network (minimum spanning tree) of haplotypes for individual sampling sites displayed a dispersive pattern: variants of MHC generally showed no clustering arrangements (Figure 3.9, 3.10, 3.11, 3.12). Those that did (Figure 3.11 and 3.12) were likely indicative with close familial relationships (see discussion).

AMOVA indicated that the overwhelming majority of genetic variation was partitioned within individuals (94.51%, $p < 0.001$). In contrast, substantially less genetic variation was attributable within populations (4.66%, $p < 0.001$) or among populations (0.83%, $p < 0.001$) (Table 3.3). In contrast to observations made with neutral loci (Chapter 1), the observed overall $F_{ST}$ among the four study sites was 0.83% ($p < 0.001$), indicating very low levels of genetic differentiation in MHC haplotypes among the four Red Junglefowl populations surveyed.

## DISCUSSION

MHC genes are among the most polymorphic genes in vertebrate animals. As a result of their diversity, MHC molecules have received considerable attention in the fields of evolutionary and conservation biology, and especially in immunogenetics. The study of fitness effects of disease resistance and the costs of adaptive immune responses in avian species has been driven primarily by commercial enterprises seeking to create more disease resistant lines of poultry (e.g. Fulton et al. 2014). Using a high-density SNP detection system, we assessed nucleotide diversity at 84 sites distributed across ~ 240 Kb of the MHC B-locus in 199 wild-caught Red Junglefowl. Genetic samples (blood) from these birds were obtained from four geographically distinct regions in South Central Vietnam, each exhibiting distinct ecological and environmental characteristics. The most striking result of this research is the extensive amount of nucleotide and haplotype variation characterized in the MHC B-locus: nearly 80% of the 398 chromosomes exhibited a unique haplotype.

Originally, we hypothesized that Red Junglefowl were likely to demonstrate extensive MHC diversity. Because they were sampled in ecologically diverse sites, recently disconnected from migration, and perhaps exposed to differing pathogen repertoires, we also predicted a presence of strong population structure similar to that observed with neutral markers (Chapter 1). Although we observed a tremendous amount of MHC haplotype variation, there was also a total lack of spatial organization even within a population. The genetic variation measured as expected heterozygosity was always greater for MHC (between 0.2753 and 0.2894, Table 3.2) than for neutral ALFP markers in the studied Red Junglefowl that previously described (between 0.1243 and 0.1916, Chapter 1).

Two other lines of evidence support the absence of spatial pattern in wild Red Junglefowl. First, genetic differentiation, as measured by $F_{ST}$ of the MHC data among the four study sites, was essentially zero (0.0083). The absence of among-population variation contrasted strongly that that observed with neutral markers (0.1400, Chapter 1). Our analytical methods also identified substantial intraspecific population structure with the neutral AFLP markers (Chapter 1). Second, the genetic covariance attributable to within (4.66%) and among-population (0.83%) levels was only a small fraction of the total within level variance. Note that variance estimation in AMOVA has been derived under several different models and they may have different outcomes in their estimated covariance components. The fixed-population model employed here is considered limited in explaining evolutionary forces causing population differentiation (Weir 1996). Nevertheless, the substantial amount of SNP variation observed at individual level (94.51%) strongly suggests diverse MHC B-locus variation occurs without regard to a geographical context in Red Junglefowl.

### *Agricultural genetic diversity*

The best-studied MHC genes in birds come from those species of agricultural significance such as chicken, pheasant, quail, duck, and turkey (e.g. Kaufman et al. 1999b, Shiina et al. 2004, Shiina et al. 2007, Chaves et al. 2009, Chaves et al. 2011). But unlike their mammalian counterparts, the genomic structure of MHC differs considerably within avian species (Hess and Edwards 2002). This makes direct comparisons more challenging and likely confounded further by significant differences in genotyping technologies (e.g. Jacob et al. 2000, Miller et al. 2004). Our current study has a distinct comparative advantage as we obtained a data set of 17 commercial lines that were genotyped with the same KASP technology at the identical MHC loci (Table 3.4). Without any doubt, the haplotype diversity of the wild-caught Red Junglefowl is increased substantially compared to commercial lines. None of the 310 Red Junglefowl MHC B-locus haplotypes overlaps with any haplotype in the commercial lines and even those of many additional heritage breeds (Fulton, pers. comm.).

Fulton and Delany (2003) brought attention to the continuing rapid decline of both commercial and heritage poultry lines. Recent research evaluating influences of commercial selection practices on the chicken genome has established convincingly a reduction of allelic diversity in chicken lines (Muir et al. 2008). Our current research also confirms this substantial reduction in genetic diversity: selective breeding practices for desirable agricultural traits, including MHC, have reduced significantly the level of genetic variation in commercial lines (Table 3.4). Importantly, the reduction in genetic diversity described by previous research is limited only to domesticated chickens, not their wild ancestors, Red Junglefowl. The unusually high polymorphism of Red Junglefowl MHC discovered in our study will further assist an understanding of adaptive polymorphism and a genetic basis of pathogen resistance (Zuckerkandl and Pauling 1965). The most convincing associations between specific MHC haplotypes and pathogen resistance and susceptibility are already known from domestic chickens. For example, resistance to a virus that causes Marek's Disease reaches 95% if an individual possesses the B-21 MHC haplotype yet is 0% with the B-19 haplotype (Biggs et al. 1968). Incredibly, a portion of one Red Junglefowl haplotype is only one SNP different from the B-21 MHC haplotype that confers 95% resistance to Marek's Disease.

### *Adaptive variation and selection*

The question as to why there is so much MHC B-locus diversity in Red Junglefowl remains unanswered. Extensive variation at MHC loci is generally thought to be maintained by balancing selection (Hess and Edwards 2002), itself modulated by host-parasite coevolution. We estimated Tajima's *D*, a comparison of the average number of nucleotide pairwise differences ($\pi$) and the number of segregating sites (*S*) (Table 3.2). In the absence of demographic changes (e.g. population expansion or contraction, high levels of inter-population migration), positive selection (selective sweeps) is indicated by negative values of *D*. Under the influence of balancing selection, alleles are kept at intermediate frequencies producing positive values of *D* because more pairwise differences exist relative to segregating sites. In all four populations, Tajima's *D* averaged about 2.1 and was statistically significant at $p < 0.05$ (Table 3.2) indicating strong evidence that substitution patterns in Red Junglefowl MHC experience balancing selection.

The exact nature and major driving mechanisms of balancing selection are often debated (Hess and Edwards 2002). In avian species, recent empirical studies support a predominant hypothesis of MHC-dependent mate choice where reproductive selection mechanisms maintain heterozygosity in natural populations (e.g. Von Schantz et al. 1989, Parker 2002, Ekblom et al. 2007). However, a method to distinguish between loci within the MHC region and closely linked loci as the target of mate choice in these studies remains unclear (Tregenza and Wedell 2000). Many other hypotheses have been proposed to account for disease-based selection to the extent of MHC diversity, particularly in model species, as previously mentioned.

The above hypotheses regarding MHC diversity and elevated polymorphism are not mutually exclusive. Given the focus of our study, we could only conclude that recombination does play a role in MHC haplotype diversity in the wild Red Junglefowl. Whether this is a historical vestige or an ongoing contemporary process cannot be determined with the samples we have currently. While long-term effects of balancing selection and evolutionary forces acting to maintain MHC diversity are not in doubt, understanding contemporary forces to MHC genes is far more interesting in Red Junglefowl, as at some sampling sites, there was a close association with domestic chickens. Domestication of the chicken is thought to have occurred in Southeast Asia, probably present-day Vietnam or South China or both (Berthouly et al. 2010, Miao et al. 2013). Mitochondria-based molecular evidence showed that the process of poultry domestication, like all other livestock species, occurred in several places and probably deployed several divergent lineages of wild ancestors (Nishibori et al. 2005, Eriksson et al. 2008). Subsequently, wild alleles of genes in natural populations of the original progenitors may have been replaced through intensive cross-breeding with domestic stocks (Peterson and

Brisbin 1998, Berthouly et al. 2010). The ubiquity of human populations and widespread occurrence of free-ranging chickens in Southeast Asia are raising fears of introgression between wild Red Junglefowl and domestic chickens.

Our findings indicate that none of analyzed Red Junglefowl haplotypes overlap with any haplotype in the commercial lines. Moreover, a small sample of local (heritage) breeds had no evidence of similar MHC haplotypes (data not shown). Together, these findings suggest allelic introgression between wild and domesticized fowls has not occurred. However, MHC haplotypes may not be the best sequence to use for these comparisons. Unique MHC haplotypes from Red Junglefowl in this study were at least affected to some degree of recombination sufficient enough to reduce the amount of LD. It is commonly believed that the two evolutionary forces attribute to distinct patterns in the MHC genes in model avian species (Edwards et al. 1995). However, the potential for recombination analysis and LD mapping in natural populations is dependent on how genetic diversity across the genome is structured in populations, about which there is so far almost no knowledge for the great majority of wild species. With balancing selection, new MHC variants are favorable and can persist over long periods in natural populations, and that could explain the observed levels of MHC haplotype variation in our Red Junglefowl samples. Particularly under circumstances where the pathogen repertoire is diverse or changes rapidly, multiple MHC loci with highly polymorphic molecules are necessary for presentation of exogenous pathogen-derived peptides to effector T cells.

### *MHC recombinants*

In contrast to mammals, the close relationship between avian MHC haplotypes and role in disease resistance may in part be due to its more compact system and general absence of non-immune related genes (Kaufman et al. 1999b). From a population genetic point of view, recombination produces new haplotypes and increases the genetic variation in a population, by breaking up existing linkage between gene loci. But the relatively small size of the chicken MHC (at least compared to MHC in mammals) may restrict recombination between different loci. For example, recombinant individuals between the MHC B–L and B–F loci are extremely rare in commercial chicken lines (Hala et al. 1989).

Our investigation of recombination in Red Junglefowl MHC suggested that while recombination within the MHC B-locus does occur, its magnitude is highly variable across both the entire MHC locus and between populations. While uncommon, at least two hotspots of recombination were identified in each of the four populations. However, despite the rarity of MHC recombination observed in domestic chickens, recombination in natural populations of Red Junglefowl populations likely contributes to increased diversity in MHC loci.

The population recombination rate also affects the extent of LD (Hill and Robertson 1968). Unfortunately, studies of MHC LD in natural populations of birds remain rare. A recent study by Edwards and Dillon (2004) examined a 40 Kb region of MHC Class II locus in Red-winged Blackbirds (*Agelaius phoeniceus*). They found evidence of high LD only across a few hundred base pairs, nearly identical to what our current study discovered. In contrast, Heifetz et al. (2005) found significant LD extending over several centimorgans in commercial populations of domestic chicken but these results may have limited relevance to comparisons with wild Red Junglefowl populations.

The expected magnitude of LD between different alleles is dependent on the age of the original mutation and the rate of recombination between loci (Stumpf and McVean 2003). Based on studies of the domestic chicken, it has been suggested that birds have high-recombination rates exclusive of MHC regions (e.g. Z-Chromosome) (Hillier et al. 2004), a phenomenon that would prevent extensive LD from occurring. The current study demonstrated that very little LD exists in MHC in Red Junglefowl. Remarkably consistent across the four sampled populations, the occurrence of high LD was restricted primarily to regions less than 1 Kb although a roughly 2 Kb region was also identified. These findings mirror those found by Edwards and Dillon (2004) and also in MHC of the wild turkeys (*Meleagris gallopavo*) (Chaves et al. 2011).

LD blocks are expected to depend on the sample populations. Generally, as effective population size ($N_e$) increases, the smaller the blocks will be. The similar magnitude and distribution of small LD blocks across four widely separated populations of Red Junglefowl suggests that not only is $N_e$ large in these populations, but also similar in size. Consistent with this interpretation are similar rates of recombination across these regions. Blocks of LD can arise by chance even when recombination rates are uniform. If the recombination hotspots identified in the Red Junglefowl populations are real, then at least some aspects of the LD blocks is transferable between populations, an event that is unlikely to occur at the large geographic scales present in my study. Importantly, this may suggest that some aspect of the environment (e.g. pathogens) is selecting for these non-random association of alleles. However, demographic transfer (e.g. dispersal) of LD blocks could certainly occur at smaller scales like the scale of my field sampling. However, if recombination rates are in fact low, then LD blocks should be more reflective of historical recombination events since only very old recombination events can result in LD block boundaries that are shared between isolated populations.

### *Sampling and haplotype diversity analyses*

Red Junglefowl have a polygynous breeding system and males are highly territorial during the breeding season. Our field method of walking snares with decoy roosters mainly lured and captured the dominant roosters of the flocks along our sampling localities (here, 'flock' in a broad sense refers to our observed social unit). The average geographic distance between the samples was 1.09 km depending on capture opportunities (min 0 m of birds in the same flock or same family, max from 14 to 26 km within a collection area). As a consequence, in most cases, we did not have knowledge about family structure that we could compare to our haplotype networks, except in the three following circumstances.

In LGO, we (presumptively) knew the relationships of two families that were sampled less than 300 m apart. Each family had a rearing hen and two young chicks. Of the six birds, we were able to match the each family member to a specific haplotype: haplotype cluster of *h*232 (hen) - *h*233 - *h*234 (chick) - *h*235 (chick) belong to one family and cluster of *h*255 (hen) - *h*254 (chick) belong to another family (insert, Figure 3.11). Haplotypes of the remaining chick in the second family were not grouped in the cluster, likely due to inheritance from different males, as multiple sire matings do occur regularly in Red Junglefowl and other avian species (Collias and Collias 1996). Alternatively, a recombination event could also have taken place. In YDN, five MHC variants of *h*289 - *h*292 - *h*277 - *h*286 - *h*279 belong to five male Red Junglefowl sampled in the same vicinity (insert, Figure 3.12). Average distance between these five birds was just more than a hundred meters whereas *h*277 and *h*279 belong to two first-year non-territorial male roosters captured together.

We previously concluded in our landscape models that the spatial neutral genetic variation in Red Junglefowl is driven more by species movement and demography than by spatial dependence on landscape features or sampling arrangements (Chapter 2). It is useful to re-evaluate the topic considering adaptive MHC variation, specifically in site CTN. Even though the sampling arrangement in CTN was not randomly distributed *per se* (Chapter 2), the Red Junglefowl inhabiting the area appeared to evenly distribute within habitat types. The park is very well protected, with no major within-site landscape barriers, compared to the other three sites. HBA is located in the foothills of the Annamite Mountain Range and the spatial sampling scheme there was strongly aggregated and influenced by habitat patchiness and steep topography. LGO is fairly small in area and features the most disturbed lowland habitats, although population densities of Red Junglefowl still remain high. YDN has deciduous forests with a major river bisecting the park, and seasonal forest fires are common. Sampling conditions, therefore, could attribute the amount of MHC diversity in CTN and quality of the sampling habitats but this topic remains an open question and needs further investigation.

Altogether, caution should be taken in interpreting the network of the MHC haplotypes in the study. With balancing selection acting on MHC diversity, new MHC variants will arise and persist but need sufficient time for recombination (or mutation) to separate markers. These forces may act at very localized physical scales (even within a single exon). From our three topological ‘family’ clusters in LGO and YDN and from the observed diversity in CTN MHC haplotypes mentioned before, we recommend the interpretation of Red Junglefowl MHC haplotype networks be used only for a determination of similarity analysis haplotype *per se*, rather than inferring any evolutionary relatedness among the birds.

### *Spatial process of adaptive genetic variation*

In a geographical context, the diversity of MHC genes is also expected to be under the influences of spatial processes and local environmental factors. In the previous study using neutral AFLP markers, we identified a metapopulation structure in Red Junglefowl wherein the birds displayed high degree of intra- and inter-specific-population patterns. Evidence of fine-scale genetic subdivision was detectable at distances as low as 5 km (Chapter 1). Spatial analyses of neutral variation in Red Junglefowl also ruled out long-distance gene flow (stepping-stone model). Instead, a metapopulation structure may have resulted from fragmentation of a formerly panmictic population. The spatial patterns of MHC variability are very different to this metapopulation neutrality structure, as described throughout in this manuscript. This has several implications in the evolution history of Red Junglefowl, as well as future management and breeding programs. First, balancing selection, perhaps modulated in part by extensive recombination, has facilitated the generation and retention of MHC polymorphism in Red Junglefowl, perhaps to an extent that counters the effects of genetic drift and gene flow that shape neutral genetic variation. Adaptively derived diversity present in the wild Red Junglefowl may be indicative of balancing selection and narrower tolerance to the underlying environments. However, our analysis so far has not been able to confirm this hypothesis, except to show that certain linkage blocks are present in some, but not all populations. Thus, neutral processes are not sufficient to explain completely spatial genetic variation in wild Red Junglefowl and currently there are few ecologically meaningful genes, except MHC, that are well-enough understood for these kind of studies. We conclude that the analysis of adaptive MHC variation in Red Junglefowl provides insights into one of the richest resources of natural genomic variation that could directly help to improve agricultural diversity.

# REFERENCES

Afanassieff, M., R. M. Goto, J. Ha, M. A. Sherman, L. W. Zhong, C. Auffray, F. Coudert, R. Zoorob, and M. M. Miller. 2001. At least one class I gene in restriction fragment pattern-Y (Rfp-Y), the second MHC gene cluster in the chicken, is transcribed, polymorphic, and shows divergent specialization in antigen binding region. Journal of Immunology **166**(5):3324-3333.

Bacon, L. D. 1987. Influence of the Major Histocompatability Complex on disease resistance and productivity. Poultry Science **66**(5):802-811.

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics **21**(2):263-265.

Beebe, W. 1926. A monograph of the Pheasants. Dover Publications, New York.

Berlin, S., L. J. Qu, X. Y. Li, N. Yang, and H. Ellegren. 2008. Positive diversifying selection in avian *Mx* genes. Immunogenetics **60**(11):689-697.

Berthouly, C., X. Rognon, T. N. Van, et al. 2010. Vietnamese chickens: a gate towards Asian genetic diversity. BMC Genetics **11**(53).

Biggs, P. M., R. J. Thorpe, and L. N. Payne. 1968. Studies on genetic resistance to mareks disease in domestic chicken. British Poultry Science **9**(1):37.

Brisbin, I. L., Jr., A. T. Peterson, R. Okimoto, and G. Amato. 2002. Characterization of the genetic status of populations of Red Junglefowl. Journal of the Bombay Natural History Society **99**(2):217-223.

Browning, B. L., and S. R. Browning. 2009. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. American Journal of Human Genetics **84**(2):210-223.

Browning, S. R., and B. L. Browning. 2011. Haplotype phasing: existing methods and new developments. Nature Reviews Genetics **12**(10):703-714.

Chaves, J. A., A. Ramis, R. Valle, A. Darji, and N. Majo. 2009. Avian influenza specific receptors expressed in the respiratory and gastrointestinal system from chickens, turkeys, ostriches, patridge, ducks and quail. Journal of Comparative Pathology **141**(4):277-277.

Chaves, L. D., G. M. Faile, J. A. Hendrickson, K. E. Mock, and K. M. Reed. 2011. A locus-wide approach to assessing variation in the avian MHC: the B-locus of the wild turkey. Heredity **107**(1):40-49.

Codon, T. 2012. Morphological detection of genetic introgression in Red Junglefowl (*Gallus gallus*). Master Thesis. Georgia Southern University, Statesboro, Georgia.

Collias, N. E., and E. C. Collias. 1996. Social organization of a Red Junglefowl, *Gallus gallus*, population related to evolution theory. Animal Behaviour **51**:1337-1354.

Delacour, J. 1977. The Pheasants of the World. 2nd edn. Spur, Hindhead, UK.

Delany, M. E. 2006. Avian genetic stocks: The high and low points from an academia researcher. Poultry Science **85**(2):223-226.

Edwards, S. V., and M. Dillon. 2004. Hitchhiking and recombination in birds: evidence from Mhc-linked and unlinked loci in Red-winged Blackbirds (Agelaius phoeniceus). Genetical Research **84**(3):175-192.

Edwards, S. V., and P. W. Hedrick. 1998. Evolution and ecology of MHC molecules: from genomics to sexual selection. Trends in Ecology & Evolution **13**(8):305-311.

Edwards, S. V., E. K. Wakeland, and W. K. Potts. 1995. Contrasting histories of avian and mammalian Mhc genes revealed by class II B sequences from songbirds. Proceedings of the National Academy of Sciences of the United States of America **92**(26):12200-12204.

Ekblom, R., S. A. Saether, P. Jacobsson, P. Fiske, T. Sahlman, M. Grahn, J. A. Kalas, and J. Hoglund. 2007. Spatial pattern of MHC class II variation in the great snipe (Gallinago media). Molecular Ecology **16**(7):1439-1451.

Eriksson, J., G. Larson, U. Gunnarsson, et al. 2008. Identification of the Yellow skin gene reveals a hybrid origin of the domestic chicken. PLOS Genetics **4**(2).

Excoffier, L., and H. E. L. Lischer. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources **10**(3):564-567.

Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of Molecular Variance Inferred from Metric Distances among DNA Haplotypes - Application to Human Mitochondrial-DNA Restriction Data. Genetics **131**(2):479-491.

FAO. 2007. The State of the World's Animal Genetic Resources for Food and Agriculture. Food and Agriculture Organisation of the United Nations, Rome.

Fearnhead, P. 2006. SequenceLDhot: detecting recombination hotspots. Bioinformatics **22**(24):3061-3066.

Fearnhead, P., and P. Donnelly. 2002. Approximate likelihood methods for estimating local recombination rates. Journal of the Royal Statistical Society Series B-Statistical Methodology **64**:657-680.

Frankham, R., J. D. Ballou, and D. A. Briscoe. 2004. A primer of conservation genetics. Cambridge University Press, Cambridge, UK; New York.

Fuller, R., and P. Garson, editors. 2000. Pheasants. Status Survey and Conservation Action Plan 2000-2004. WPA/BirdLie/SSC Pheasant Specialist Group. IUCN, Gland. Switzerland and Cambrigde, UK and the World Pheasant Association, Reading, UK.

Fulton, J. E., J. Arango, R. A. Ali, E. B. Bohorquez, A. R. Lund, C. M. Ashwell, P. Settar, N. P. O'Sullivan, and M. D. Koci. 2014. Genetic variation within the *Mx* Gene of commercially selected chicken lines reveals multiple Haplotypes, recombination and a protein under selection pressure. PLOS One **9**(9):14.

Fulton, J. E., and M. E. Delany. 2003. Poultry genetic resources - Operation rescue needed. Science **300**(5626):1667-1668.

Fulton, J. E., H. R. Juul-Madsen, C. M. Ashwell, A. M. McCarron, J. A. Arthur, N. P. O'Sullivan, and R. L. Taylor. 2006. Molecular genotype identification of the *Gallus gallus* Major Histocompatibility Complex. Immunogenetics **58**(5-6):407-421.

Fumihito, A., T. Miyake, S. I. Sumi, M. Takada, S. Ohno, and N. Kondo. 1994. One subspecies of the Red Junglefowl *Gallus gallus gallus* suffices as the matriarchic ancestor of all domestic breeds. Proceedings of the National Academy of Sciences of the United States of America **91**(26):12505-12509.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, et al. 2002. The structure of haplotype blocks in the human genome. Science **296**(5576):2225-2229.

Granevitze, Z., J. Hillel, G. H. Chen, N. T. K. Cuc, M. Feldman, H. Eding, and S. Weigend. 2007. Genetic diversity within chicken populations from different continents and management histories. Animal Genetics **38**(6):576-583.

Groeneveld, L. F., J. A. Lenstra, H. Eding, et al. 2010. Genetic diversity in farm animals - a review. Animal Genetics **41**:6-31.

Guangxin, E., R. N. Sha, S. C. Zeng, C. Wang, J. F. Pan, and J. L. Han. 2014. Genetic variability, evidence of potential recombinational event and selection of LEI0258 in chicken. Gene **537**(1):126-131.

Guillemot, F., J. F. Kaufman, K. Skjoedt, and C. Auffray. 1989. The Major Histocompatibility Complex in the chicken. Trends in Genetics **5**(9):300-304.

Hala, K., R. Sgonc, C. Auffray, and G. Wick. 1989. Typing of MHC haplotypes in os chicken by means of RLFP analysis. Pages 177-186  Bhogal, B. S. And G. Koch.

Heifetz, E. M., J. E. Fulton, N. O'Sullivan, H. Zhao, J. C. M. Dekkers, and M. Soller. 2005. Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. Genetics **171**(3):1173-1181.

Hess, C. M., and S. V. Edwards. 2002. The evolution of the major histocompatibility complex in birds. Bioscience **52**(5):423-431.

Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics **38**(6):226-231.

Hillier, L. W., W. Miller, E. Birney, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432**(7018):695-716.

Holderegger, R., U. Kamm, and F. Gugerli. 2006. Adaptive vs. neutral genetic diversity: implications for landscape genetics. Landscape Ecology **21**(6):797-807.

Hosomichi, K., T. Shiina, S. Suzuki, et al. 2006. The major histocompatibility complex (Mhc) class IIB region has greater genomic structural flexibility and diversity in the quail than the chicken. Bmc Genomics **7**(1):322.

Hunt, H. D., R. M. Goto, D. N. Foster, L. D. Bacon, and M. M. Miller. 2006. At least one YMHCI molecule in the chicken is alloimmunogenic and dynamically expressed on spleen cells during development. Immunogenetics **58**(4):297-307.

Izadi, F., C. Ritland, and K. M. Cheng. 2011. Genetic diversity of the major histocompatibility complex region in commercial and noncommercial chicken flocks using the LEI0258 microsatellite marker. Poultry Science **90**(12):2711-2717.

Jaap, R. G., and W. F. Hollander. 1954. Wild type as standard in poultry genetics. Poultry Science **33**((1)):94-100.

Jacob, J. P., S. Milne, S. Beck, and J. Kaufman. 2000. The major and a minor class II beta-chain (B-LB) gene flank the Tapasin gene in the B-F/B-L region of the chicken major histocompatibility complex. Immunogenetics **51**(2):138-147.

Johnsgard, P. A. 1999. The Pheasants of the world: biology and natural history. Smithsonian Institution Press, Washington, D.C.

Kaufman, J., J. Jacob, I. Shaw, B. Walker, S. Milne, S. Beck, and J. Salomonsen. 1999a. Gene organisation determines evolution of function in the chicken MHC. Immunological Reviews **167**:101-117.

Kaufman, J., S. Milne, T. W. F. Gobel, B. A. Walker, J. P. Jacob, C. Auffray, R. Zoorob, and S. Beck. 1999b. The chicken B locus is a minimal essential major histocompatibility complex. Nature **401**(6756):923-925.

Letunic, I., and P. Bork. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Research **39**:W475-W478.

Lewontin, R. C. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49**((1)):49-67.

Librado, P., and J. Rozas. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics **25**(11):1451-1452.

Longmire, J. L., M. Maltbie, and R. Baker. 2000. Use of "Lysis Buffer" in DNA isolation and its implication for museum collections. Texas Tech University Museum.

Mekchay, S., P. Supakankul, A. Assawamakin, A. Wilantho, W. Chareanchim, and S. Tongsima. 2014. Population structure of four Thai indigenous chicken breeds. BMC Genetics **15**(40).

Miao, Y. W., M. S. Peng, G. S. Wu, et al. 2013. Chicken domestication: an updated perspective based on mitochondrial genomes. Heredity **110**(3):277-282.

Miller, M. M., L. D. Bacon, K. Hala, H. D. Hunt, S. J. Ewald, J. Kaufman, R. Zoorob, and W. E. Briles. 2004. 2004 Nomenclature for the chicken major histocompatibility (B and Y ) complex. Immunogenetics **56**(4):261-279.

Muir, W. M., G. K. S. Wong, Y. Zhang, et al. 2008. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. Proceedings of the National Academy of Sciences of the United States of America **105**(45):17312-17317.

Ngo, T. K. C., H. Simianer, H. Eding, H. V. Tieu, V. C. Cuong, C. B. A. Wollny, L. F. Groeneveld, and S. Weigend. 2010. Assessing genetic diversity of Vietnamese local chicken breeds using microsatellites. Animal Genetics **41**(5):545-547.

Nishibori, M., T. Shimogiri, T. Hayashi, and H. Yasue. 2005. Molecular evidence for hybridization of species in the genus Gallus except for Gallus varius. Animal Genetics **36**(5):367-375.

Parker, T. H. 2002. Benefits of female mate choice in the red junglefowl. Ph.D Thesis. The University of New Mexico, Ann Arbor.

Peterson, A. T., and I. L. Brisbin. 1998. Genetic endangerment of wild Red Junglefowl *Gallus gallus*. Bird Conservation International **8**(4):387-394.

Posada, D. 2002. Evaluation of methods for detecting recombination from DNA sequences: Empirical data. Molecular Biology and Evolution **19**(5):708-717.

Posada, D., and K. A. Crandall. 2001. Intraspecific gene genealogies: trees grafting into networks. Trends in Ecology & Evolution **16**(1):37-45.

Prim, R. C. 1957. Shortest connection networks and some generalizations. Bell System Technical Journal **36**(6):1389-1401.

Reich, D. E., M. Cargill, S. Bolk, et al. 2001. Linkage disequilibrium in the human genome. Nature **411**(6834):199-204.

Rogers, S., I. Shaw, N. Ross, V. Nair, L. Rothwell, J. Kaufman, and P. Kaiser. 2003. Analysis of part of the chicken Rfp-Y region reveals two novel lectin genes, the first complete genomic sequence of a class I alpha-chain gene, a truncated class II beta-chain gene, and a large CR1 repeat. Immunogenetics **55(**2):100-108.

Rohlf, F. J. 1973. Hierarchical clustering using minimum spanning tree. Computer Journal **16**(1):93-95.

Rokas, A., E. Ladoukakis, and E. Zouros. 2003. Animal mitochondrial DNA recombination revisited. Trends in Ecology & Evolution **18**(8):411-417.

Semagn, K., R. Babu, S. Hearne, and M. Olsen. 2014. Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. Molecular Breeding **33**(1):1-14.

Shiina, T., W. E. Briles, R. M. Goto, K. Hosomichi, K. Yanagiya, S. Shimizu, H. Inoko, and M. M. Miller. 2007. Extended gene map reveals tripartite motif, C-type lectin, and Ig superfamily type genes within a subregion of the chicken MHC-B affecting infectious disease. Journal of Immunology **178**(11):7162-7172.

Shiina, T., S. Shimizu, K. Hosomichi, S. Kohara, S. Watanabe, K. Hanzawa, S. Beck, J. K. Kulski, and H. Inoko. 2004. Comparative genomic analysis of two avian (quail and chicken) MHC regions. Journal of Immunology **172**(11):6751-6763.

Stephens, M., N. J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics **68**(4):978-989.

Storey, A. A., J. S. Athens, D. Bryant, et al. 2012. Investigating the global dispersal of chickens in rrehistory using ancient mitochondrial DNA signatures. PLOS One **7**(7).

Stumpf, M. P. H., and G. A. T. McVean. 2003. Estimating recombination rates from population-genetic data. Nature Reviews Genetics **4(**12):959-968.

Tajima, F. 1989. Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**(3):585-595.

Tregenza, T., and N. Wedell. 2000. Genetic compatibility, mate choice and patterns of parentage: Invited review. Molecular Ecology **9**(8):1013-1027.

Von Schantz, T., G. Goransson, G. Andersson, I. Froberg, M. Grahn, A. Helgee, and H. Wittzell. 1989. Female choice selects for a viability-based male trait in Pheasants. Nature **337**(6203):166-169.

Weir, B. S. 1996. Genetic data analysis II : methods for discrete population genetic data. Sinauer Associates, Sunderland, Massachussets.

Wigginton, J. E., D. J. Cutler, and G. R. Abecasis. 2005. A note on exact tests of Hardy-Weinberg equilibrium. American Journal of Human Genetics **76**(5):887-893.

Worley, K., J. Collet, L. G. Spurgin, C. Cornwallis, T. Pizzari, and D. S. Richardson. 2010. MHC heterozygosity and survival in red junglefowl. Molecular Ecology **19**(15):3064-3075.

Zuckerkandl , E., and L. Pauling. 1965. Molecules as documents of evolutionary history. Journal of Theoretical Biology **8**(2):357-&.

Table 3.1: SNP diversity in the four sampling sites.

*Heterozygosities have observed values (Obs H) and predicted values (Pred H) with Hardy-Weinberg p values (HWpval). MAF is minor allele frequency (of at least 0.05).*

| Locus Name | Position | CTN | | | | | HBA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Obs H | Pred H | HW pval | MAF | Alleles | Obs H | Pred H | HW pval | MAF | Alleles |
| **SNP-1** | **30189** | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C |
| **SNP-2** | **30246** | 0.0650 | 0.0630 | 1.0000 | 0.033 | C:T | 0.0890 | 0.0850 | 1.0000 | 0.045 | C:T |
| **SNP-3** | **43262** | 0.3910 | 0.4230 | 0.7947 | 0.304 | G:A | 0.2140 | 0.3370 | 0.0213 | 0.214 | G:A |
| **SNP-4** | **48730** | 0.0220 | 0.0220 | 1.0000 | 0.011 | G:T | 0.0890 | 0.0850 | 1.0000 | 0.045 | G:T |
| **SNP-5** | **49108** | 0.4570 | 0.4880 | 0.8268 | 0.424 | C:T | 0.4290 | 0.4480 | 0.9120 | 0.339 | C:T |
| **SNP-6** | **55684** | 0.6090 | 0.4850 | 0.1717 | 0.413 | T:C | 0.5000 | 0.4690 | 0.8881 | 0.375 | T:C |
| **SNP-7** | **57047** | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C |
| **SNP-8** | **59015** | 0.1740 | 0.1590 | 1.0000 | 0.087 | G:C | 0.1430 | 0.1330 | 1.0000 | 0.071 | G:C |
| **SNP-9** | **64656** | 0.5870 | 0.4940 | 0.3648 | 0.446 | G:A | 0.4290 | 0.4990 | 0.3875 | 0.482 | A:G |
| **SNP-10** | **68273** | 0.5000 | 0.4320 | 0.5152 | 0.315 | A:G | 0.4640 | 0.4360 | 0.9249 | 0.321 | A:G |
| **SNP-11** | **69257** | 0.0000 | 0.0000 | 1.0000 | 0.000 | G:G | 0.0000 | 0.0000 | 1.0000 | 0.000 | G:G |
| **SNP-12** | **75065** | 0.1960 | 0.2110 | 0.9834 | 0.120 | A:T | 0.2320 | 0.2820 | 0.3462 | 0.170 | A:T |
| **SNP-13** | **81752** | 0.5220 | 0.4230 | 0.2417 | 0.304 | C:T | 0.4820 | 0.4300 | 0.5987 | 0.312 | C:T |
| **SNP-14** | **85359** | 0.0870 | 0.0830 | 1.0000 | 0.043 | G:C | 0.0000 | 0.1330 | 0.0000 | 0.071 | G:C |
| **SNP-15** | **89076** | 0.0430 | 0.0430 | 1.0000 | 0.022 | C:T | 0.0180 | 0.0180 | 1.0000 | 0.009 | C:T |
| **SNP-16** | **94952** | 0.2170 | 0.2870 | 0.2229 | 0.174 | C:T | 0.2320 | 0.3840 | 0.0090 | 0.259 | C:T |
| **SNP-17** | **95066** | 0.1740 | 0.1590 | 1.0000 | 0.087 | C:T | 0.0710 | 0.0690 | 1.0000 | 0.036 | C:T |
| **SNP-18** | **95599** | 0.3480 | 0.3640 | 0.9926 | 0.239 | G:A | 0.2500 | 0.2190 | 0.7917 | 0.125 | G:A |
| **SNP-19** | **95934** | 0.1090 | 0.1030 | 1.0000 | 0.054 | C:G | 0.0180 | 0.0180 | 1.0000 | 0.009 | C:G |
| **SNP-20** | **97054** | 0.1520 | 0.1410 | 1.0000 | 0.076 | G:A | 0.3040 | 0.2570 | 0.4780 | 0.152 | G:A |
| **SNP-21** | **100610** | 0.3700 | 0.4150 | 0.6334 | 0.293 | G:A | 0.3390 | 0.3470 | 1.0000 | 0.223 | G:A |
| **SNP-22** | **100714** | 0.5220 | 0.4760 | 0.7964 | 0.391 | C:T | 0.4460 | 0.4420 | 1.0000 | 0.330 | C:T |
| **SNP-23** | **101288** | 0.5220 | 0.4910 | 0.9671 | 0.435 | T:G | 0.5180 | 0.4640 | 0.6118 | 0.366 | T:G |
| **SNP-24** | **101676** | 0.0430 | 0.0430 | 1.0000 | 0.022 | T:G | 0.0000 | 0.0000 | 1.0000 | 0.000 | T:T |
| **SNP-25** | **102094** | 0.3480 | 0.3150 | 0.9155 | 0.196 | T:C | 0.1610 | 0.1770 | 0.8378 | 0.098 | T:C |
| **SNP-26** | **103030** | 0.4570 | 0.4320 | 1.0000 | 0.315 | C:T | 0.4290 | 0.3750 | 0.5289 | 0.250 | C:T |
| **SNP-27** | **103447** | 0.4570 | 0.3750 | 0.3081 | 0.250 | C:T | 0.4110 | 0.4000 | 1.0000 | 0.277 | C:T |
| **SNP-28** | **104523** | 0.2170 | 0.2580 | 0.5159 | 0.152 | C:T | 0.2140 | 0.1910 | 1.0000 | 0.107 | C:T |

| Locus Name | Position | CTN | | | | | HBA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Obs H | Pred H | HW pval | MAF | Alleles | Obs H | Pred H | HW pval | MAF | Alleles |
| **SNP-29** | **108085** | 0.3910 | 0.4850 | 0.2776 | 0.413 | T:C | 0.3750 | 0.4300 | 0.4712 | 0.312 | T:C |
| **SNP-30** | **109194** | 0.4780 | 0.4400 | 0.8613 | 0.326 | C:T | 0.4110 | 0.4730 | 0.4373 | 0.384 | C:T |
| **SNP-31** | **109949** | 0.5430 | 0.4710 | 0.5163 | 0.380 | C:T | 0.3570 | 0.4360 | 0.2613 | 0.321 | C:T |
| **SNP-32** | **111557** | 0.0000 | 0.0000 | 1.0000 | 0.000 | T:T | 0.0000 | 0.0000 | 1.0000 | 0.000 | T:T |
| **SNP-33** | **112362** | 0.4780 | 0.4050 | 0.4385 | 0.283 | G:A | 0.3040 | 0.3840 | 0.1994 | 0.259 | G:A |
| **SNP-34** | **112488** | 0.4350 | 0.4850 | 0.6343 | 0.413 | G:A | 0.4640 | 0.4970 | 0.7649 | 0.464 | A:G |
| **SNP-35** | **116315** | 0.1300 | 0.1590 | 0.5647 | 0.087 | A:G | 0.1430 | 0.1330 | 1.0000 | 0.071 | A:G |
| **SNP-36** | **116687** | 0.3260 | 0.3280 | 1.0000 | 0.207 | T:C | 0.2320 | 0.2820 | 0.3462 | 0.170 | T:C |
| **SNP-37** | **117533** | 0.5220 | 0.4540 | 0.5391 | 0.348 | G:A | 0.5180 | 0.4540 | 0.4900 | 0.348 | G:A |
| **SNP-38** | **119134** | 0.4780 | 0.4230 | 0.6584 | 0.304 | C:T | 0.4460 | 0.4730 | 0.8322 | 0.384 | C:T |
| **SNP-39** | **120084** | 0.1960 | 0.2430 | 0.3967 | 0.141 | T:C | 0.2320 | 0.3050 | 0.1628 | 0.188 | T:C |
| **SNP-40** | **120680** | 0.4130 | 0.3960 | 1.0000 | 0.272 | T:C | 0.1960 | 0.2320 | 0.4732 | 0.134 | T:C |
| **SNP-41** | **122472** | 0.0220 | 0.0630 | 0.0659 | 0.033 | G:A | 0.0890 | 0.0850 | 1.0000 | 0.045 | G:A |
| **SNP-42** | **124279** | 0.1520 | 0.1410 | 1.0000 | 0.076 | T:C | 0.0890 | 0.0850 | 1.0000 | 0.045 | T:C |
| **SNP-43** | **124920** | 0.1740 | 0.1590 | 1.0000 | 0.087 | C:T | 0.1790 | 0.1630 | 1.0000 | 0.089 | C:T |
| **SNP-44** | **125176** | 0.2390 | 0.2430 | 1.0000 | 0.141 | G:T | 0.3750 | 0.3050 | 0.1976 | 0.188 | G:T |
| **SNP-45** | **125504** | 0.5000 | 0.4470 | 0.6938 | 0.337 | C:T | 0.4110 | 0.3840 | 0.9352 | 0.259 | C:T |
| **SNP-46** | **125845** | 0.3480 | 0.4050 | 0.4877 | 0.283 | C:T | 0.4640 | 0.4990 | 0.7440 | 0.482 | C:T |
| **SNP-47** | **126281** | 0.5220 | 0.4660 | 0.6773 | 0.370 | G:A | 0.4290 | 0.4080 | 1.0000 | 0.286 | G:A |
| **SNP-48** | **136539** | 0.3910 | 0.4910 | 0.2444 | 0.435 | T:G | 0.3390 | 0.5000 | 0.0260 | 0.491 | G:T |
| **SNP-49** | **136733** | 0.3040 | 0.3860 | 0.2583 | 0.261 | G:T | 0.3930 | 0.4940 | 0.1805 | 0.446 | G:T |
| **SNP-50** | **137666** | 0.3480 | 0.3860 | 0.6989 | 0.261 | T:C | 0.2680 | 0.3050 | 0.5601 | 0.188 | T:C |
| **SNP-51** | **138420** | 0.3700 | 0.4150 | 0.6334 | 0.293 | C:T | 0.3210 | 0.3920 | 0.2779 | 0.268 | C:T |
| **SNP-52** | **141694** | 0.3910 | 0.4050 | 1.0000 | 0.283 | C:T | 0.5180 | 0.4730 | 0.7223 | 0.384 | C:T |
| **SNP-53** | **142652** | 0.2170 | 0.1940 | 1.0000 | 0.109 | A:C | 0.1070 | 0.1010 | 1.0000 | 0.054 | A:C |
| **SNP-54** | **143396** | 0.2390 | 0.2430 | 1.0000 | 0.141 | C:T | 0.3040 | 0.4000 | 0.1247 | 0.277 | C:T |
| **SNP-55** | **147308** | 0.2610 | 0.2870 | 0.7901 | 0.174 | G:A | 0.4460 | 0.4160 | 0.8833 | 0.295 | G:A |
| **SNP-56** | **148723** | 0.4350 | 0.4760 | 0.7165 | 0.391 | G:A | 0.4640 | 0.4770 | 1.0000 | 0.393 | G:A |

| Locus Name | Position | CTN | | | | | HBA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Obs H | Pred H | HW pval | MAF | Alleles | Obs H | Pred H | HW pval | MAF | Alleles |
| **SNP-57** | **149999** | 0.1300 | 0.1220 | 1.0000 | 0.065 | T:C | 0.0710 | 0.0690 | 1.0000 | 0.036 | T:C |
| **SNP-58** | **156169** | 0.1520 | 0.1410 | 1.0000 | 0.076 | G:A | 0.1610 | 0.1770 | 0.8378 | 0.098 | G:A |
| **SNP-59** | **157474** | 0.1960 | 0.2110 | 0.9834 | 0.120 | G:A | 0.1610 | 0.1480 | 1.0000 | 0.080 | G:A |
| **SNP-60** | **158754** | 0.1300 | 0.1220 | 1.0000 | 0.065 | C:T | 0.0890 | 0.2050 | 0.0018 | 0.116 | C:T |
| **SNP-61** | **161219** | 0.3910 | 0.4230 | 0.7947 | 0.304 | G:A | 0.3040 | 0.4540 | 0.0243 | 0.348 | G:A |
| **SNP-62** | **163249** | 0.3260 | 0.4810 | 0.0494 | 0.402 | A:G | 0.5000 | 0.4770 | 0.9935 | 0.393 | A:G |
| **SNP-63** | **163580** | 0.2830 | 0.3280 | 0.5463 | 0.207 | G:A | 0.2860 | 0.3750 | 0.1381 | 0.250 | G:A |
| **SNP-64** | **168754** | 0.4570 | 0.4150 | 0.8147 | 0.293 | C:G | 0.5180 | 0.4420 | 0.3633 | 0.330 | C:G |
| **SNP-65** | **169823** | 0.3040 | 0.4050 | 0.1608 | 0.283 | G:A | 0.2140 | 0.3570 | 0.0097 | 0.232 | G:A |
| **SNP-66** | **170413** | 0.2830 | 0.2730 | 1.0000 | 0.163 | G:A | 0.1430 | 0.1630 | 0.7115 | 0.089 | G:A |
| **SNP-67** | **171391** | 0.5000 | 0.4150 | 0.3310 | 0.293 | C:G | 0.3930 | 0.3570 | 0.7745 | 0.232 | C:G |
| **SNP-68** | **175929** | 0.3260 | 0.3280 | 1.0000 | 0.207 | G:A | 0.3210 | 0.2700 | 0.3925 | 0.161 | G:A |
| **SNP-69** | **176828** | 0.0430 | 0.0430 | 1.0000 | 0.022 | G:T | 0.0540 | 0.0520 | 1.0000 | 0.027 | G:T |
| **SNP-70** | **177699** | 0.3700 | 0.3750 | 1.0000 | 0.250 | G:A | 0.3210 | 0.3570 | 0.6434 | 0.232 | G:A |
| **SNP-71** | **181329** | 0.1960 | 0.2110 | 0.9834 | 0.120 | C:T | 0.3390 | 0.2820 | 0.3173 | 0.170 | C:T |
| **SNP-72** | **182702** | 0.2830 | 0.4600 | 0.0174 | 0.359 | C:A | 0.4460 | 0.4300 | 1.0000 | 0.312 | C:A |
| **SNP-73** | **190516** | 0.0000 | 0.0000 | 1.0000 | 0.000 | G:G | 0.0180 | 0.0180 | 1.0000 | 0.009 | G:A |
| **SNP-74** | **199162** | 0.5220 | 0.4540 | 0.5391 | 0.348 | G:A | 0.3750 | 0.3470 | 0.9088 | 0.223 | G:A |
| **SNP-75** | **201654** | 0.5000 | 0.4320 | 0.5152 | 0.315 | A:G | 0.3930 | 0.4770 | 0.2656 | 0.393 | A:G |
| **SNP-76** | **207680** | 0.0000 | 0.0000 | 1.0000 | 0.000 | G:G | 0.0000 | 0.0000 | 1.0000 | 0.000 | G:G |
| **SNP-77** | **208958** | 0.1520 | 0.1410 | 1.0000 | 0.076 | A:T | 0.0180 | 0.0180 | 1.0000 | 0.009 | A:T |
| **SNP-78** | **213445** | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C |
| **SNP-79** | **225208** | 0.0000 | 0.0000 | 1.0000 | 0.000 | G:G | 0.0000 | 0.1330 | 0.0000 | 0.071 | G:A |
| **SNP-80** | **232805** | 0.5650 | 0.4990 | 0.5953 | 0.478 | G:A | 0.4820 | 0.4960 | 0.9936 | 0.455 | A:G |
| **SNP-81** | **234371** | 0.5220 | 0.4050 | 0.1161 | 0.283 | T:C | 0.5180 | 0.4990 | 1.0000 | 0.473 | T:C |
| **SNP-82** | **236165** | 0.4780 | 0.4990 | 0.9505 | 0.478 | C:T | 0.5180 | 0.4730 | 0.7223 | 0.384 | T:C |
| **SNP-83** | **238448** | 0.3480 | 0.3860 | 0.6989 | 0.261 | C:T | 0.3210 | 0.3370 | 0.9537 | 0.214 | C:T |
| **SNP-84** | **240933** | 0.0650 | 0.0630 | 1.0000 | 0.033 | G:C | 0.0000 | 0.0000 | 1.0000 | 0.000 | G:G |

| Locus Name | Position | LGO | | | | | YDN | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Obs H | Pred H | HW pval | MAF | Alleles | Obs H | Pred H | HW pval | MAF | Alleles |
| **SNP-1** | **30189** | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C |
| **SNP-2** | **30246** | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C |
| **SNP-3** | **43262** | 0.3680 | 0.4780 | 0.2450 | 0.395 | G:A | 0.4660 | 0.4670 | 1.0000 | 0.371 | G:A |
| **SNP-4** | **48730** | 0.0260 | 0.0260 | 1.0000 | 0.013 | G:T | 0.1380 | 0.1580 | 0.6900 | 0.086 | G:T |
| **SNP-5** | **49108** | 0.3160 | 0.3320 | 1.0000 | 0.211 | C:T | 0.4660 | 0.4880 | 0.8807 | 0.422 | C:T |
| **SNP-6** | **55684** | 0.2890 | 0.4970 | 0.0179 | 0.461 | T:C | 0.4310 | 0.4820 | 0.5460 | 0.405 | T:C |
| **SNP-7** | **57047** | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C |
| **SNP-8** | **59015** | 0.1580 | 0.1450 | 1.0000 | 0.079 | G:C | 0.0520 | 0.0820 | 0.1716 | 0.043 | G:C |
| **SNP-9** | **64656** | 0.2890 | 0.4410 | 0.0659 | 0.329 | A:G | 0.5520 | 0.4950 | 0.5836 | 0.448 | A:G |
| **SNP-10** | **68273** | 0.2630 | 0.3880 | 0.1021 | 0.263 | A:G | 0.3620 | 0.4210 | 0.4024 | 0.302 | A:G |
| **SNP-11** | **69257** | 0.0000 | 0.0000 | 1.0000 | 0.000 | G:G | 0.0000 | 0.0000 | 1.0000 | 0.000 | G:G |
| **SNP-12** | **75065** | 0.0790 | 0.1230 | 0.2612 | 0.066 | A:T | 0.1030 | 0.1280 | 0.4552 | 0.069 | A:T |
| **SNP-13** | **81752** | 0.3680 | 0.4880 | 0.2060 | 0.421 | C:T | 0.5000 | 0.5000 | 1.0000 | 0.491 | C:T |
| **SNP-14** | **85359** | 0.2890 | 0.2480 | 0.8606 | 0.145 | G:C | 0.0860 | 0.1130 | 0.3492 | 0.060 | G:C |
| **SNP-15** | **89076** | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C | 0.0340 | 0.0340 | 1.0000 | 0.017 | C:T |
| **SNP-16** | **94952** | 0.2630 | 0.4320 | 0.0346 | 0.316 | C:T | 0.3790 | 0.4280 | 0.5225 | 0.310 | C:T |
| **SNP-17** | **95066** | 0.0260 | 0.0760 | 0.0800 | 0.039 | C:T | 0.0690 | 0.0670 | 1.0000 | 0.034 | C:T |
| **SNP-18** | **95599** | 0.1580 | 0.1450 | 1.0000 | 0.079 | G:A | 0.3100 | 0.3480 | 0.5873 | 0.224 | G:A |
| **SNP-19** | **95934** | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C | 0.1210 | 0.1720 | 0.1338 | 0.095 | C:G |
| **SNP-20** | **97054** | 0.0790 | 0.0760 | 1.0000 | 0.039 | G:A | 0.1550 | 0.1430 | 1.0000 | 0.078 | G:A |
| **SNP-21** | **100610** | 0.2370 | 0.4580 | 0.0064 | 0.355 | G:A | 0.2930 | 0.3750 | 0.1693 | 0.250 | G:A |
| **SNP-22** | **100714** | 0.4210 | 0.4990 | 0.4717 | 0.474 | C:T | 0.3280 | 0.4750 | 0.0307 | 0.388 | C:T |
| **SNP-23** | **101288** | 0.4470 | 0.5000 | 0.6859 | 0.487 | T:G | 0.3620 | 0.4880 | 0.0764 | 0.422 | T:G |
| **SNP-24** | **101676** | 0.0000 | 0.0000 | 1.0000 | 0.000 | T:T | 0.1030 | 0.0980 | 1.0000 | 0.052 | T:G |
| **SNP-25** | **102094** | 0.1320 | 0.1670 | 0.5222 | 0.092 | T:C | 0.2590 | 0.2740 | 0.9252 | 0.164 | T:C |
| **SNP-26** | **103030** | 0.3160 | 0.3610 | 0.6496 | 0.237 | C:T | 0.2760 | 0.2620 | 1.0000 | 0.155 | C:T |
| **SNP-27** | **103447** | 0.2890 | 0.3750 | 0.2846 | 0.250 | C:T | 0.5170 | 0.4900 | 0.9363 | 0.431 | C:T |
| **SNP-28** | **104523** | 0.1840 | 0.4000 | 0.0031 | 0.276 | C:T | 0.1550 | 0.2250 | 0.0803 | 0.129 | C:T |

| Locus Name | Position | LGO | | | | | YDN | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Obs H | Pred H | HW pval | MAF | Alleles | Obs H | Pred H | HW pval | MAF | Alleles |
| **SNP-29** | **108085** | 0.2630 | 0.4650 | 0.0150 | 0.368 | T:C | 0.4660 | 0.4820 | 0.9537 | 0.405 | T:C |
| **SNP-30** | **109194** | 0.3680 | 0.4110 | 0.7218 | 0.289 | C:T | 0.3620 | 0.4070 | 0.5463 | 0.284 | C:T |
| **SNP-31** | **109949** | 0.2630 | 0.3610 | 0.1894 | 0.237 | C:T | 0.3100 | 0.4140 | 0.0982 | 0.293 | C:T |
| **SNP-32** | **111557** | 0.0000 | 0.0000 | 1.0000 | 0.000 | T:T | 0.0000 | 0.0000 | 1.0000 | 0.000 | T:T |
| **SNP-33** | **112362** | 0.2630 | 0.2290 | 1.0000 | 0.132 | G:A | 0.4480 | 0.4850 | 0.7056 | 0.414 | G:A |
| **SNP-34** | **112488** | 0.4470 | 0.5000 | 0.6859 | 0.487 | G:A | 0.3970 | 0.4210 | 0.8254 | 0.302 | G:A |
| **SNP-35** | **116315** | 0.2110 | 0.1880 | 1.0000 | 0.105 | A:G | 0.1720 | 0.1580 | 1.0000 | 0.086 | A:G |
| **SNP-36** | **116687** | 0.1580 | 0.1450 | 1.0000 | 0.079 | T:C | 0.0860 | 0.1720 | 0.0090 | 0.095 | T:C |
| **SNP-37** | **117533** | 0.2890 | 0.4830 | 0.0250 | 0.408 | G:A | 0.5340 | 0.4960 | 0.7984 | 0.457 | A:G |
| **SNP-38** | **119134** | 0.4210 | 0.4940 | 0.5006 | 0.447 | C:T | 0.4140 | 0.4520 | 0.6696 | 0.345 | C:T |
| **SNP-39** | **120084** | 0.3680 | 0.4500 | 0.3944 | 0.342 | T:C | 0.3100 | 0.2620 | 0.4196 | 0.155 | T:C |
| **SNP-40** | **120680** | 0.0530 | 0.0510 | 1.0000 | 0.026 | T:C | 0.3100 | 0.4410 | 0.0441 | 0.328 | T:C |
| **SNP-41** | **122472** | 0.0000 | 0.0000 | 1.0000 | 0.000 | G:G | 0.1030 | 0.0980 | 1.0000 | 0.052 | G:A |
| **SNP-42** | **124279** | 0.2370 | 0.2840 | 0.5477 | 0.171 | T:C | 0.1380 | 0.1580 | 0.6900 | 0.086 | T:C |
| **SNP-43** | **124920** | 0.0530 | 0.1450 | 0.0163 | 0.079 | C:T | 0.0340 | 0.0340 | 1.0000 | 0.017 | C:T |
| **SNP-44** | **125176** | 0.2110 | 0.3880 | 0.0148 | 0.263 | G:T | 0.2760 | 0.2380 | 0.6031 | 0.138 | G:T |
| **SNP-45** | **125504** | 0.2890 | 0.4410 | 0.0659 | 0.329 | C:T | 0.2760 | 0.3660 | 0.1164 | 0.241 | C:T |
| **SNP-46** | **125845** | 0.2110 | 0.4320 | 0.0042 | 0.316 | C:T | 0.4310 | 0.4460 | 0.9645 | 0.336 | C:T |
| **SNP-47** | **126281** | 0.3160 | 0.4940 | 0.0449 | 0.447 | G:A | 0.4660 | 0.4460 | 1.0000 | 0.336 | G:A |
| **SNP-48** | **136539** | 0.3160 | 0.4320 | 0.1701 | 0.316 | T:G | 0.5170 | 0.4950 | 0.9879 | 0.448 | T:G |
| **SNP-49** | **136733** | 0.2630 | 0.4780 | 0.0110 | 0.395 | G:T | 0.3450 | 0.3660 | 0.8517 | 0.241 | G:T |
| **SNP-50** | **137666** | 0.1840 | 0.3470 | 0.0153 | 0.224 | T:C | 0.3970 | 0.3570 | 0.7074 | 0.233 | T:C |
| **SNP-51** | **138420** | 0.2890 | 0.4000 | 0.1672 | 0.276 | C:T | 0.5000 | 0.4880 | 1.0000 | 0.422 | C:T |
| **SNP-52** | **141694** | 0.5000 | 0.4970 | 1.0000 | 0.461 | C:T | 0.4830 | 0.4990 | 0.9542 | 0.483 | C:T |
| **SNP-53** | **142652** | 0.0000 | 0.0510 | 0.0267 | 0.026 | A:C | 0.1900 | 0.1720 | 1.0000 | 0.095 | A:C |
| **SNP-54** | **143396** | 0.2370 | 0.2840 | 0.5477 | 0.171 | C:T | 0.3450 | 0.4280 | 0.2136 | 0.310 | C:T |
| **SNP-55** | **147308** | 0.1580 | 0.3010 | 0.0195 | 0.184 | G:A | 0.1210 | 0.1720 | 0.1338 | 0.095 | G:A |
| **SNP-56** | **148723** | 0.3160 | 0.4990 | 0.0409 | 0.474 | A:G | 0.3790 | 0.4620 | 0.2503 | 0.362 | G:A |

| Locus Name | Position | LGO | | | | | YDN | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Obs H | Pred H | HW pval | MAF | Alleles | Obs H | Pred H | HW pval | MAF | Alleles |
| **SNP-57** | **149999** | 0.0000 | 0.0000 | 1.0000 | 0.000 | T:T | 0.0690 | 0.0670 | 1.0000 | 0.034 | T:C |
| **SNP-58** | **156169** | 0.0530 | 0.1000 | 0.1589 | 0.053 | G:A | 0.0340 | 0.0340 | 1.0000 | 0.017 | G:A |
| **SNP-59** | **157474** | 0.3420 | 0.3170 | 1.0000 | 0.197 | G:A | 0.1550 | 0.2250 | 0.0803 | 0.129 | G:A |
| **SNP-60** | **158754** | 0.1840 | 0.1670 | 1.0000 | 0.092 | C:T | 0.0520 | 0.0820 | 0.1716 | 0.043 | C:T |
| **SNP-61** | **161219** | 0.2630 | 0.3880 | 0.1021 | 0.263 | G:A | 0.5000 | 0.5000 | 1.0000 | 0.491 | G:A |
| **SNP-62** | **163249** | 0.2110 | 0.4320 | 0.0042 | 0.316 | A:G | 0.4310 | 0.3750 | 0.4774 | 0.250 | A:G |
| **SNP-63** | **163580** | 0.1580 | 0.1880 | 0.6721 | 0.105 | G:A | 0.2240 | 0.2250 | 1.0000 | 0.129 | G:A |
| **SNP-64** | **168754** | 0.2890 | 0.4410 | 0.0659 | 0.329 | C:G | 0.4140 | 0.4140 | 1.0000 | 0.293 | C:G |
| **SNP-65** | **169823** | 0.3950 | 0.4720 | 0.4503 | 0.382 | G:A | 0.2070 | 0.4280 | 0.0002 | 0.310 | G:A |
| **SNP-66** | **170413** | 0.1050 | 0.1450 | 0.3836 | 0.079 | G:A | 0.2590 | 0.2740 | 0.9252 | 0.164 | G:A |
| **SNP-67** | **171391** | 0.3420 | 0.3750 | 0.8181 | 0.250 | C:G | 0.4660 | 0.4340 | 0.8705 | 0.319 | C:G |
| **SNP-68** | **175929** | 0.4210 | 0.3880 | 1.0000 | 0.263 | G:A | 0.2410 | 0.4000 | 0.0069 | 0.276 | G:A |
| **SNP-69** | **176828** | 0.0530 | 0.0510 | 1.0000 | 0.026 | G:T | 0.0690 | 0.0670 | 1.0000 | 0.034 | G:T |
| **SNP-70** | **177699** | 0.3160 | 0.3320 | 1.0000 | 0.211 | G:A | 0.2590 | 0.2250 | 0.7083 | 0.129 | G:A |
| **SNP-71** | **181329** | 0.1320 | 0.1670 | 0.5222 | 0.092 | C:T | 0.2410 | 0.2380 | 1.0000 | 0.138 | C:T |
| **SNP-72** | **182702** | 0.3680 | 0.4650 | 0.3047 | 0.368 | C:A | 0.4830 | 0.4900 | 1.0000 | 0.431 | C:A |
| **SNP-73** | **190516** | 0.0260 | 0.0260 | 1.0000 | 0.013 | G:A | 0.0000 | 0.0000 | 1.0000 | 0.000 | G:G |
| **SNP-74** | **199162** | 0.2110 | 0.3880 | 0.0148 | 0.263 | G:A | 0.5340 | 0.4990 | 0.8275 | 0.474 | G:A |
| **SNP-75** | **201654** | 0.4740 | 0.4990 | 0.9477 | 0.474 | A:G | 0.3970 | 0.3920 | 1.0000 | 0.267 | A:G |
| **SNP-76** | **207680** | 0.0000 | 0.0000 | 1.0000 | 0.000 | G:G | 0.0340 | 0.0340 | 1.0000 | 0.017 | G:A |
| **SNP-77** | **208958** | 0.0000 | 0.0510 | 0.0267 | 0.026 | A:T | 0.1210 | 0.1130 | 1.0000 | 0.060 | A:T |
| **SNP-78** | **213445** | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C | 0.0000 | 0.0000 | 1.0000 | 0.000 | C:C |
| **SNP-79** | **225208** | 0.0000 | 0.0000 | 1.0000 | 0.000 | G:G | 0.0000 | 0.0000 | 1.0000 | 0.000 | G:G |
| **SNP-80** | **232805** | 0.3420 | 0.4410 | 0.2664 | 0.329 | A:G | 0.3100 | 0.4620 | 0.0223 | 0.362 | A:G |
| **SNP-81** | **234371** | 0.2110 | 0.4650 | 0.0016 | 0.368 | T:C | 0.3280 | 0.4070 | 0.2172 | 0.284 | T:C |
| **SNP-82** | **236165** | 0.3420 | 0.5000 | 0.0855 | 0.487 | T:C | 0.3620 | 0.5000 | 0.0551 | 0.491 | T:C |
| **SNP-83** | **238448** | 0.2110 | 0.3010 | 0.1674 | 0.184 | C:T | 0.3620 | 0.4570 | 0.1721 | 0.353 | C:T |
| **SNP-84** | **240933** | 0.1320 | 0.1230 | 1.0000 | 0.066 | G:C | 0.0000 | 0.0000 | 1.0000 | 0.000 | G:G |

Table 3.2: Haplotype diversity and statistics of Red Junglefowl in the four sampling sites.

| **Site** | **CTN** | **HBA** | **LGO** | **YDN** |
|---|---:|---:|---:|---:|
| Sample | 46 | 56 | 39 | 58 |
| Total haplotypes | 92 | 112 | 78 | 116 |
| Unique haplotypes | 92 | 82 | 48 | 91 |
| Haplotype diversity $H_d$ | 100% | 98.97% | 97.86% | 99.39% |
| Gene diversity | 76 | 76 | 71 | 75 |
| Average number of difference $K$ | 24.5745 | 23.9067 | 23.4393 | 24.1789 |
| Nucleotide diversity $\pi$ | 0.2926 | 0.2846 | 0.2790 | 0.2878 |
| Tajima's $D$ | 2.1236 | 2.1381 | 2.0641 | 2.2986 |
| Segregation site $S$ | 76 | 76 | 71 | 75 |
| Recombination parameter $\hat{\rho}$ | 0.0065 | 0.0020 | 0.0013 | 0.0024 |
| Recombination rate (average) $\rho$ | 1.3100 | 1.3000 | 1.13000 | 1.3400 |
| Expected heterozygosity $H_E$ | 0.2894 | 0.2821 | 0.2753 | 0.2854 |

*Gene diversity is number of polymorphic loci (out of total 84 loci). All Tajima' D for neutrality tests are significant ($p < 0.05$), indicating balancing selection.*

Table 3.3: Analysis of Molecular Variance (AMOVA).

| **Source of Variation** | ***Degree of freedom*** | **Sum of Squares** | **Variance components** | **Percentage of variation** |
|---|---:|---:|---:|---:|
| Among *K* populations | 3 | 2.791 | 0.00417 | 0.83% |
| Among *N* individuals within *K* populations | 195 | 101.187 | 0.02370 | 4.66% |
| Within *N* individuals | 199 | 94.000 | 0.47236 | 94.51% |
| **Total** | 397 | 197.977 | 0.49980 | 100.00% |

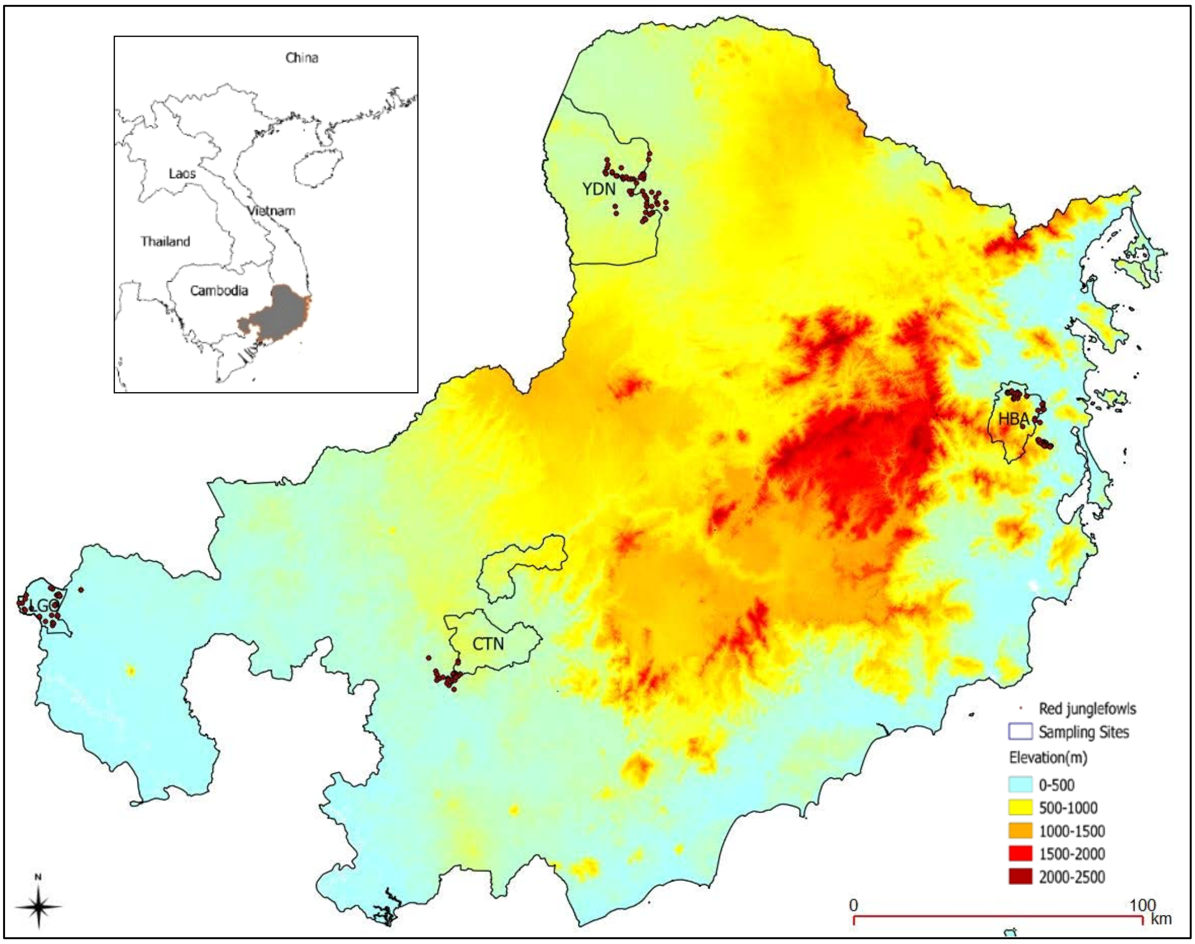
*N = 199 Red Junglefowl, K = 4 populations.*

Table 3.4: Comparison of MHC haplotypes recorded in different studies.

*Red Junglefowl (this study) and in commercial chicken lines (unpubl. data).*

| **Line type** | **No. Samples** | **No. Haplotypes** | **Haplotype percentage** |
|---|---:|---:|---:|
| Red Junglefowl | 199 | 310 | 77.89% |
| Broiler-UAB-AMC-1957 | 71 | 8 | 5.63% |
| Broiler-UAB-AMC-1978S | 64 | 5 | 3.91% |
| Broiler-UAB-AMC-1978D | 78 | 10 | 6.41% |
| Broiler-UGA-ACRB | 100 | 11 | 5.50% |
| Broiler-UGA-ARB | 71 | 4 | 2.82% |
| Broiler-UAR-RB | 54 | 7 | 6.48% |
| Standard-UAB-BPR | 76 | 4 | 2.63% |
| Standard-UAB-SBPR | 80 | 4 | 2.50% |
| Standard-USK-BPR | 96 | 2 | 1.04% |
| Standard-UAB-SRIR | 80 | 4 | 2.50% |
| Standard-UAB-WL | 72 | 3 | 2.08% |
| Standard-UAB-LS | 77 | 3 | 1.95% |
| Standard-UAB-NH | 73 | 4 | 2.74% |
| Standard-Ill-NH | 94 | 3 | 1.60% |
| Standard-UAB-BL | 76 | 1 | 0.66% |
| Synthetic-Ill-PC | 92 | 3 | 1.63% |
| Synthetic-USK-EPI | 97 | 9 | 4.64% |

Figure 3.1: Sampling sites in South Central Vietnam.



*With the putative Annamite landscape barriers. Sampling sites: Cat Tien National Park and Dong Nai Nature Reserve (here after CTN, as the two sites are connected), Hon Ba Nature Reserve (HBA), Lo Go Sa Mat National Park (LGO), and Yok Don National Park (YDN).*

Figure 3.2: Diagram of MHC B-locus in chickens (after Shiina et al. 2007).

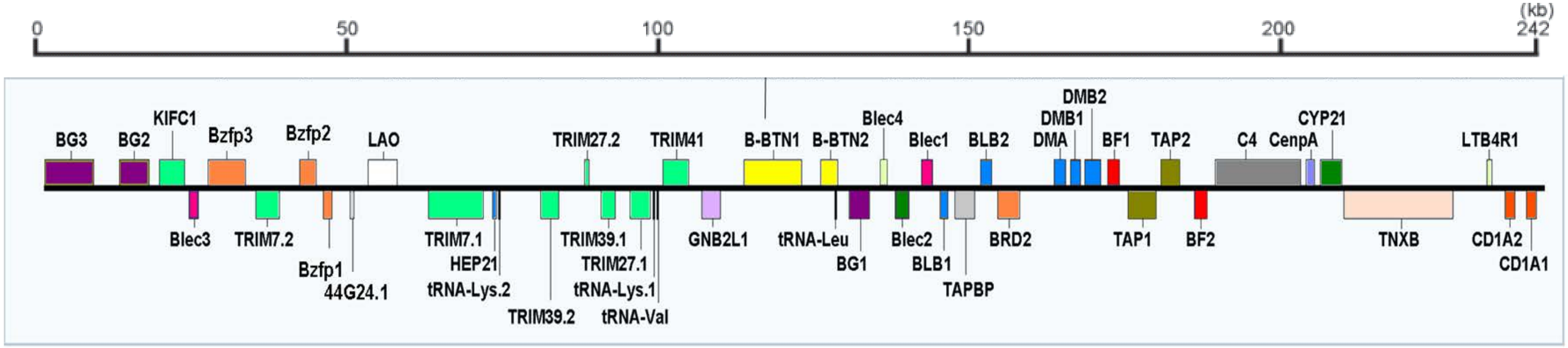
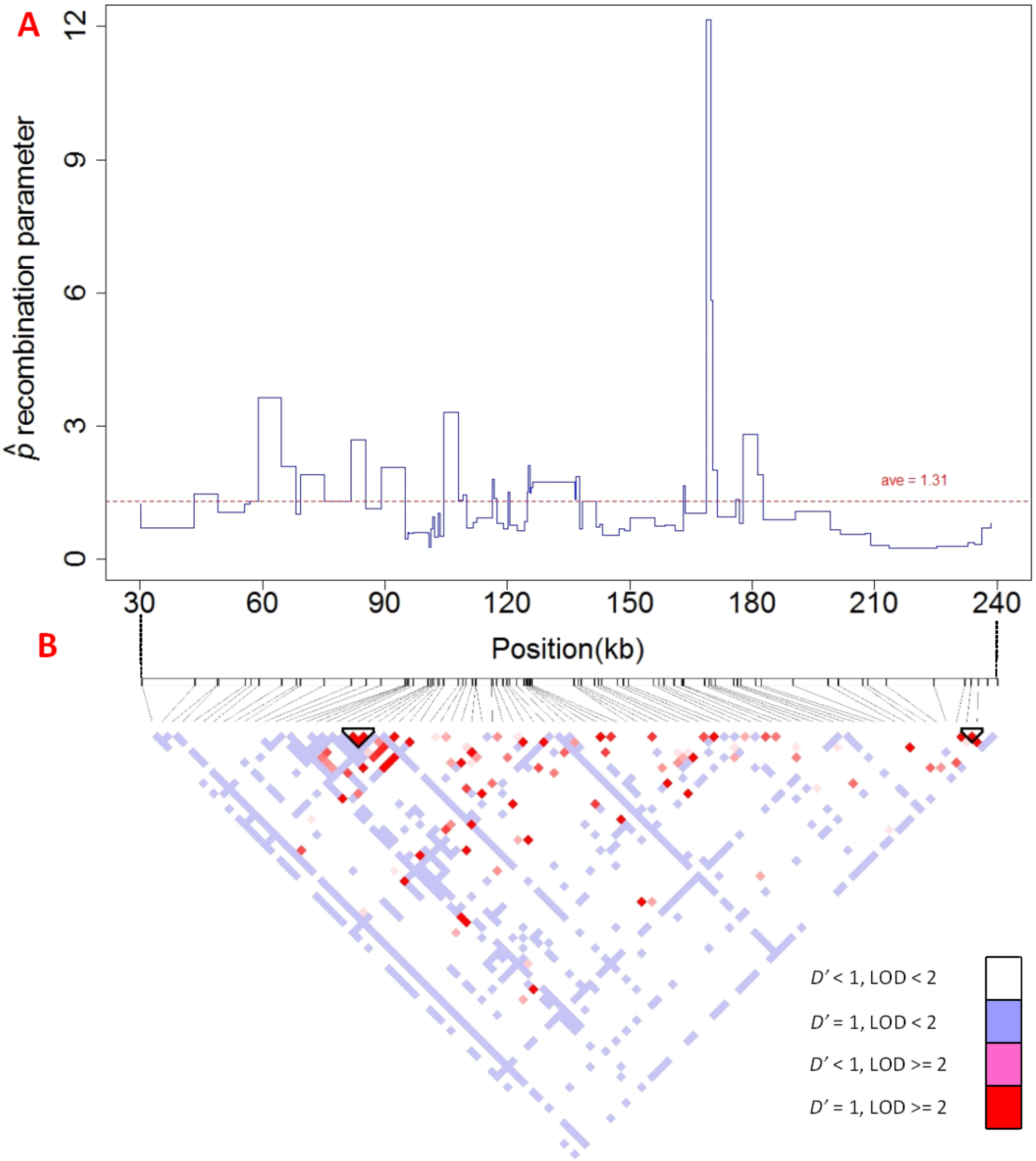
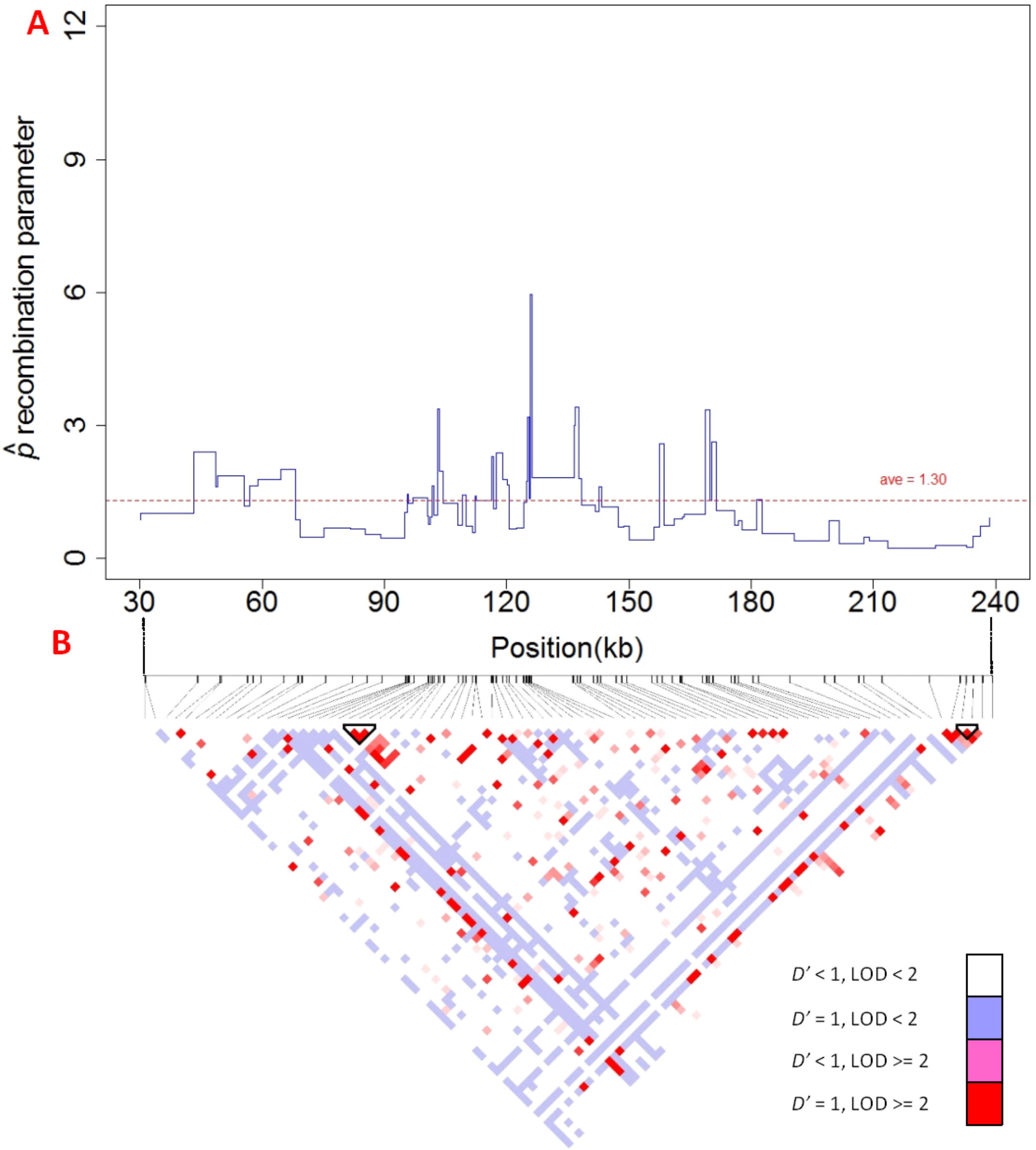


Figure 3.3:  MHC recombinants in CTN.



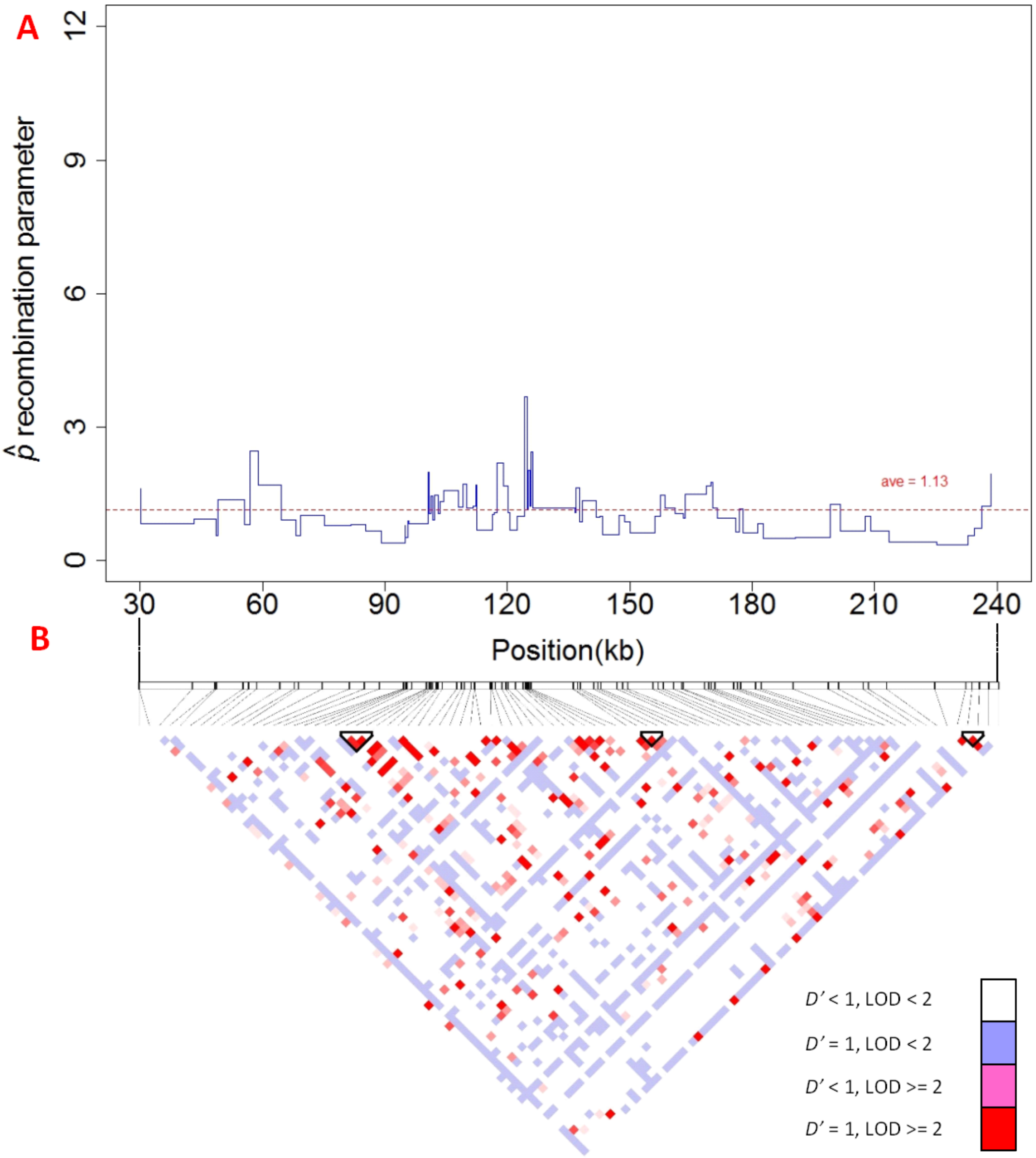
*(A) Recombination. (B) Linkage Disequilibrium. Estimate of recombination rates $\rho$ (per base pair) - as the factor by which recombination between any two loci exceeds the background recombination parameter $\hat{\rho}$. D' is the pairwise disequilibrium coefficient and LOD score is logarithm (base 10) of odds for linkage. Black 'triangle' is haplotype block with high LD.*

Figure 3.4: MHC recombinants in HBA.



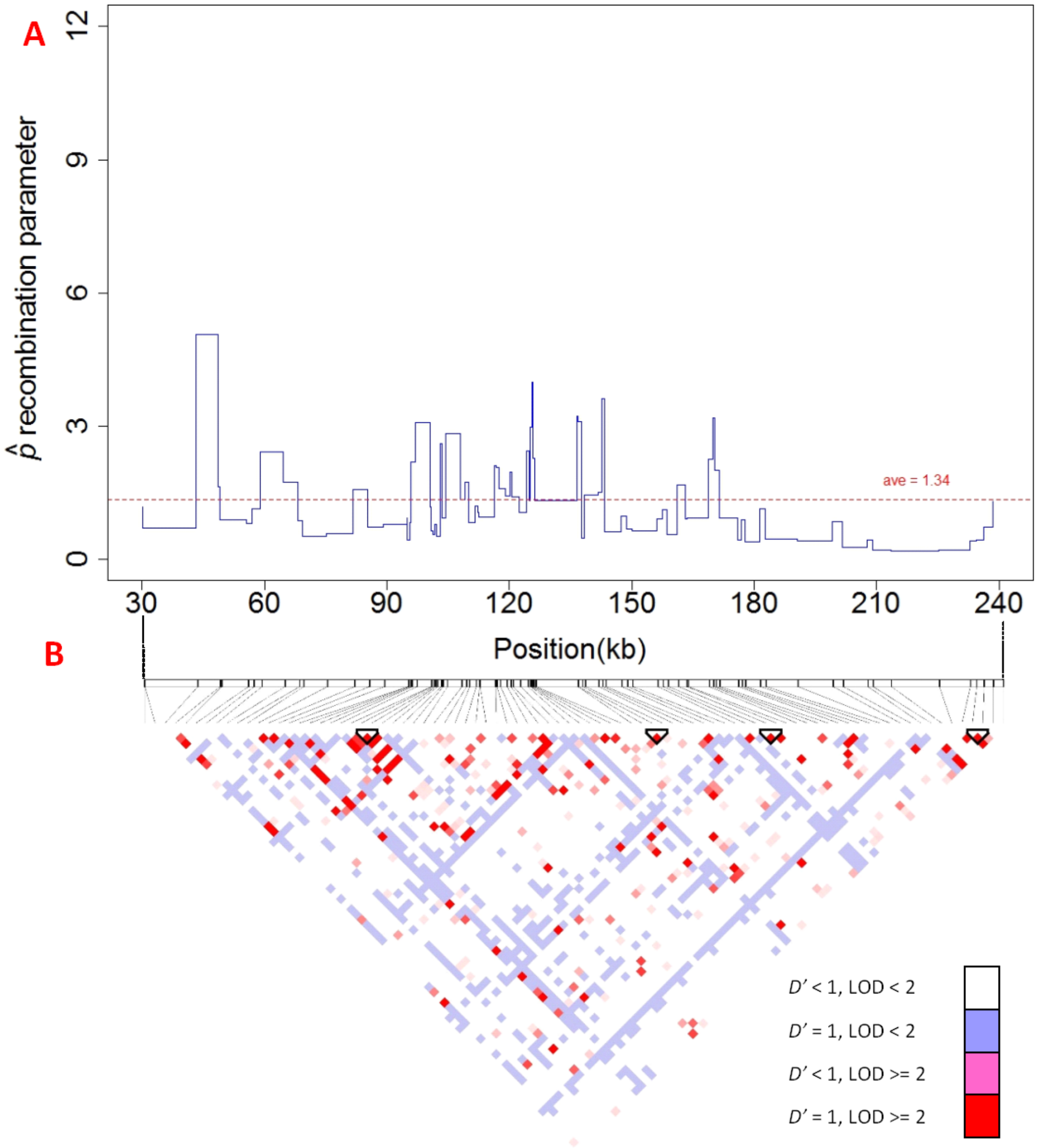
*(A) Recombination. (B) Linkage Disequilibrium. Estimate of recombination rates $\rho$ (per base pair) - as the factor by which recombination between any two loci exceeds the background recombination parameter $\hat{\rho}$ . D' is the pairwise disequilibrium coefficient and LOD score is logarithm (base 10) of odds for linkage. Black 'triangle' is haplotype block with high LD.*

Figure 3.5: MHC recombinants in LGO.



*(A) Recombination. (B) Linkage Disequilibrium. Estimate of recombination rates $\rho$ (per base pair) - as the factor by which recombination between any two loci exceeds the background recombination parameter $\hat{\rho}$. D' is the pairwise disequilibrium coefficient and LOD score is logarithm (base 10) of odds for linkage. Black 'triangle' is haplotype block with high LD.*

Figure 3.6: MHC recombinants in YDN.



*(A) Recombination. (B) Linkage Disequilibrium. Estimate of recombination rates* $\rho$ *(per base pair) - as the factor by which recombination between any two loci exceeds the background recombination parameter* $\hat{\rho}$ *. D' is the pairwise disequilibrium coefficient and LOD score is logarithm (base 10) of odds for linkage. Black 'triangle' is haplotype block with high LD.*

Figure 3.7: MHC recombination estimated by *SequenceLDhat*.

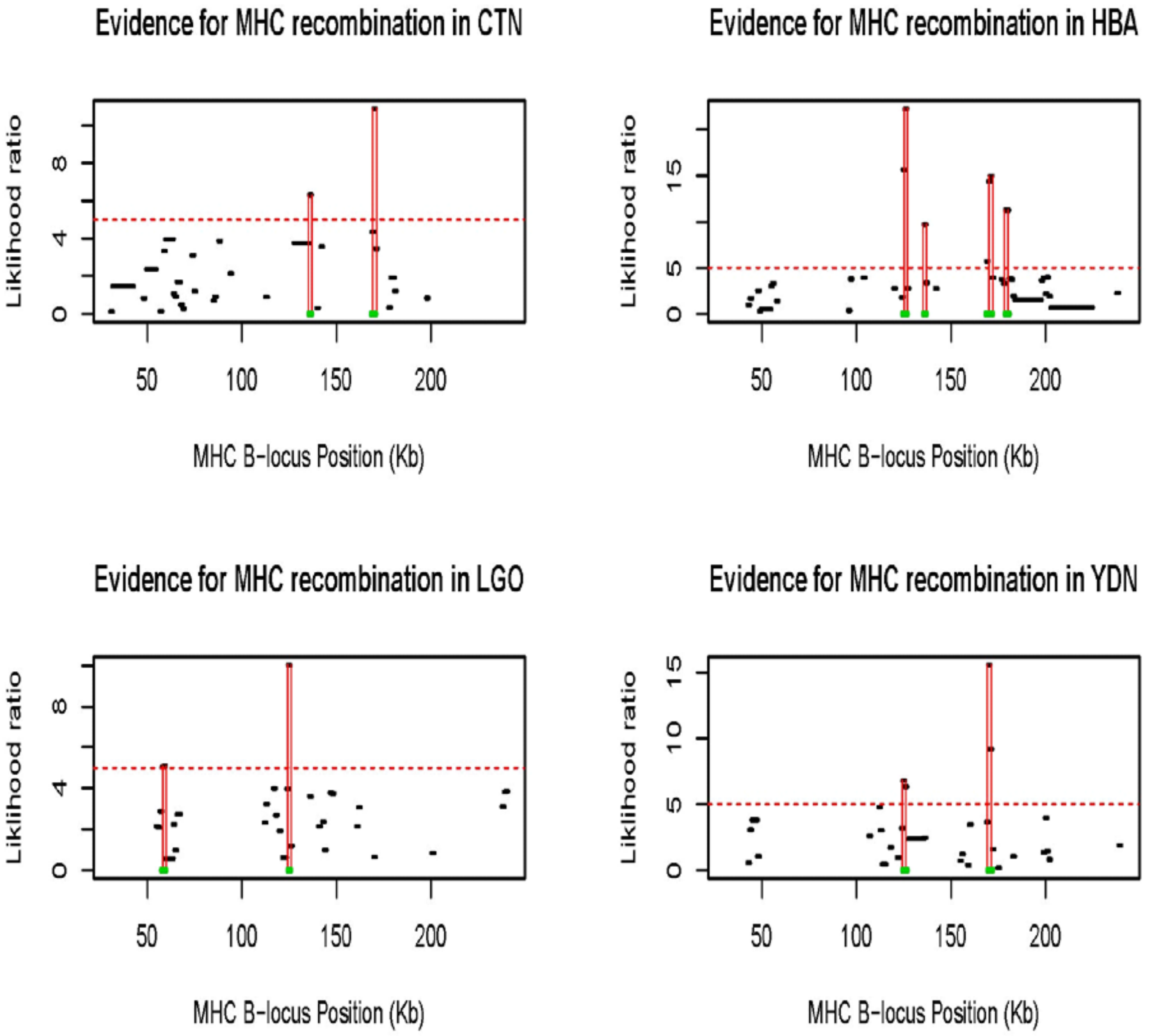


Figure 3.8: Haplotype network for four field sites.

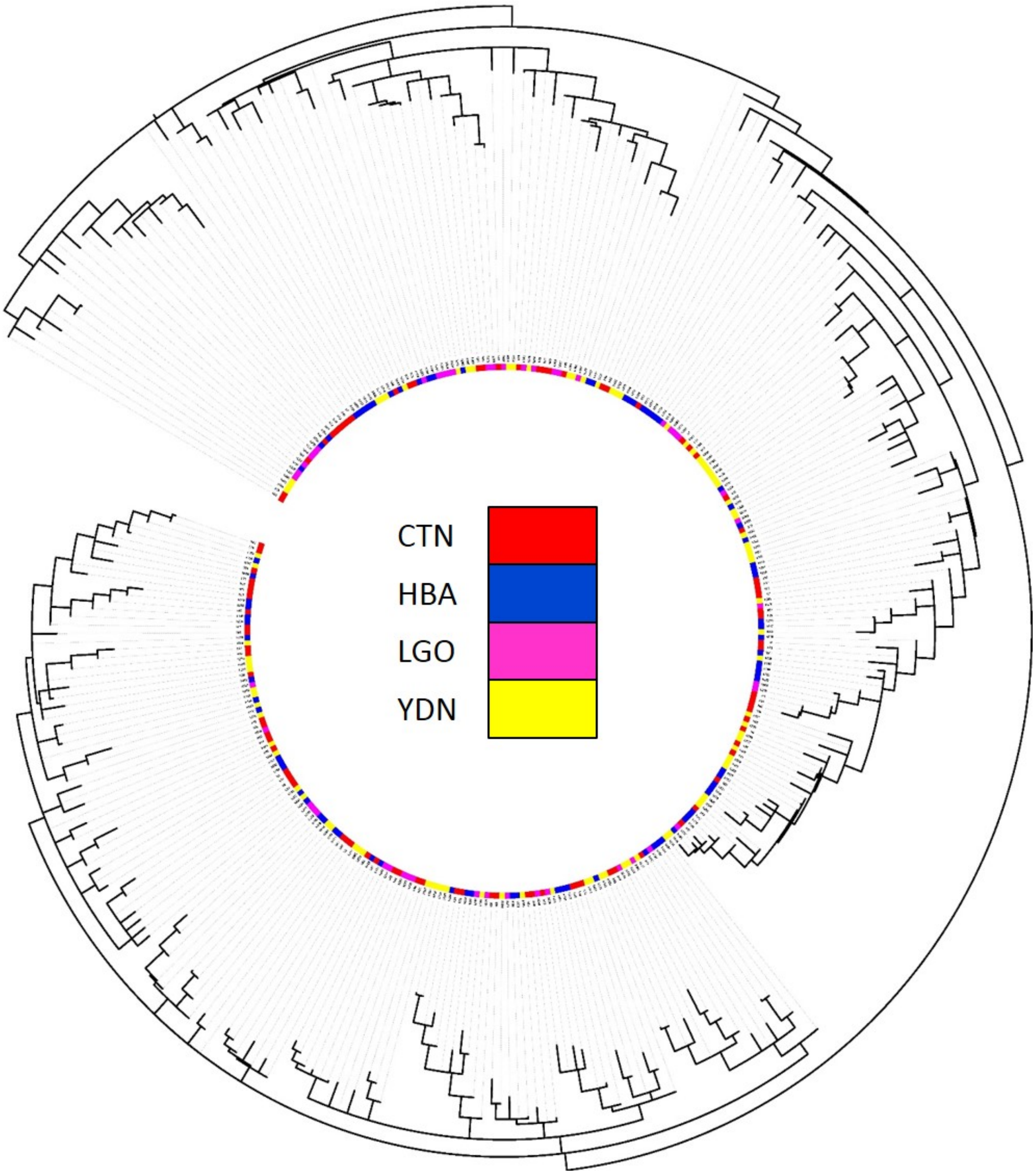


Figure 3.9: Haplotype network in CTN.

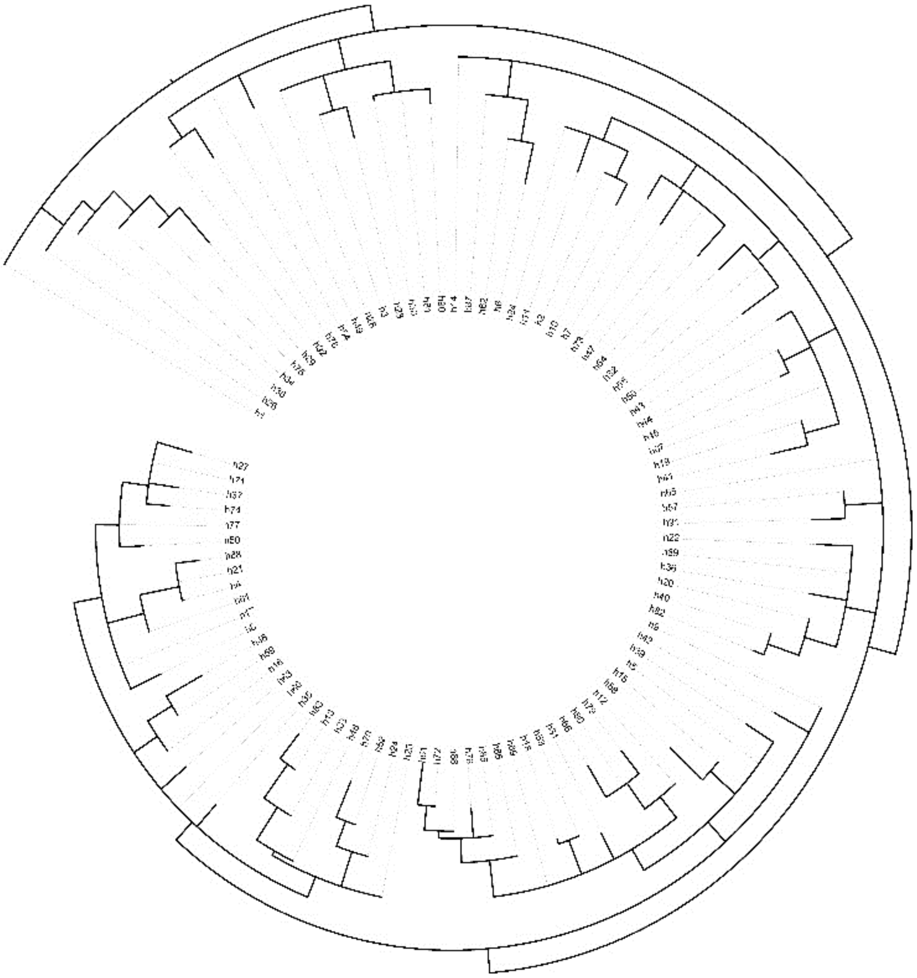


Figure 3.10: Haplotype network in HBA.

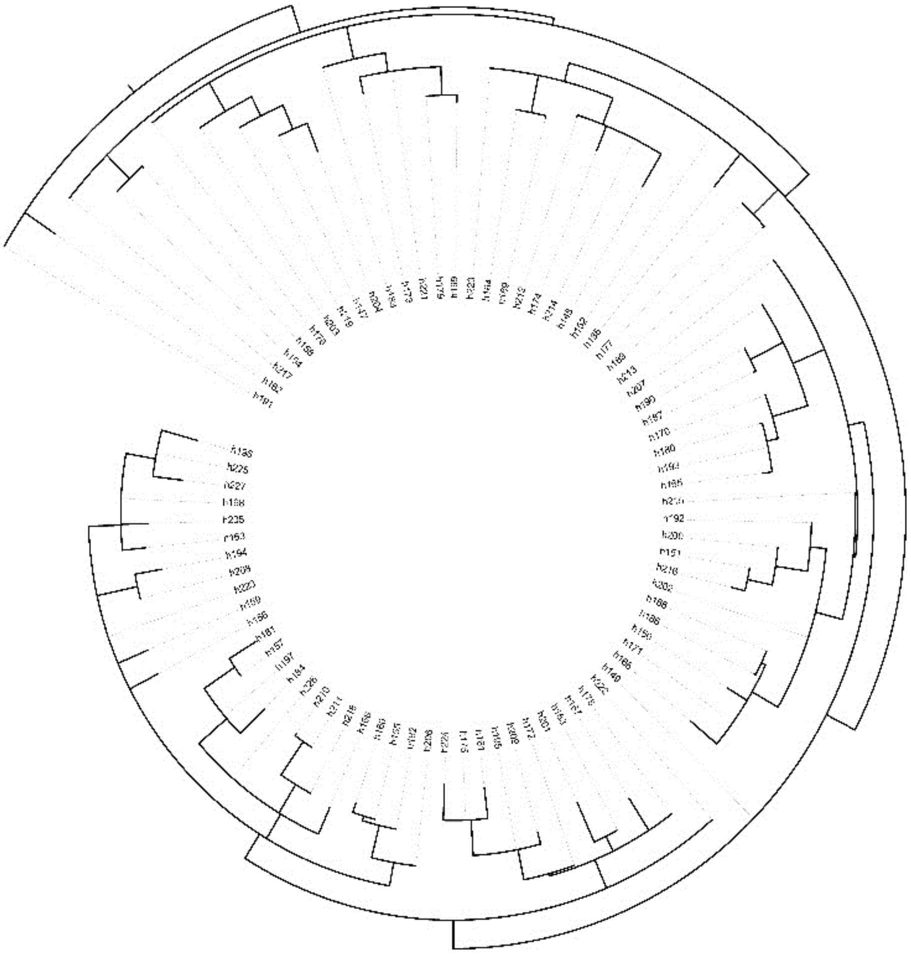


Figure 3.11: Haplotype network in LGO.

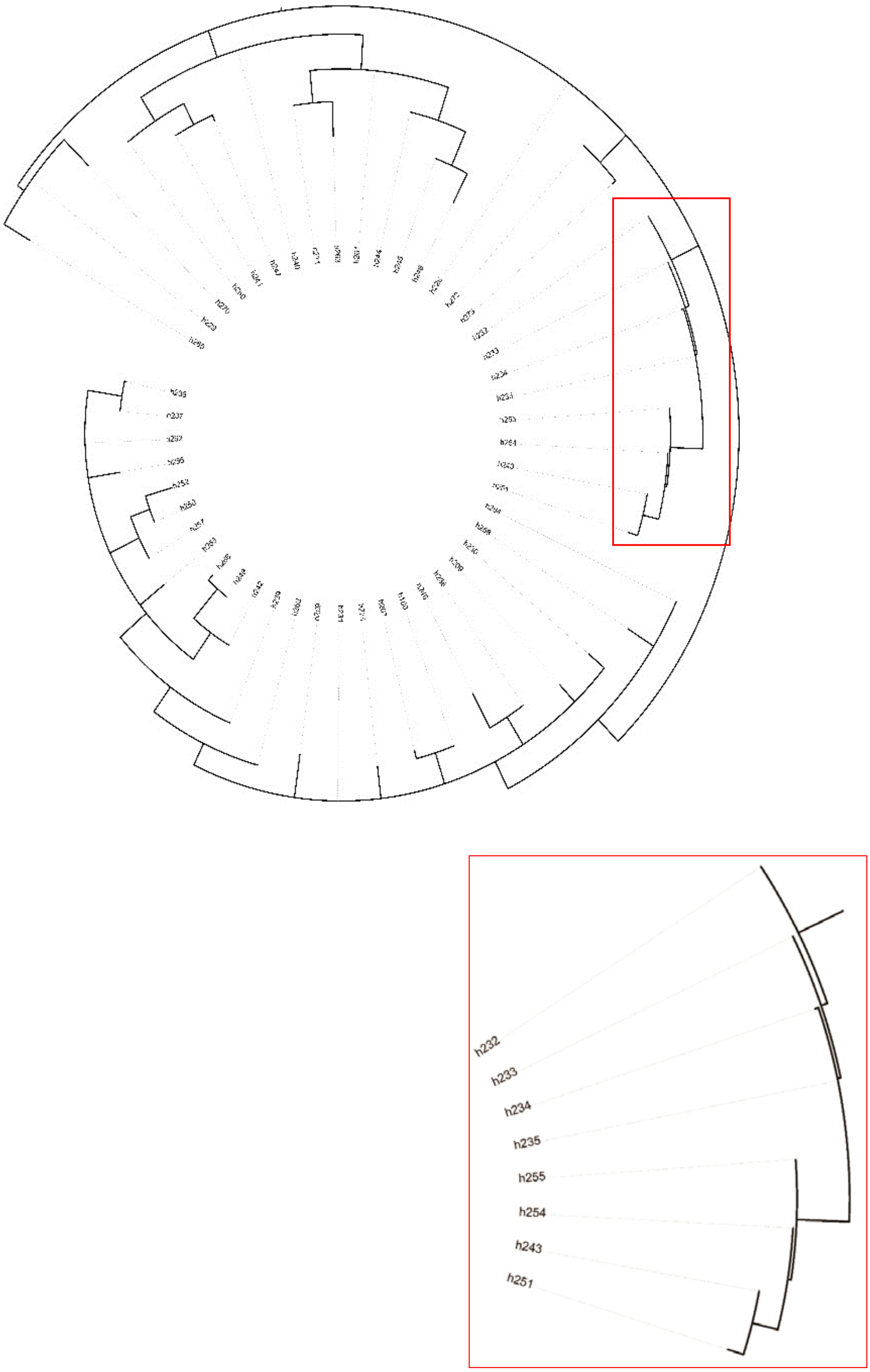
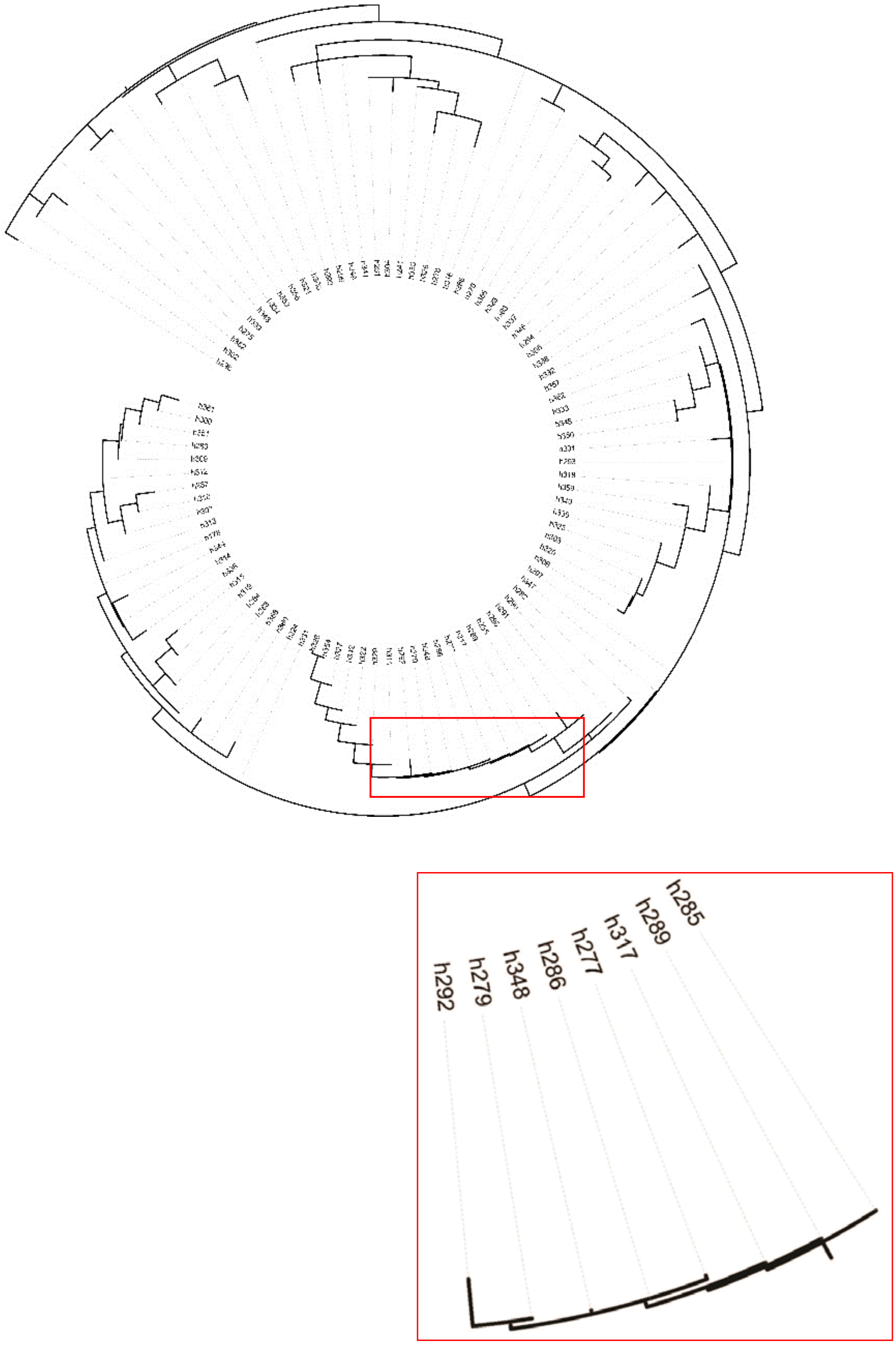


Figure 3.12: Haplotype network in YDN.


h292
h279
h348
h286
h277
h317
h289
h285