Inferring histories of adaptive divergence with gene flow: genetic, demographic and geographic effects

by

Qixin He

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in the University of Michigan
2015

Doctoral Committee:

Professor L. Lacey Knowles, Chair
Assistant Professor Timothy Y. James
Professor Mercedes Pascual
Professor Mark L. Wilson

2015

Dedication

To my beloved grandpa.

Acknowledgments

This work would have been impossible without the help of many people.

First and foremost, I am very thankful to my advisor, L. Lacey Knowles, for all the help and support. She has inspired me in seeking out challenging research projects, more ingenious ways to tackle difficulties, and devoted patience in developing my scientific writing. She always believed in my potential to work on challenging problems, which helped strengthen my confidence to continue my research career. She was instrumental in creating a harmonious and synergetic atmosphere in the lab group, in which we have a sense of belonging.

I would also like to thank my doctoral committee, Mercedes Pascual, Mark Wilson and Timothy James, for their help and support. Their insightful comments and suggestions helped me to improve my thesis. The inspiring discussions also helped extend my work to a broader context and accessible to a wider audience.

The Knowles lab has always been a second family to me. They also provided me with tremendous help in research. Many colleagues from the EEB department and other universities have also helped me with my research. I am thankful for the valuable discussions with Tim Connallon, Huateng Huang, Trevor Bedford, Mark Kirkpatrick and Alex Kondrashov for the theoretical formulation of my thesis. I would also like to thank John McCormack, Amanda Zellmer, Dan Edwards, Diego Alvarado-Serrano, Lucy Tran, Pavel Klimov and Haley Lanier in scientific discussions, analysis advice and critical manuscript reading. Furthermore, I would like to thank Antony J. Cornel, Yooksoon Jin, Kevin Njabo, Seraphin Menzepoh in preparing and conducting field work. Kelsey Gibbons, Jen-Pan Huateng, Robert Massatti, Andrea Thomaz, Anna Papadopoulou, Tristan McKnight, Carlos Munoz, Raquel Marchan, Pamela Murillo and Mark Christie have been extremely helpful in providing critiques on writing, research proposals, and presentations.

I would like to thank the EEB staff, especially Norah Daugherty, Robbin Murrell, Jane Sullivan, Gale Kuhnlein, John Megahan for administrative assistance.

My family has always been very supportive of my choices and ambitions. My grandfather was the inspirational figure for me to pursue research, a dream which he bestowed upon me to finish as he couldn't complete it due to political climate. He taught me the

value of pursuing the scientific truth. I will never forget the memory of him lying on a ward bed, teaching me organic chemistry days before he succumbed to pneumonia. My parents always encouraged me to explore every aspects of life and decide my future on my own. Special thanks to my five-year roommate in Ann Arbor, Anna Abhilasha, and my boyfriend, Kalyan Nadella, for their unconditional support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

LIST OF ABBREVIATIONS

LD linkage disequilibrium

SNP single nucleotide polymorphism

ENM ecological niche modeling

ABC approximate bayesian computation

PCA principle component analyses

DAPC discriminant analysis of principle components

RFLP restriction fragment length polymorphism

Ne effective population size

SFS site frequency spectrum

ABSTRACT

Inferring histories of adaptive divergence with gene flow: genetic, demographic and geographic effects

by

Qixin He

Chair: Professor L. Lacey Knowles

As genomic data is increasingly available even for non-model organisms, the traditional boundaries among fields such as phylogenetics, phylogeography and genetics of adaptation are disappearing. This thesis provides a synthetic framework for studying ecological genomics, which considers selective processes (such as adaptation to new niches) and neutral processes (such as population size changes due to environmental shifts) simultaneously. Conventionally, studies that look for targets of selection on a genome assume a simple demographic model without validations from the species' ecological or phylogeographic histories. The work demonstrates that one cannot reliably identify selection unless realistic demographic histories are inferred for the species or even a specific genomic region. In particular, I investigate the evolutionary history of large polymorphic inversions in *Anopheles gambiae*, which maintains adaptive divergence among ecologically divergent populations. By modeling the unique origin and introgression histories of each inversion, I am able to identify target regions of selection within inversions through training discriminant functions with pure drift versus selection simulations. The thesis also extends the existing theory of local adaptation model via chromosomal inversions to consider the source

of inversion variation, as well as evaluates the likelihood of such adaptations under different parameter spaces. The findings are particularly important for understanding mosaic genomic evolution in the early stages of speciation, where accumulation of divergence is dampened by gene flow. Finally, I examine how historical events, such as habitat contractions or recolonization, influence current genetic pattern and the application of spatially-explicit demographic modeling under Approximate Bayesian Computation statistics to distinguish different phylogeographic scenarios. The work represents a flexible framework for researchers to translate dynamic phylogeographic hypotheses into testable coalescent models by integrating all the available information of the species, such as distribution records, habitat preference, paleo-climate models, and competition between species. In general, with the amount of information as well as inherent heterogeneity of genomic data, this thesis contributes to the ongoing paradigm shift from studying separate evolutionary processes towards a holistic analysis of the interactions of selective and neutral processes under a rigorous statistical framework.

CHAPTER 1

Introduction

Technology advances have fostered an exciting era of a burgeoning field: ecological genomics (Ekblom and Galindo, 2010). Six years ago, most phylogeographic studies relied on developing microsatellites or mitochondrial DNA for inferring genetic structures and demographic histories. During the following years, as costs for individual reads plunged thanks to next-generation sequencing technologies, pioneering studies on developing genomic sequencing libraries for non-model organisms emerged (Baird et al., 2008; Peterson et al., 2012). Today, generating genomic libraries has become a common practice for a doctorate study.

As with the big shift in the magnitudes of data analysis, the traditional boundaries among fields are also disappearing, in particular, phylogenetics (inter-species evolutionary relationships) versus population genetics (intra-species evolutionary relationships), phylogeography (historical processes that shape contemporary geographical distribution and relatedness among populations) versus genetics of adaptation (targeting specific genes or genetic mechanisms that confer local adaptations). Not surprisingly, population genomic data collected from carefully designed experiments might be able to decipher demographic history and relatedness among populations/species, test alternative hypotheses of historical processes and identify loci involved in environmental adaptation at the same time. This paradigm shift, however, is also required by the complex nature of big data so that these seemingly separate questions cannot be answered correctly without the consideration of

each other. Unlike traditional multilocus studies that focus on unlinked neutral genes/regions, genomic data has a much higher heterogeneity in terms of linkage, mutational scenarios, and selection, which poses new challenges for data analyses. For example, most species tree estimation methods assume neutral evolution. A new intra-polymorphism-aware phylogenetic model (De Maio et al., 2013) using polymorphism and divergence data together can account for incomplete lineage sorting as well as estimate mutation and fixation biases caused by selection on different genomic regions. Phylogeography is usually estimated from neutrally-evolving loci (e.g., microsatellites) with independent histories. Genomic data might violate both assumptions. In extreme cases, signals of genetic structures might be dominated by SNPs/loci from tightly linked regions (e.g., inversions) that do not actually correspond to the neutral evolution history (see Chapter 2). On the other hand, the small portion of loci that are under selection might be of high interest. Without considering specific demographic histories, genomic scans for detecting loci involved in ecological adaptation often suffer from high false positive rates (Lotterhos and Whitlock, 2014).

My thesis aims to answer questions regarding ecological genomics with new methodologies and theories. In Chapter 2, I developed a new approach to detect selection within inversions with the aid of inversion-specific demographic histories. In Chapter 3, I developed a theory to evaluate the importance of standing variation versus new mutations in adaptive divergence aided by polymorphic inversions. In Chapter 4, I synthesized a new approach, iDDC, to test support for different phylogeographic scenarios by integrating distributional, demographic, and coalescent models that generate predictions for species-specific patterns of genetic variation. In the last chapter, I summarize the basic findings, limitations of the methods and future directions. Although the thesis may methodologically oriented, the development of new methods were inspired by the specific research goals of my study systems, which are briefly introduced in the following sections.

1.1 Maintenance of adaptation via chromosomal inversions

How fast can adaptive divergence occur upon environmental change or range expansion? Past population genetics studies have thoroughly explored the two most important factors, sources of variation and the probability of fixation of favorable mutations, theoretically (e.g., Ewens, 2004; Fisher, 1930; Kimura, 1983; Orr, 1998) and empirically (e.g., Bradshaw et al., 1995; Colosimo et al., 2005; Karasov et al., 2010). These, however, are not the sole factors because when adaptation requires two or more genes, maintenance of such haplotypes of favorable alleles will be hard when gene flow is still common at the early stage of divergence (Nosil et al., 2009). Under such scenario, any mechanism that can protect the co-adapted genotype from shuffling with maladaptive alleles in recombination will be preferred in adaptation (Yeaman, 2013; Yeaman and Whitlock, 2011).

Chromosomal inversions are rearrangements of a chromosome segment in which it is entirely reversed from end to end. This will cause disorder when recombination occurs between heterokaryotypes (i.e., homologous chromosome pairs which have alternative arrangements) so that recombination is largely suppressed between heterokaryotypes while it is free between homokaryotypes. The traditional view of chromosomal speciation posits that when different inversions get established in parapatric populations, the inverted region will become a hotspot for accumulating positively selected genes and speciation genes (Fig. 1.1a) because recombination inside the region is drastically reduced between hybrids (Noor et al., 2001; Rieseberg, 2001). A more recent view argues the reverse. Inversions can capture alleles that are already locally adapted, but are maladaptive among populations that inhabit ecologically dissimilar habitats (Fig. 1.1b). In this later case, the selective advantage of an inversion is generated by recombination load from the maladapted gene flow (Kirkpatrick and Barton, 2006; Manoukis et al., 2008). This hypothesis (Fig. 1.1b) revived studies of the evolution of chromosomal inversions in recent years because it has no presumption of peculiar genetic interactions inside inverted regions and can be readily applied to many common scenarios.

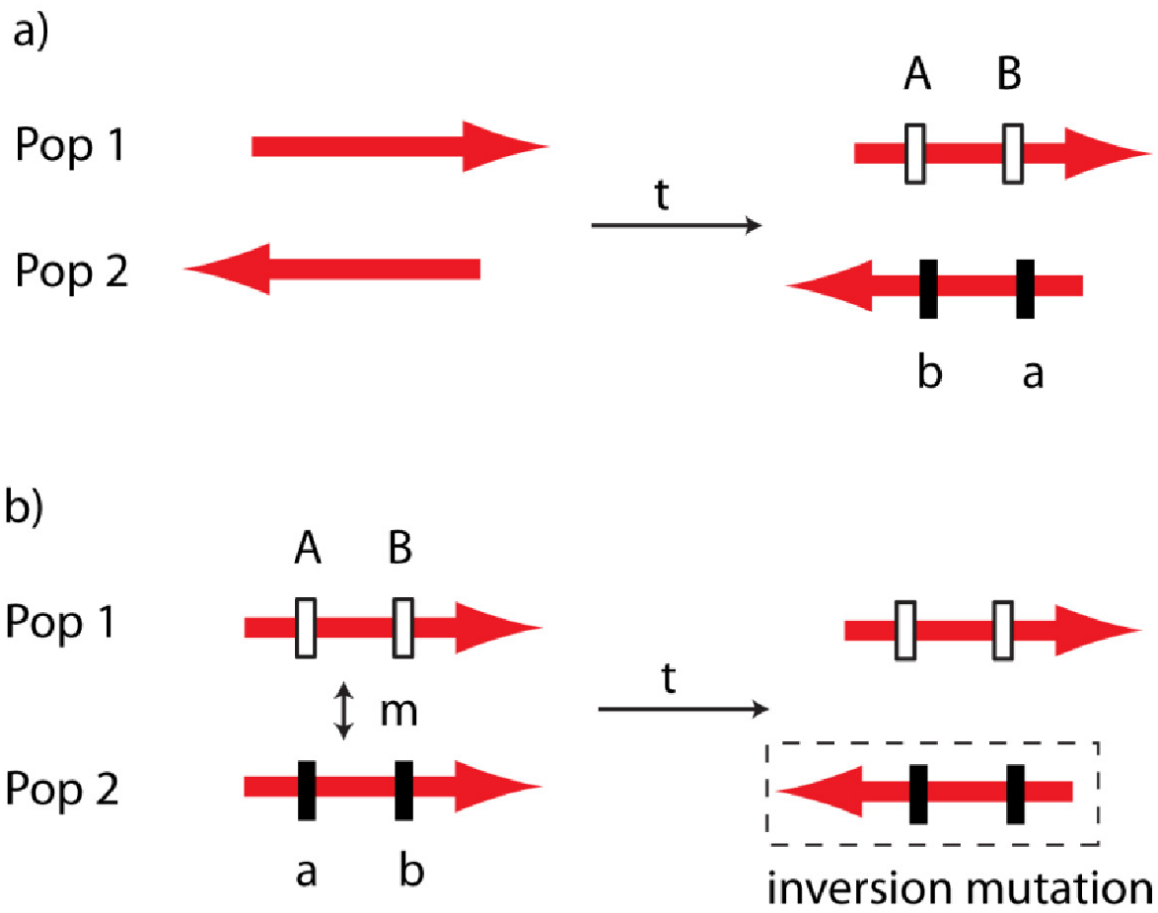

Figure 1.1: schematic representation of maintenance of inversion polymorphism in populations. Direction of red arrowheads indicate alternative chromosomal rearrangements. a) inversions are fixed in different populations. They become hotspot for accumulation of adaptive/speciation loci; b) populations accumulated adaptive loci. Upon secondary contact, inversion mutation occurred to protect co-adaptive genotypes.

The local adaptation model for inversions has found support from empirical evidence, where chromosomal inversions promoted divergence by creating physical linkage among locally adapted loci that are responsible for alternative complex traits such as wing patterns in Batesian mimicry (Joron et al., 2011), diapause timing (Feder et al., 2003) and annual/perennial life-history shift (Lowry and Willis, 2010). I used this model as the basis of Chapter 2 and Chapter 3. In Chapter 2, I developed a new approach to detect adaptive loci inside inversions that give the selective advantage of inversions. In Chapter 3, I focused on the sources of inversion variation and the likelihood of such adaptation under different

genetic or demographic conditions.

1.2 Polymorphic inversions in *Anopheles gambiae* populations

Anopheles gambiae is widely distributed across sub-Saharan Africa near human dwelling sites (Coluzzi et al., 1979; della Torre et al., 2005). The mosquito's extensive distribution throughout a heterogeneous environment (Fig. 1.2a) might indicate local divergence and strong population structures within the species. However, neutral markers (e.g., Lehmann et al., 1997, 2000) usually showed no apparent population structures or failed to reject simple isolation-by-distance test, except for the populations distributed across the Great Rift Valley in East Africa (Lehmann et al., 1998, 1999; Zhong et al., 2006), which acts as a strong barrier to gene flow. Estimations of effective population size from genetic polymorphism data of local demes are usually large (Donnelly et al., 2002; Lehmann et al., 1997), indicating that gene flow among demes is high and populations within a large area can be viewed as being panmictic. *An. gambiae* also shares common genetic variation with the sister species *An. arabiensis* (Besansky et al., 1997), and introgression between these two species is not rare (Donnelly et al., 2004; della Torre et al., 2002). These facts together indicate that its speciation and subsequent rapid adaptation were fairly recent so that retention of ancestral polymorphism is prevalent. In contrast, genetic markers that do exhibit divergence among populations are usually those that reside near polymorphic inversions or centromere-proximate regions (Lanzaro et al., 1998; Oliveira et al., 2008; Onyabe and Conn, 2001; Temu and Yan, 2005; Turner, 2005; Turner and Hahn, 2007; White et al., 2010). Local adaptation to different ecotypes of the species, yet extensive intraspecific gene flow, produced accentuated divergence in some part of the genome, dubbed "genomic islands", while little differentiation in other parts, which fits the hotly debated "divergence with gene flow" model (Nosil et al., 2009).

Anopheles gambiae giles have a super diverse inversion system. Within *An. gambiae s.s.*, there are seven major polymorphic inversions (Coluzzi et al., 2002). All of the polymorphic inversions are distributed on the second chromosome (Fig. 1.2b). The left arm harbors only one large inversion 2La, which had its origin from An. arabiensis (Besansky et al., 2003). The right arm has six polymorphic inversions (2Rb, 2Rc, 2Ru, 2Rd, 2Rbk, 2Rj), among which 2Ru, 2Rd and 2Rbk are partially overlapping with each other (Coluzzi et al., 2002). The ecological role of chromosomal inversions was first documented by Coluzzi et al. (1985), who observed that the frequency of 2La increases linearly with the aridity level. Similar trends were observed in 2Rb (Simard et al., 2009). It has been shown that different chromosomal forms do have unique environmental niches based on large association studies between distribution of chromosomal forms and environmental variables (Bayoh et al., 2001; Costantini et al., 2009; Simard et al., 2009; Yawson et al., 2007).

Theories have shown that old polymorphic inversions must be maintained by divergent selection (Guerrero et al., 2012). Therefore, how can we detect adaptive loci in inversions that show stable clines in *An. gambiae*? And how prevalent are they in inversions? In Chapter 2, I analyzed genomic data from wild collected *An. gambiae* populations to answer these questions.

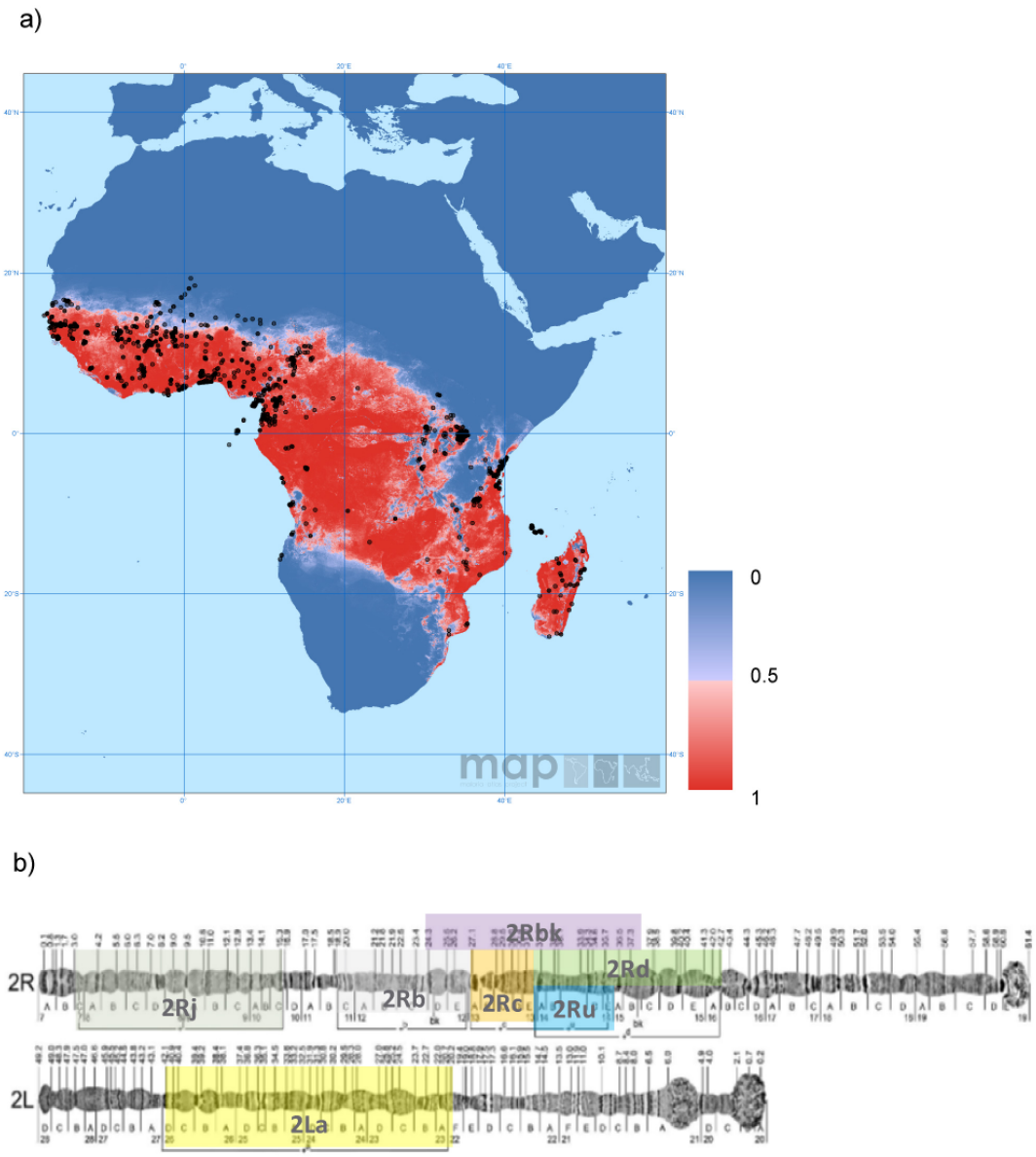

Figure 1.2: Distribution and inversions of *Anopheles gambiae*.a) distribution of *Anopheles gambiae s.s.* (adapted from Sinka et al. 2010); b) polymorphic inversions on chromosome 2 (adapted from George et al. 2010).

1.3 iDDC as a tool for evaluating phylogeographic hypotheses

Population genetic structures can be a representative of contemporary habitat suitability and connectivity, a remnant of historical processes or both (Knowles and Alvarado-

Serrano, 2010). Thus, regression-based tests (e.g., mantel test, dbRDA, CCA) that directly correlate genetic divergence among populations or individuals with contemporary environmental factors or geographical barriers may overlook or misidentify the impact of historical processes, because temporal shifts of habitats (Carnaval et al., 2009) and the demographic processes of range contraction, expansion and recolonization leave signatures in intraspecific genetic diversity and divergence patterns (Excoffier et al., 2009a). Spatially explicit demographic modeling, however, can directly model these processes and generate genetic markers for empirical comparisons (Currat and Excoffier, 2004). This approach is extended from standard coalescent models (Kingman, 1982) in that all the coalescent events occur within demes and movements of individuals across demes are tracked based on local carrying capacities and migration rates (Fig. 1.3).

Spatially explicit demographic modeling can be used to expand the tradition of intuiting qualitative phylogeographic hypotheses from ecological niche models, ENMs (reviewed in Knowles, 2009) to incorporating quantitative information about variation in the habitat suitabilities across space and time. As a consequence, predicted patterns of genetic variation are species-specific, reflecting the interaction between the physical environment and biological parameters (e.g., local population sizes and migration rates) that determines the level and pattern of gene flow across the landscape (see Brown and Knowles, 2012; Knowles, 2009; Morgan et al., 2011). In Chapter 4, I designed a novel approach, iDDC modeling, that integrates distributional, demographic, and coalescent models to generate predictions for species-specific patterns of genetic variation. These simulations can then be incorporated into a statistical testing framework, Approximate Bayesian Computation ABC (Beaumont et al., 2002), to select for scenarios that fit empirical data the best. Additionally, I assessed the quality of parameter estimates using pseudo-observed datasets (pods) (see Bertorelle et al., 2010; Robert et al., 2011). The methods proposed here can be generally applied to different biological systems that have experienced non-static demographic history.

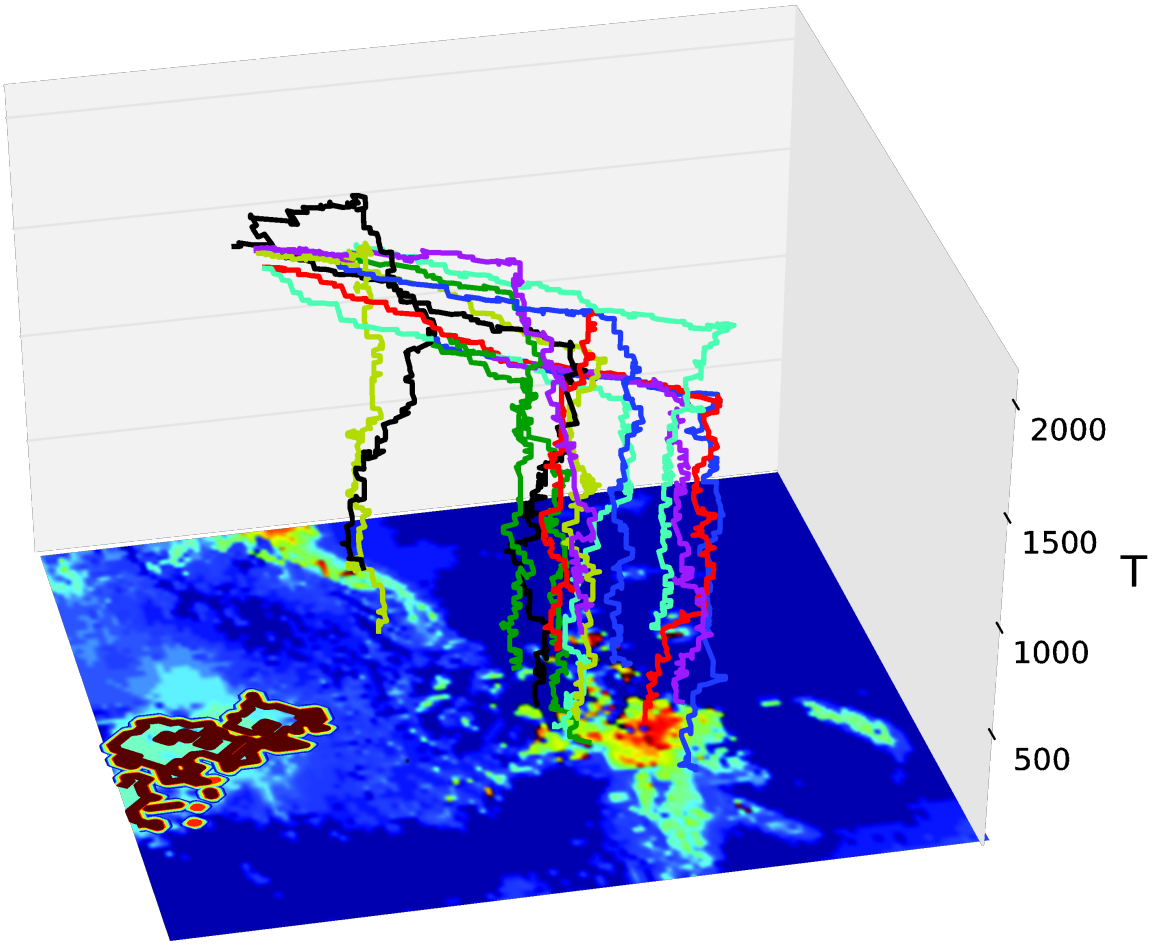

Figure 1.3: Spatially explicit coalescent history. Different lineages are indicated by different colors. Colors on the 2-D surface represent suitabilities across the landscape, with blue denoting less suitable and red, more suitable. Lineages travel across demes and coalesce at different times backward.

CHAPTER 2

Generating neutral expectations for specific genomic regions increase the detection of selection targets and reveal different roles of polymorphic inversions in local adaptation of *Anopheles gambiae*

2.1 Abstract

Chromosomal inversions are important structural changes that may facilitate divergent selection when they capture co-adaptive loci. However, identifying selection targets within inversions can be challenging because of their high degree of differentiation between heterokaryotypes (hence, reducing the power of detection with methods like F_{ST}-based outliers). Likewise, neutral expectations built from collinear regions cannot be used when detecting selection in inversions because of differences in the demographic histories of inverted compared to collinear regions. Here, we developed a new approach that uses discriminant functions to classify loci that are under selection (or drift) informed from inversion-specific demographic histories. We demonstrate that this approach has much higher power than traditional F_{ST} outlier analysis and analyze data we collected in a classic

Dipteran species with polymorphic inversion clines - *Anopheles gambiae* s. str, a malaria vector species from sub-Saharan Africa. Polymorphic inversions within or between populations in the species are thought to contribute to rapid adaptation to different microhabitats in this widely distributed species. We collected wild mosquito samples from transitional ecozones between forest and savanna in Cameroon and identified genome-wide SNPs using barcoded Rad sequencing. Contrary to minimal geographic structure among populations in collinear regions, individuals are clustered by SNPs from 2La and 2Rb, two polymorphic inversions with inverted chromosomes predominating in dry habitats and the standard chromosomes in wet habitats. Although both inverted forms were introgressed from a sister species, *Anopheles arabiensis*, the two inversions showed very different origin from demographic reconstruction. The divergence time between the non-inverted and inverted forms of 2La dated back before the species divergence, whereas 2Rb originated within *Anopheles arabiensis*. Predicting the status of selection or drift based on discriminant functions built from simulated training sets showed that in standard form ($2L+^{a}$) that associates with wet habitat exhibits much more selection signatures than inverted form (2La) that associates with dry habitat.

2.2 Introduction

Detecting the signature of natural selection from molecular data has been a central focus of geneticists, not only because of the deeper understanding of molecular evolution it brought, but also because of its potential in revealing important functional information (Nielsen et al., 2007). With the advances in sequencing technologies, many new methods for identifying selected sites and regions of the genomes are burgeoning now. These approaches capture different attributes of the signature of selection, including patterns of exceptional long haplotypes (Sabeti et al., 2002, 2007; Voight et al., 2006), surges in linkage desequilibirum LD (Kim and Nielsen, 2004; Wang et al., 2006), skewed site frequency

spectrums (Carlson et al., 2005); Nielsen et al. (2005); Ronen et al. (2013). But when the goal is to identify selection under spatially divergent selection among populations, many studies rely on F_{ST} outlier tests (Antao et al., 2008; Beaumont and Balding, 2004; Bonhomme et al., 2010; Foll and Gaggiotti, 2008; Gunther and Coop, 2013) to scan for regions with exceptionally high divergence between ecologically divergent populations, especially when whole genomic resources are not available for detailed estimates of linkage disequilibrium (LD) or haplotypes. However, the power of such methods becomes inherently limited when adaptive mutations occur in regions with reduced recombination, such as with chromosomal inversions, which is especially problematic for studying adaptive divergence given the important role inversions play through the maintenance of co-adaptive genotypes (Kirkpatrick and Barton, 2006; Yeaman, 2013).

Identifying selection targets within alternative inversions can be challenging because of two characteristics that distinguish them from collinear regions (i.e., genomic regions without inversions). First, in genomic regions with inversions there is an overall high divergence between the inverted and non-inverted chromosomal arrangement (e.g., Cheng et al., 2012), which diminishes the power to detect selection (Beaumont, 2005; Lotterhos and Whitlock, 2014). Second, given that the demographic dynamics vary across the genome - that is, the genome is mosaic, with certain regions for example experiencing more or less gene flow depending on whether it is captured by an inversion, relying on a single neutral parameterization (either with simulations under the island-model used in FDIST2 (Beaumont and Nichols, 1996) or the coancestry matrix in FLK (Bonhomme et al., 2010) will necessarily be a mis-specification for expected patterns of divergence that can further exacerbate the difficulties with detecting targets of selection.

In this study, we develop a new generic approach for locating a selection signature when population divergence (or speciation) is promoted by inversions that limit the power of traditional tests for selection. By first estimating region specific demographic history, we are able to do demography-adjusted selection tests (see also Rafajlovic et al., 2014).

We then build a discriminant function using combinations of summary statistics from sequences simulated under neutral and selected scenario. Empirical sites were assigned into neutral or selection scenarios through predictions from the discriminant function. We apply the newly developed approach to *Anopheles gambiae*, which like other Dipeteran species, has a long history of research on inversion polymorphisms (Coluzzi et al., 1979). As a widespread species with large population size, it lacks in general significant population structure, except for the geographic structure observed at genomic regions characterized by seven commonly segregating inversions on the chromosome 2 (Czeher et al., 2010; Lanzaro et al., 1998; Lehmann et al., 1998). Here we focus on two large inversions, 2La and 2Rb, that exhibit stable clines, but which defy detection of regions/genes within the genomic regions associated with the inversions because of high LD and F_{ST} across the regions (White et al., 2007a). The high frequency of 2La, which spans 21.4 Mb, in dry geographic areas (savannas) compared to wet areas (forests) (White et al., 2007b), and a predictable cycle in the frequency of 2La during dry and wet seasons identifies its role in adaptive divergence. Similar trends observed in 2Rb (Simard et al., 2009), a 7Mb inversion, suggest that this genomic region also contains sites contributing to adaptive divergence. In addition to the general difficulties with identifying the targets of selection within inverted regions discussed above, another potential complication is tied to the origins of these two inversions. Both inversions are thought to represent examples of introgression from a sister species, *An. arabiensis* (Besansky et al., 2003), which is sympatric with *An. gambiae* in arid savanna areas (Neafsey et al., 2010; White et al., 2009). This suggests that the inversions might be old enough to have very different genetic background, which exacerbates the high divergence problem between heterokaryotypes aforementioned. Nevertheless, despite these challenges, as our analyses demonstrate, we are able to not only identify approximate targets of selection in the genome, but also to make statistical statements about the selective history, and specifically what lineage underwent selective divergence with respect to the ancestral state. We discuss the general applicability of our procedure for tests of selection

in other taxa and specific genomic regions, even those regions that have vastly different history than the rest of the genome because of the mosaic nature of the genome.

2.3 Results

We collected genomic data in 259 *An. gambiae* individuals and 8 *An. arabiensis* from six sites in Cameroon along a gradient of wet to dry habitats (see Appendix A materials for specimen identification results). From individually-barcoded double digest Radseq libraries (Peterson et al., 2012), and two lanes of 100bp paired-end sequencing on Illumina HiSeq2000 platform, 25,966 loci (i.e., RADtag from the genomic prep) were mapped onto Chromosome 2, 3 and the X, after filtering for ambiguously mapped reads and loci with low coverage per sample or low presence across samples. Although we recognize that this represents a small proportion of the genome (about 1%), the goal of this manuscript is to demonstrate the promise of the approach for detecting approximate targets of selection, as opposed to providing a full analysis of the proportion of sites under selection within inversions contributing to adaptive divergence, which would require additional sequencing that is beyond the scope of this study.

Our new approach of targeting specific selection within inversion involves several procedures (summarized in Fig. 2.1a). First we estimated divergence time and introgression rate between the two species from collinear regions (Fig. 2.1b) using site frequency spectrum (Excoffier et al., 2013). We then estimated inversion specific parameters, such as gene flux rate between alternative inversions and the age of inversion mutations (Fig. 2.1c). This framework was then used for tests of selection against neutral expectations that are region specific by discriminant functions based on summary statistics estimated from empirical versus simulated data informed from inferred demographic history.

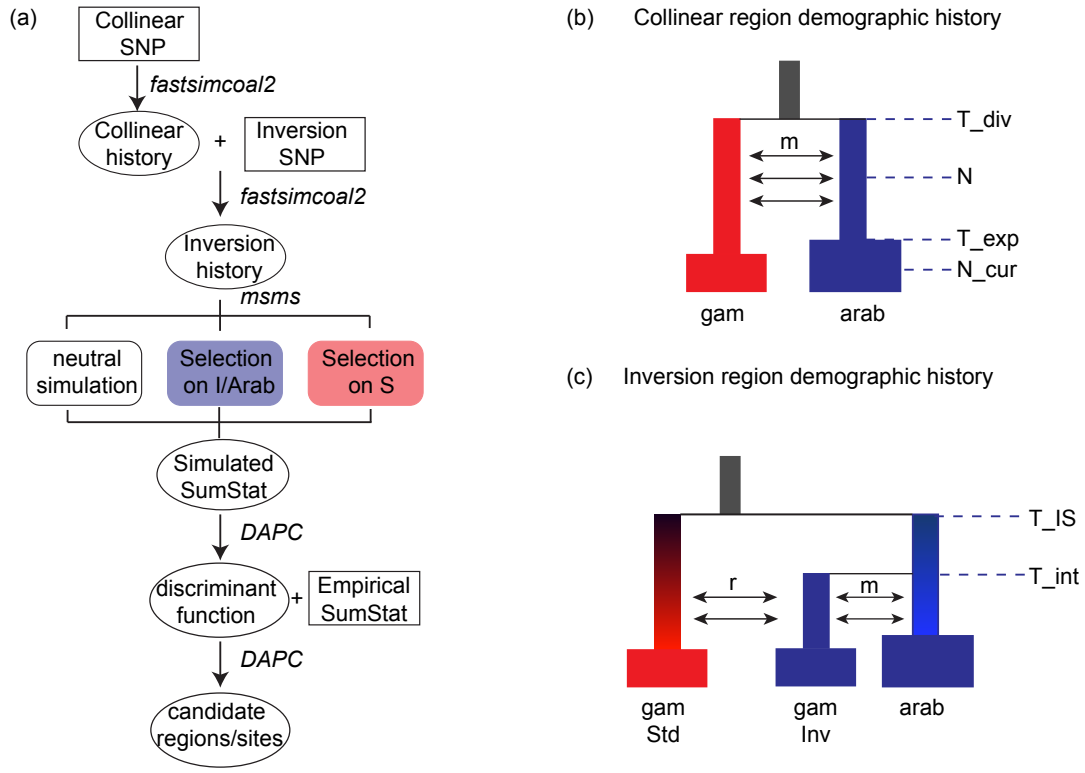

Figure 2.1: Schematic illustration of study design and population demographic scenarios of *Anopheles gambiae* and *An. arabiensis*. a) procedures involved in detection of targets of selection; b) in collinear regions, *An. gambiae* and *An. arabiensis* diverged from a common ancestor at time T_div with a population size Ne. They experienced recent population expension at time T_exp to the current population size N_cur. The two species have constant gene flow since divergence; c) in regions with alternative arrangements, the arrangement that were prevalent in *An. gambiae* (Std) split with the alternative arrangement (Inv) in *An. arabiensis* at time T_IS; at time T_int, Inv introgressed into *An. gambiae* population. Alternative arrangements have reduced recombination rate (r). Inv have a similar gene flow rate with *An. arabiensis* as collinear region.

2.3.1 Establishing a demographic null model for tests of selection

Principal Components Analysis (PCA) and Discriminant Analysis of Principal Components (DAPC) analyses of SNP data showed a lack of population genetic structure in the species in the collinear regions. Specifically, the PCA showed no apparent geographic structure among the six sampled populations (Fig. A.3) and one group (K = 1) received the highest support in the DAPC analysis (see supplementary text for details about tests

of geographic structure). Therefore individuals were pooled across populations to estimate a demographic history that could be used to establish a null expectation for patterns of genomic divergence under a model of drift. Specific demographic histories were inferred for different genomic regions from the region-specific site frequency spectrum (SFS) under a composite-likelihood approach implemented in fastsimcoal2 (Excoffier et al., 2013). Specifically, for collinear regions a demographic model was inferred that included parameters for the time of species divergence (T_div) between *An. gambiae* and *An. arabiensis*, gene flow (m) between the two species, as well as population expansion (N_cur, T_exp) (Fig. 2.1b). We used the population size of *An. gambiae* as a fixed parameter to estimate other free parameters (Fig. 2.1b, c); this was set to ∼Ne = 750,000 using a mutation rate of 3.5E-9 per base per generation (estimation from Drosophila resequencing, Keightley et al., 2009) given that *An. gambiae* was estimated to have a nucleotide diversity (π) of 0.01024 ± 4.0E-5 from all Radtags. Estimations showed that the two species diverged fairly recently around 87K years ago (∼1.4Ne generations ago, Table 2.1; assuming 12 generations a year (Lehmann et al., 1998)) with small but constant genetic exchange (on the order of 1E-7 to 1E-8, which is about exchanging 0.01 to 0.2 individuals per generation).

In contrast to the collinear regions, SNPs from 2La (2L: 20524058-42165532) and 2Rb (2R: 19023925-26758676) were distributed into one of three clusters in the PCAs (with the first PC explaining 20.3% and 11.9% of the total variance in 2La and 2Rb, respectively; Fig. A.4 a,d). The results from the DAPC also supported K=3 as the most likely number of genetic clusters (Fig. A.4 b, e). Individuals that form the three clusters identified from these analyses correspond to the three genotypes associated with the inverted genomic region: namely, the inverted homokaryotypes (referred to as I/I hereafter), the heterokaryotypes (I/S), and the standard (i.e., non-inverted) homokaryotypes (S/S) based on comparison with molecular karyotyping results (Fig. A.4c,f). Separate demographic histories were therefore inferred for the different genomic arrangements (i.e., inverted versus standard chromosomal regions) for both the 2La and 2Rb regions, with a recombination parameter to accommodate

Table 2.1: Estimations of population genetic and demographic parameters using region specific SFS implemented in fastsimcoal2.

Regions	Parameters	Point estimates	Confidence Interval	
Collinear	current population size (arabiensis)	4,202,600	2,253,300	66,070,400
	population size before expansion	677,800	608,700	813,500
	introgression rate	1.75E-07	1.52E-08	2.70E-07
	expansion time	186,800	136,900	228,700
	species divergence time	1,052,500	824,600	1,240,400
2La	ancestral population size	1,028,000	1,006,000	1,267,100
	Gene flux rate	7.99E-08	9.25E-08	1.08E-07
	bottleneck population size	29,900	26,500	101,500
	time to bottleneck	339,900	319,600	351,100
	bottleneck duration	300	100	3,200
	divergence time	2,353,800	2,236,100	2,568,100
2Rb	ancestral population size	980,800	1,041,600	1,148,900
	Gene flux rate	4.79E-07	4.64E-07	5.15E-07
	bottleneck population size	8,400	3,500	29,400
	time to bottleneck	552,200	513,700	580,000
	bottleneck duration	39,600	19,400	87,700
	divergence time	591,800	570,100	625,200

occasional gene flux between alternative karyotypes (Fig. 2.1c).

Inversions were modeled as introgressed from *An. arabiensis* into *An. gambiae* after the species' divergence time (see demographic model in Fig. 2.1c). The age of inversion mutation and their introgression time will strongly influence the baseline divergence expected for detecting selection. Interestingly, the coalescent time of S and I (T_IS, Fig. 2.1c) of 2La was around 3Ne generations ago (196K years, Table 2.1), which is much more ancient than species divergence time (T_div). On the contrary, T_IS of 2Rb is similar to T_div. Despite being old, 2La was also found to be introgressed more recently ($\sim$0.5 N) into *An. gambiae* than 2Rb ($\sim$0.8N), but with shorter and less dramatic bottleneck than 2Rb (Table 2.1).

2.3.2 Signature of selection within inversions

A modified F_{ST} outlier scan based on inversion specific demographic history was first performed. Due to structural constraints of inversions, double recombination rates are

higher in the center of an inversion versus breaking points. We therefore divided inversion regions into 150kb segments and adjusted recombination rates from our average estimates based on local 2Mb F_{ST} estimations (details see methods) and then simulated 1000 neutral cases per region. We found that because the divergence of I and S of 2La was really old, F_{ST} estimations between I and S for neutrally evolved DNA segments can be really high (see the 95 percentile and 99 percentile of simulation estimates in Fig. 2.2a). The power of differentiating selection from neutral genes are really impaired for F_{ST} estimates because 95 percentile of the simulations can reach $\sim$0.8. The distribution of empirical F_{ST} estimates of 2La is fairly flat compared to collinear regions where most F_{ST} estimations are concentrated around 0 (Fig. 2.2c). This is less exacerbated in younger 2Rb (Fig. 2.2b). Yet, the empirical F_{ST} distribution has much longer tail than the collinear F_{ST} distribution (Fig. 2.2d).

Since outlier analysis based solely on F_{ST} measures have limited power to differentiate selection from drift, we designed a new approach to make use of all summary statistics. 1000 simulations of 50kb sequences containing selected locus were ran for each of the three scenarios: 1) neutral; 2) selection occurs on sites associated with S; 3) selection occurs on sites associated with I and *Anopheles arabiensis* (Fig. 2.1a). Discriminant function (DAPC) were built to differentiate three scenarios from summary statistics (Fig. 2.3a). In order to minimize random effects from individual Radtags, average summary statistics of random loci sampled across the simulated 50kb regions at a similar density as empirical data were calculated to build the discriminant function. The power (true positives) for three scenarios is above 0.9 (Table 2). Moreover, in contrast to the problems with false positive of putatively targets of selection, such cases represented a minority of incorrect assignments (Table 2), making it a more conservative test (i.e., more failures involved the incorrect assignment of selected sites to neutral sites, rather than the other way around). Interestingly, heterozygosity (H) and π contributed more of the variation in differentiating each scenario compared to F_{ST} (Fig. 2.3d).

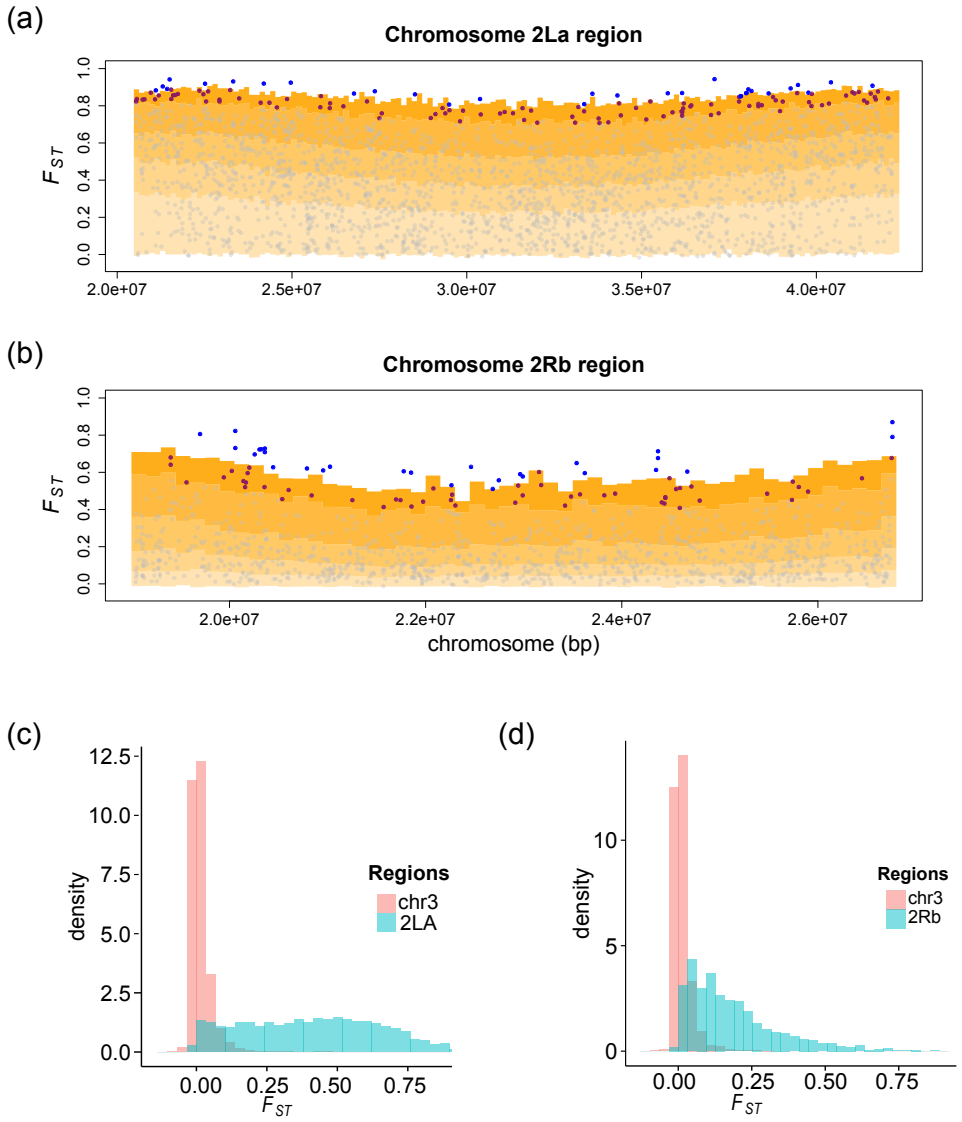

Figure 2.2: Outlier analysis of inverted region scan. a) and b), dots represent FST measures of each Radtag locus between Std and Inv chromosomes in *An. gambiae* populations along the region. Shades of yellow show quantiles of 25%, 50%, 75%, 95%, 99% respectively, of simulated values of divergence measures under reconstructed demographic histories with region-adjusted recombination rate. c) and d) empirical distributions of FST between individuals with Std and Inv chromosomes on inverted region (green) or collinear region (red). a) and c), 2La regions; b) and d), 2Rb regions.

After empirical estimates of summary statistics were transformed into discriminant scores and assignment of scenarios were predicted, we identified regions that were sug-

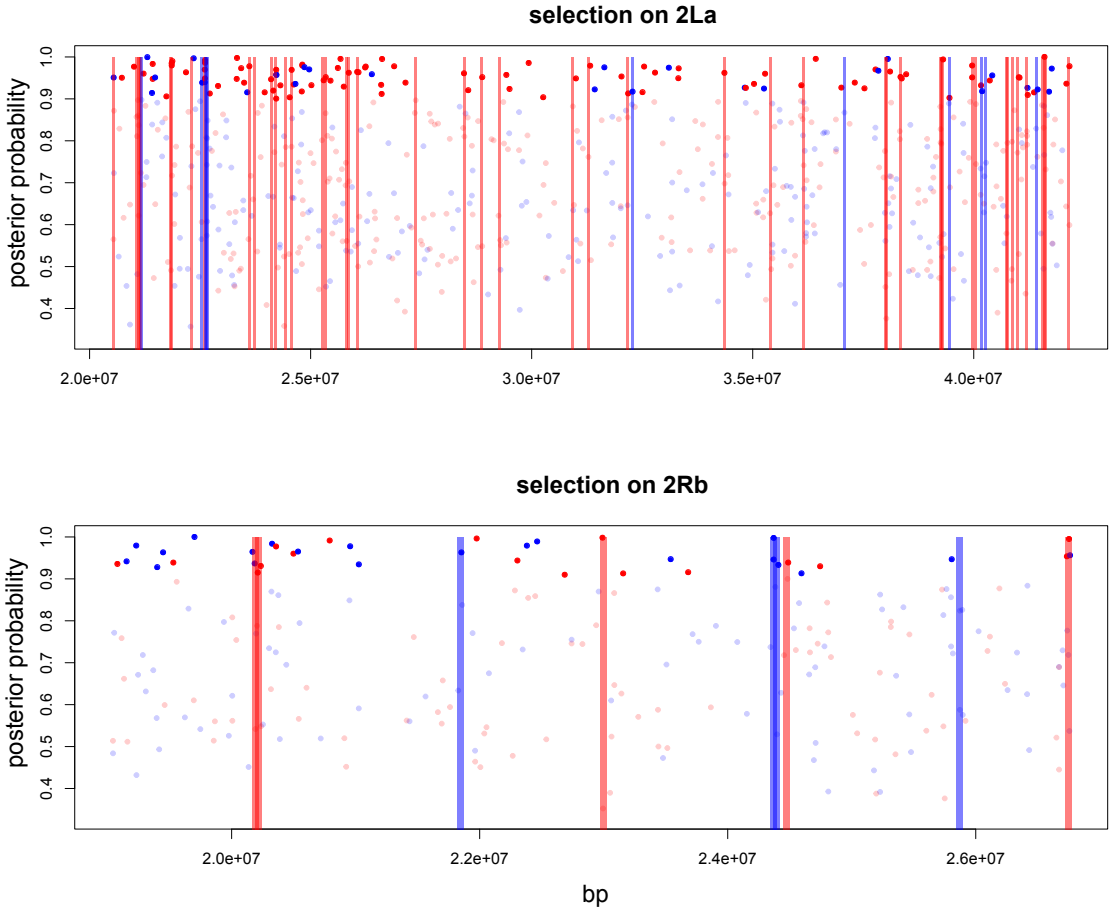

Figure 2.3: Candidate loci and regions under selection. All the dots show Radtag loci that have highest posterior probability to be assigned as either experienced selection in the Std lineage (red) or Inv lineage (blue). Dots with solid color are the ones with larger than 0.9 assignment posterior probability. Bars show regions that have highest posterior probability to be assigned as either experienced selection in the Std lineage (red) or Inv lineage (blue). Width of each bar corresponds to 50kb in our analysis.

Table 2.2: Power of differentiating selection form drift using average summary statistics of loci in a 5kb segment with/without selected locus inside.

		Prediction			
Inversion	True Scenario	Neutral	Selection in I and Arab	selection in S	Correct Assignment % (Power)
2La	Neutral	936	31	33	93.6%
	Selection in I and Arab	65	927	8	92.7%
	selection in S	56	2	942	94.2%
2Rb	Neutral	981	9	10	98.1%
	Selection in I and Arab	72	927	1	92.7%
	selection in S	70	0	930	93.0%

gestive of under selection (regions shown as bars in Fig 3a, b). We found that on 2Rb, four regions contain selected sites associated with I (red bars on Fig 3c), while four contain selection associated with S (blue bars in Fig. 2.3c). These regions span some loci that were identified as F_{ST} outliers, but not all of them. In 2La, we recovered an overwhelmingly higher number of regions identified as being selected in S compared to those associated with I (Fig. 2.3b).

The increased power of applying multiple summary stats and converting them into PCs and transforming into discriminant functions is a good way to detect selection when the expected changes in different statistics can be verbalized but hard to summarize into one statistics. For example, loci/region that were identified as under selection in I will have a lower π_I and π_{Arab} compared to π_S, higher F_{ST} between S and I than that of S and Arab, whereas loci/region that were identified as under selection in S will have a lower π of S compared to I, but higher F_{ST} (S_Arab) than F_{ST} (S_I). By definition, if a single statistic had been used to detect the signature of selection, clearly some cases of selection would have been missed because the signature of selection can manifest itself through a diversity of summary statistics.

2.4 Discussion

In this study, we recovered different demographic histories for collinear and inverted regions between two largely sympatric species, *An. gambiae* and *An. arabiensis*. The collinear region experienced ongoing introgressive hybridization since the species divergence, as well as the recent population expansion (11K to 19K) after the last glacial maximum. In inverted regions, standard arrangements in *An. gambiae* had limited gene flux with inverted arrangements, and the divergence time between two arrangements of 2La dated back before the species divergence. Genomic scan of selection signature based on neutral simulations showed that the age of inversions played a significant role in the power

of detection because of the increasing baseline divergence and the of divergence estimators in neutral sequences in older inversions. Predicting the status of selection or drift based on discriminant functions built from simulated training sets seems to be much more promising than traditional genomic scans.

2.4.1 Age of inversions and its influence on F_{ST} outlier analysis

Our estimation of the older divergence of 2L+^{a} from 2La corroborates a recent karyotype phylogeny for the *Anopheles gambiae* complex (Kamali et al., 2012) which predicted that 2L+^{a} evolved in the ancestor of *An. gambiae* and *An. arabiensis* and alternative karyotypes got fixed in the two species after their divergence. The coalescent time of 2L+^{a} and 2La is estimated to be around 3Ne, which was predicted to be a good age to detect selection from estimating coalescent time between alternative arrangements (Guerrero et al., 2012). However, our F_{ST} outlier analysis showed much less power in 2La region than in the younger 2Rb region. The discrepancy might follow from the fact that coadapted genotypes have evolved after the species divergence and that the introgression age of 2La from *An. arabiensis* into *An. gambiae* is too short compared to its origin time so that gene flux between heterokaryotypes is not frequent enough to reduce the divergence between neutrally evolving regions inside inversions. 2Rb, in which the background divergence is not significantly older than the time of introgression, showed clearer patterns of selection signals of regions with significantly higher divergence.

2.4.2 Detecting selection in regions with reduced recombination rates

The traditional indexes (F_{ST}, D_{XY}) in genomic scan to quantify inter-population differentiation are still the most effective way to detect targets of divergent selection in most cases (e.g., Soria-Carrasco et al. (2014). With the availability of population genomic data, such genomic scan studies have become a common practice (reviewed in Nosil and Feder (2012) and many studies reported regions of elevated divergence, termed genomic islands

(e.g., Harr, 2006; Nadeau et al., 2012; Turner, 2005). While common practice does not accommodate regions with reduced recombination or different demographic histories, we tried to circumvent the overall high divergence problem in inversions by generating null expectations from neutral evolution in selection identification. However, this is not entirely ideal because the detection power is highly diminished when variability of divergence estimators among neutral sequences increases with the age of divergence.

Our second approach which uses discriminant function to predict selection is similar to an ABC parameter estimation process, in which simulations close to observations are retained, and linear functions are built to estimate parameters from PCs transformed from summary statistics (Beaumont et al., 2002; Wegmann et al., 2009). The difference is that we are choosing models (neutral or under selection) instead of estimating parameters because our simulations are fixed with specific parameters estimated from reconstructed demographic histories. This approach has the great advantage of jointly considering all summary statistics together across different arrangements/species instead of relying on F_{ST} measures alone. Since radtag sequences are relatively short ($\sim$100bp), simulated sequences still show a great degree of variation in each scenario (e.g., drift alone, selection in I and selection in S) so that discriminant functions do not differentiate them completely (Fig. 2.3a). Nevertheless, it still gives $\sim$0.8 true positive rate in assigning to correct scenarios. We further filtered out loci that have less than 0.9 posterior probability in the assignment of selection scenarios to be conservative. With this approach, we identified more loci under selection in 2La than 2Rb (Fig. 2.3b, c). More interestingly, the signature overwhelmingly showed more prevalent selection in sites associated with S rather than I. This might not be surprising given 2La's origin history. Recent molecular studies have shown that 2La is the ancestral arrangement that stayed in *Anopheles arabiensis* (Sharakhov et al., 2006), from which 2L+$^{\mathrm{a}}$ arose and got fixed in *Anopheles gambiae*. Therefore, during the species divergence period, 2L+$^{\mathrm{a}}$ might have carried co-adaptive genotypes that got preferentially selected in forest environment and the arrangement got fixed in the species. Later on, when

the species expanded from forest to savanna (Coluzzi et al., 2002), 2La that conferred higher fitness in dryer habitat in sympatric *An. arabiensis* introgressed into *An. gambiae*. Our data suggest that the alternative arrangements have each facilitated either wet or dry habitat adaptation at different stages of the species history. In this study, however, we do not intend to pinpoint the exact genes that are under selection given the relatively low density of Radtag markers ($\sim$10kb between adjacent markers).

Currently, local changes in linkage disequilibrium inside inversions are hard to detect in our data. Yet, LD or extended haplotype tests might be informative with individuals of same arrangements because although recombination between heterokaryotypes is highly reduced, it is not reduced within same karyotypes (LDhat (Auton and McVean, 2007) did not find significant changes in recombination rates within same karyotypes in our data). Therefore, if individually-barcoded whole genome sequences are available, selection signatures such as changes in linkage disequilibrium within the same karyotype can be detected (e.g., Lang et al., 2012)). Nevertheless, our demography-informed discriminant analysis is still powerful for genomic regions with different evolution histories. In addition, this approach can also be applied to non-model species with no genome reference and sparse genomic markers.

2.4.3 Adaptation from mosaic genotypes

With the increasing availability of physical maps among different species, the important role of chromosomal inversions in maintaining adaptive divergence has been shown to be prevalent in many systems, such as controlling flowering differences in *Mimulus guttatus* between different ecotypes (Lowry and Willis, 2010) and wing patters that form Batesian mimicry in *Heliconius numata* (Joron et al., 2011). The unique aspect of adaptation via polymorphic inversions in Sub-Saharan mosquito species *Anopheles gambiae*, is the prevalent introgression and sharing of inversions among sibling species (Besansky et al., 2003; Coluzzi et al., 2002), which posed challenges in recovering phylogenetic rela-

tionships within *Anopheles gambiae* complex (Besansky et al., 2003, 1994; Bhutkar et al., 2007; White et al., 2011).

Our study showed how different sets of adaptive loci for different habitat (e.g., dry vs. wet) can be maintained in alternative rearrangements throughout a widespread species. The fact that the species complex do not have complete reproductive isolation and that they "borrow" pre-adapted inversions from each other while exploring new environments provide an interesting example of how adaptation leads to mosaic genotypes instead of new species, especially for species with big populations and high connectivity. This mode of adaptation coincides with recent theories predicting that when gene flow persists between populations that are under divergent selection, mechanisms that can reduce or suppress recombination, or increasing linkage between co-adaptive and maladapted genotypes will be advantageous because the gene complex can avoid being swapped (Aeschbacher and Brger, 2014; Barton, 1995; Kirkpatrick and Barton, 2006; Yeaman and Whitlock, 2011).

2.5 Material and Methods

2.5.1 Sample collection and DNA extraction

Mosquitoes were collected indoor at each site using either aspirators or insecticide spray, and then were individually preserved in 0.5ml tubes containing 100% ethanol. Ethanol preserved samples were shipped back to our lab in the Ruthven Zoology Museum, University of Michigan. Morphologies of each sample were examined according to Gillies and Meillon (1968) and Gillies and Coetzee (1987) under dissecting microscope before the body was grinded for DNA extraction. QIAamp DNA Mini Kit was used to extract DNA, whose yield ranges from 10-200 ng per sample.

2.5.2 Molecular identification of species and karyotypes

The species status within the complex of each sample (*gambiae*/*coluzzi*/*arabiensis*) was determined by a PCR-RFLP method following (Fanello et al., 2002). Briefly, the species was identified by the difference in the number of bands and/or fragment length after HhaI digestion of part of the integenic spacer (IGS) of the ribosomal DNA PCR products. The presence of inversions were determined by PCR of unique breakpoint regions of alternative arrangements. For 2La, primers were chosen to amplify a 492 bp region of 2La distal breakpoint and a 207 bp product from 2L+^{a} proximal breakpoint (White et al., 2007b). For 2Rb, three primers amplify a 429 bp fragment on 2Rb breakpoint and a 630bp fragment on 2R+b breakpoint (Lobo et al., 2010). If only one of the two PCR bands is present on a gel electrophoresis, then the sample is considered to be homokaryotype of one arrangement; alternatively, if both bands are present with similar brightness, the sample is considered to be heterokaryotype.

2.5.3 ddRAD library preparation and sequence analysis

Genomic DNA from each sample was individually barcoded and processed into a reduced complexity library for Illumina sequencing using a double digestion Restriction Associated DNA sequencing procedure (ddRADseq; for details see Peterson et al., 2012). Briefly, DNA was digested with the two most frequent restriction enzymes, MluCI and MseI, to maximize the number of unique short fragments. The digested products were then ligated by part of Illumina adaptor sequences and unique barcodes. Ligation products were pooled among samples and size-selected between 340 and 420 base pairs (excluding adaptor lengths) using a Pippin Prep (Sage Science) machine. The targeted-size ligation products were amplified by iProof^{TM} High-Fidelity DNA Polymerase (BIO-RAD) with 12 cycles to add complete Illumina adaptors. The library was sequenced in two lanes on the Illumina HiSeq2000 platform to generate paired-end 100 base pair reads. Sequences were identified to each sample based on the barcodes. Only reads with an average quality score

of at least 30 (Phred) and an unambiguous barcode and restriction cut site were retained.

After filtering, sequences were mapped to the *Anopheles gambiae* reference genome AGam30 (Holt et al., 2002) using bwa-mem algorithm in bwa with default settings (Li and Durbin, 2009) and mappings were retained with quality scores above 10 using samtools (Li, software no pub). SNPs were called from mapped contigs and genotypes were assigned using a maximum-likelihood statistical model (Catchen et al., 2011; Hohenlohe et al., 2012) with the Stacks v1.03 pipeline (Catchen et al., 2013); default settings were used except where noted below. Specifically, loci (termed as "stacks" in the program) were identified from genomic locations with mappings of at least 5 copies of RAD sequences in each individual using the PSTACKS program, to ensure credible calling of heterozygous SNPs in an individual (Catchen et al., 2013). A catalog of loci were built with the CSTACKS program from the PSTACKS output files across individuals to check the presence or absence of a particular locus at a genomic location. We retained loci which are present in at least 50% of all the samples and no more than two haplotypes per locus within each sample.

2.5.4 Population genetic structure of collinear and inverted regions

Geographic structure of *An. gambiae* populations were examined by measuring population divergence and performing principal component analysis principle component analyses (PCA). We performed separate analyses on collinear regions and inverted regions of each chromosome. Weir and Cockerham's F_{ST} 1984 and nucleotide diversity (π) were estimated on a per-site basis and windowed basis (150kb per window, 50kb step) along the genome using the POPULATIONS program in the Stacks pipeline (Catchen et al., 2013). SNPs were thinned to be at least 1000bp apart on the genome and imported into adegenet 1.4-1 package (Jombart, 2008) in R (2011) for PCA analyses. Only SNPs that are present in all populations and at least 80% of all individuals were included in the study. Missing genotype were filled with the average value. Based on PCA results, discriminant analysis of principal components (DAPC; Jombart et al., 2010) were run to determine genetic clus-

ters within the species without prior assumptions on the model of population subdivision. DAPC runs K-means clustering on the transformed PCs to identify groups of individuals that maximizes between-group genetic variation while minimizes within-group variation. Best supported number of clusters is then determined by the comparison of model likelihoods through Bayesian Information Criterion (BIC) similar to the program STRUCTURE (Falush et al., 2003; Pritchard et al., 2000). The agreement between genetic clusters inferred from inverted regions and molecular karyotyping assignment was also checked to assess the reliability of PCR identification methods.

2.5.5 Demographic history of collinear and inverted regions

Inference of demographic history implemented in fastsimcoal2 (Excoffier et al., 2013) calculates composite-likelihood of joint-SFS across populations under user-specified demographic scenario by parameterized simulations sampled from priors and optimizes the parameter estimation through a conditional maximization algorithm (ECM). The derived and ancestral states of the SNPs were inferred from the comparison of four species in the *An. gambiae* complex: *An. gambiae* s.s., *An. arabiensis*, *A. quadriannulatus*, and *A. merus*. Whole genome scaffolds of the latter three species were mapped to gambiae genome using MUMmer 3.23 (Delcher et al., 1999, 2002) and unique alignments were kept. Majority state of the diallelic SNP among four species was considered ancestral and multi-states SNPs were filtered. In order to maximize the number of SNPs included and ensure reasonable running time for each scenario, we excluded individuals with less complete sequencing coverage and down sampled SNPs in each case to infer region specific demographic histories. Confidence intervals of parameter estimation were obtained by 100 parametric bootstrapping runs from the point estimations.

2.5.6 Coalescent history between *An. gambiae* and *An. arabiensis* in collinear regions

We first estimated the divergence time, introgression rate and recent population expansion of *An. gambiae* and *An. arabiensis* using joint-SFS built from Chromosome 3 as proxies for collinear regions (X chromosome has a different effective population size and different selection regime). SNPs were down sampled to 40 copies in *An. gambiae* and 6 in *An. arabiensis*. We fixed the population size effective population size (Ne) of *An. gambiae* through nucleotide diversity estimations from Radtag and used one variable SNP per Radtag (if there is) to estimate the other parameters relative to Ne because selecting random sites from Radtags to include invariable sites would result in too less variable SNPs to inform the demographic model correctly (see also Excoffier et al., 2013). Unsure of the magnitude and timing of introgression between the two species, we performed test runs to choose the most likely scenario of introgression: 1) there was a onetime introgression from *An. arabiensis* into *An. gambiae* after their ancestors' divergence; 2) gene flow has continued after their ancestors diverged into two species (Fig. 2.1b). The first scenario was proposed based on the hypotheses that both 2La and 2Rb were introgressed from *An. arabiensis* into *An. gambiae* (Besansky et al., 2003). However, results showed that the first scenario is significantly less likely than the second scenario because of unlikely parameter estimations and lower likelihood. Therefore, in the following analyses we assume continued gene flow between the two species.

2.5.7 Coalescent history between *An. gambiae* and *An. arabiensis* in regions with inversion polymorphisms

We fixed the parameters that have been estimated from collinear region models and focused on inversion specific parameters because collinear regions have a larger SNP dataset. After the introgression of inversions from *An. arabiensis* to *An. gambiae*, individuals of

An. gambiae with the inversion (abbreviated as I in the following description) can freely recombine with *An. arabiensis*, which implies that the introgression rate estimated from collinear regions also applies. However, recombination is severely reduced between individuals of *An. gambiae* with standard chromosomes (abbreviated as S in the following description) and I (Fig. 2.1c). We estimated the divergence time between standard chromosomes and inverted chromosomes, time of introgression for inversions, and recombination rates between alternative rearrangements using joint-SFS built from 2La or 2Rb regions while fixing other parameters that were estimated from the collinear region. SNPs were down sampled to 20 copies in *I* and *S*, and 6 in *An. arabiensis*.

2.5.8 Selection signature in inversion regions

Based on reconstructed demographic histories for inversion regions, we obtained neutral expectations of population genomic measures through simulations. These measures can then be compared against empirical data to detect selection. We first applied an outlier analyses similar to traditional approaches. One difficulty lies in the heterogeneity of recombination rates inside inversions between heterokaryotypes (i.e., recombination rates are higher in the center and decrease sharply towards the breaking point region, Navarro et al., 1997), whereas the recombination rates estimated from SFS demographic modeling was an average. In order to make neutral simulations more realistic, we adjusted the recombination rate to be higher in the center region and lower on the two sides. First, empirical 2Mb F_{ST} windows with 150kb-step were calculated and fed into a smooth spline function in R to get fitted values for each 150kb segment. Recombination rate was adjusted for each segment to the value that generates the fitted F_{ST} value. 1000 demographic simulations were carried out using estimated parameters for each segment to generate 100bp DNA sequences. Two population divergence measures between I and S were estimated for the sequences, F_{ST} and $\mathrm{D_{XY}}$. Lastly, empirical measures of each loci were compared against the range of values from simulations to identify outliers.

Our second approach is to utilize sets of summary statistics to detect selection. Three scenarios were run for 1000 replications under the current demographic model using msms (Ewing and Hermisson, 2010) : a) pure neutral evolution; b) selection occurred on the branch of *An. arabiensis* and inverted chromosomes (I); c) selection occurred on the branch of *An. gambiae* and continued in standard chromosomes. Selection started from the time when the two species diverged with selection coefficient ranging from 0.01 to 0.0001 (Fig. 2.2c). Selected locus is located in the center of a 50kb-long simulated region. We tested two ways of building discriminant functions: 1) based on summary statistics of one neutral locus that is close to the selected locus; 2) based on an average of summary statistics across several neutral loci of a region that contains selected locus. For the first approach, we sampled 100bp long sequences that are located 5Kb away from the selected locus (an average distance between empirical adjacent Radtags in our data). 12 summary statistics, including heterozygozity, theta_pi of each population, F_{ST} and $\mathrm{D_{XY}}$, were calculated for each simulated sequence. For the second approach, we sampled random sets of 100bp short sequences across the entire 50kb according empirical Radtag distributions and calculated an average of these summary statistics. In both cases, discriminant functions using principle components transformed from summary statistics (DAPC; Jombart et al., 2010) were built based on the simulated training sets to differentiate three scenarios. We then used the discriminant function to predict which scenario each loci belonged to based on their empirical estimations of summary statistics.

CHAPTER 3

Rapid adaptation with gene flow via a reservoir of chromosomal inversion variation?

3.1 Abstract

The increased recognition of frequent divergence with gene flow has renewed interest in chromosomal inversions as a source for promoting adaptive divergence. Inversions can suppress recombination between heterokaryotypes so that local adapted inversions will be protected from introgression with the migrants. However, we do not have a clear understanding of the conditions for which adaptive divergence is more or less likely to be promoted by inversions when the availability of inversion variation is considered. Standing genetic variation, as opposed to new mutations, could offer a quick way to respond to sudden environmental changes, making it a likely avenue for rapid adaptation. For a scenario of secondary contact between locally-adapted populations, we might intuit that standing inversion variation would predominate over new inversion mutations in maintaining local divergence. Our results show that this is not always the case. Maladaptive gene flow, as both a demographic parameter and the cause for selection that favors locally-adapted inversions, differentiates the dynamics of standing inversion variation from that of segregating point mutations. Counterintuitively, in general, standing inversion variation will be less important to the adaptation than new inversions under the demographic and genetic conditions that are more conducive to adaptive divergence via inversions.

3.2 Introduction

Allopatric populations usually accumulate locally adaptive alleles after a period of environmental change (Nosil et al., 2009; Papadopulos et al., 2011). How fast can the adaptive divergence occur?The two most important factors, sources of variation and the probability of fixation of favorable mutations, have been thoroughly discussed from classical population genetic theories (e.g., Ewens, 2004; Fisher, 1930; Kimura, 1983; Orr, 1998) to empirical data on patterns of genetic variation (e.g., Bradshaw et al., 1995; Colosimo et al., 2005; Karasov et al., 2010). Less is clear about the mechanisms that maintain these adaptive loci in the face of maladaptive gene flow after secondary contact, which is common at the early stage of adaptive divergence (Nosil et al., 2009). Specifically, gene flow from ecologically dissimilar populations will dilute locally-adapted loci and disrupt the combinations of alleles by recombination. Under such scenario, any mechanism that can lower the effective gene flow rate or protect the good combination of alleles from shuffling with bad alleles in recombination will be preferred in adaptation (Yeaman and Whitlock, 2011). Rearrangements in chromosomes-inversions-can serve as such a mechanism because they can suppress recombination between heterokaryotypes so that local chromosomes carrying the adaptive alleles within an inversion will be protected from introgression of the maladapted genes carried by the migrants (Kirkpatrick, 2011; Kirkpatrick and Barton, 2006; Manoukis et al., 2008; Navarro and Barton, 2003; Noor et al., 2001; Rieseberg, 2001).

Empirical evidence has shown that many adaptive loci are associated with inversions, especially complex traits such as wing patterns (Joron et al., 2011), diapause timing (Feder et al., 2003) and annual/perennial life-history shift (Lowry and Willis, 2010). However, it is not clear whether these inversions become established in the population because of the maladaptive gene flow. Although theoretically possible, and despite the appeal of such a hypothesis, we do not have a clear understanding of the conditions (genetic or demographic) for which adaptive divergence is more or less likely to be promoted by inversions. Similar to adaptive point mutations, the rate of adaptation via locally-adapted inversions is

determined by the availability of inversion variation (i.e., either new inversion mutations or pre-exisiting standing variation of inversions) and the probability of establishment of favorable inversions. Both aspects were modeled or simulated separately in several studies (Feder et al., 2011; Kirkpatrick and Barton, 2006; Manoukis et al., 2008). What is missing for evaluating the contribution of inversions to adaptation is critical information on the rate of adaptation that considers the probability of inversions capturing a locally-adapted genotype, as well as the likely contribution of standing inversion variation versus new inversion mutations.

Here we develop a theory for rapid local adaptation under gene flow via chromosomal inversions such that we are able to predict when (i.e., genetic or demographic conditions) the rate of adaptation by inversions will be higher. Moreover, we can evaluate the likelihood of standing inversion variation contributing to adaptive divergence. By expanding the repertoire of models of local adaptation, our work contributes to a growing body of work for predicting when different mechanisms are likely to promote rapid evolution (e.g., Hermisson and Pennings, 2005; Kirkpatrick and Barton, 2006; Przeworski et al., 2005; Scoville and Pfrender, 2010). Moreover, by focusing on the potential contribution of new mutational input versus standing genetic variation, the general rules derived from the developed theory takes on special significance given the difficulty for such distinctions based on empirical evaluations of molecular data (reviewed in Barrett and Schluter, 2008).

Similar to standing variation of point mutations that facilitate rapid adaptation under sudden environmental change (Orr and Betancourt, 2001; Przeworski et al., 2005) and bottlenecks (Hermisson and Pennings, 2005; Orr and Unckless, 2008), standing inversion variation can be readily established in the population without a prolonged waiting time for the occurrence of an inversion and will suffer less random loss compared to new inversions if the mean frequency is larger than 1/2N. However, unlike point mutations, inversions do not confer a fitness difference directly. Instead, they reduce recombination cost and create linkage disequilibrium among selected loci. Therefore, the chance of local adaptation from a

new inversion through indirect selection might be very low considering that the probability of an inversion mutation capturing coadapted genotypes would be smaller after the onset of gene flow, let alone the possible stochastic loss of the single mutation. This contrasts with standing inversion variation (i.e., inversions that captures good combinations of alleles before the onset of gene flow). Moreover, standing inversion variation will have less chance of harboring deleterious mutations because they have been under purifying selection before the onset of gene flow. These factors might enhance the importance of standing inversion variation in the scenario of secondary contact.

Our key finding is that when inversions facilitate divergence with gene flow, higher gene flow increases the contribution from standing inversion variation. Yet, under the demographic and genetic conditions that are more conducive to adaptive divergence via inversions, new inversions become a more important source. We discuss how this counter-intuitive result (and one that differs from a recent study; Feder et al., 2011) can only be understood by explicitly considering the dynamics of adaptation from inversion variation, highlighting the importance and utility of analytical models for studying adaptation. By considering a broad range of selective values of alleles, instead of assuming weak selection (Kirkpatrick and Barton, 2006), our model and simulation also include predictions about (i) the characteristics of inversions contributing to adaptation (e.g., the selective benefit of alleles and its relationship with migration and number of loci involved) and (ii) conditions when the relative importance of standing inversion variation as a source of maintaining adaptive divergence might be increased.

3.3 Models and Methods

Consider a situation in which a peripheral population is receiving maladaptive gene flow from the central population across a heterogeneous environment such that alleles at two (or more) loci confer a selective benefit in the peripheral population but are maladap-

tive in the central population (Fig. 3.1). If the genetics of local adaptation is based on more than one locus, the maintenance of adapted alleles in the gene flow will depend not only on the selective benefit of an allele, but also on whether recombination will break up coadapted alleles (i.e., produce genotypes with a combination of adaptive and maladaptive alleles). If a chromosomal inversion captures the locally adapted alleles, with the introduction of maladapted alleles by gene flow (Fig. 3.1), there is a selective advantage to adapted alleles captured in an inversion because of suppressed recombination, whereas in the standard (i.e., non-inverted) chromosome locally adapted alleles can freely recombine with maladapted migrant alleles in heterokaryotypes, thereby breaking up locally adapted genotypes (Kirkpatrick and Barton, 2006).

Based on the scenario described above, we focus on comparing the dynamics and probabilities of maintenance of divergence from new inversions and standing inversion variation. Consider the simplest scenario of a single inversion mutation that captures locally coadapted alleles of two loci in a diploid population, where alleles *A* and *B* each have a homozygous fitness advantage (s) with dominance coefficient(h) over the maladapted alleles a and b from a different population (Fig. 3.1A). The two loci are linked on the same chromosome with recombination fraction *r* (Fig. 3.1A). These loci have independent influence on the individual fitness (i.e., multiplicative fitness is assumed, meaning no epistasis, Table 3.1). Before onset of migration, the populations are monomorphic with haplotype *AB*. With migration, maladapted alleles (*ab*, shown in black) replace m proportion of the locally adapted *AB* individuals each generation, and form recombinants (*Ab*, *aB*) with homokaryotes (i.e., with the standard, non-inverted chromosome). We define inversions mutations that capture the coadapted genotype and occur after gene flow started as new inversions (NI, Fig. 3.1B), and the ones that segregate in the population before gene flow as standing inversion variation (SIV, Fig. 3.1C). The standing inversion variations are selectively neutral until the start of gene flow because individual fitness is determined only by the alleles. Denoted as *AB**, recombination between inversions and other standard karyotypes is sup-

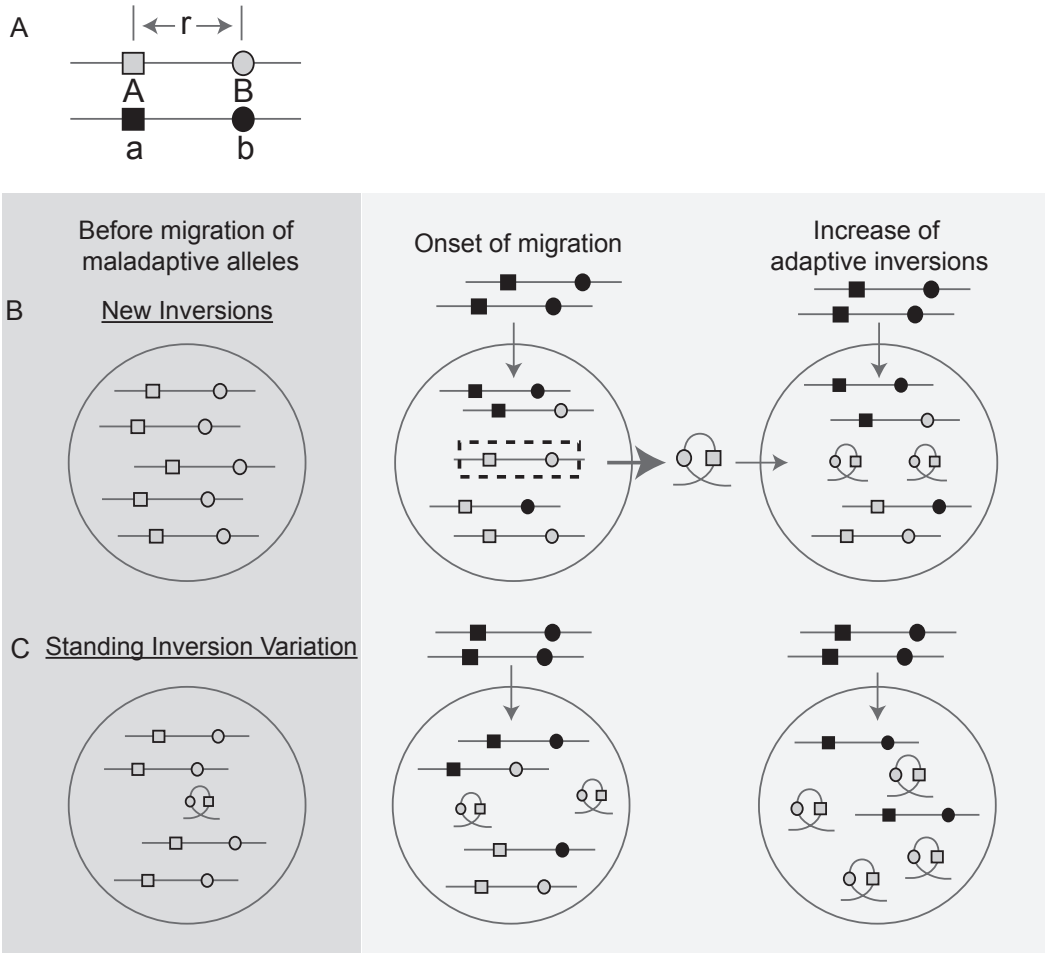

Figure 3.1: Illustration of the processes involved in adaptation from inversions under divergence with gene flow. (A) For two loci (shown by a square and a circle), alleles *A* and *B* (shown in grey) are locally adapted compared with the maladapted alleles, *a* and *b* (shown in black); the two loci are linked on the same chromosome with a recombination fraction, *r*. (B) Adaptation from new inversion and (C) standing inversion variation when divergence occurs with gene flow. See method section for the explanation of the process.

pressed. Because migrants can only form recombinants with standard chromosomes, the inversion *AB** will become more advantageous (if it survives the initial stochastic loss) as

Table 3.1: Fitness of offspring from different parental genotypes, where the haplotype *AB* is locally adapted (see Fig. 3.1) assuming loci have independent fitness effects (i.e., multiplicative fitness, or no epistasis).

	BB	*Bb*	*bb*
AA	$w_{1,1} = (1+s)^2$	$w_{1,2} = (1+s)(1+hs)$	$w_{2,2} = (1+hs)^2$
Aa	$w_{1,3} = (1+s)(1+hs)$	$w_{1,4} = w_{2,3} = (1+s)$	$w_{2,3} = (1+hs)$
aa	$w_{3,3} = (1+hs)^2$	$w_{3,4} = (1+hs)$	$w_{4,4} = 1$

the proportion of recombinants, *Ab* and *aB*, is built up. When adaptation from inversions is successful, all the other genotypes, *AB*, *Ab*, *aB*, will be replaced by *AB** with *ab* left in the population if gene flow continues (in this study this stage is referred to as establishment of the inversion).

The rate of spreading of an inversion depends upon the following parameters in a deterministic model: gene flow rates, allele effect sizes, number of locally-adapted loci and rate of recombination between them. When the allele effect size is small (i.e., $s \ll 1$), it can be omitted from the analytical approximation (see Eq 1. in Kirkpatrick and Barton, 2006). However, if allele effect size is not negligible, linkage disequilibrium (LD) between adapted loci needs to be considered to calculate the frequency of genotypes. Here we relaxed the assumption, and derived analytical approximations to examine the relationship between gene flow (m) and allele effect size (s) in determining the rate of adaptation via inversions. Comparisons between the probabilities of adaptation from new inversions with that from standing inversion variation are also evaluated in the context of the availability of these two sources under different parameter spaces.

All analytical equations were tested using time-forward simulations (Mathematica and Matlab code available upon request). At each generation, with a population of N diploid hermaphrodically reproducing individuals (Ne = 5000), migration (m) occurs first, followed by selection of individuals according to their fitnesses (Table 3.1); recombination occurs at rate *r* as gametes are formed meiotically, and then the next generation of dipoloids is randomly drawn from the pool of gametes. Each individual in a population is represented as two linear chromosomes of n loci with same allele effect (s) and no dominance (h =0.5).

Recombination is suppressed in heterokaryotypes. Free recombination is assumed between loci (i.e., r = 0.5), except for simulations that explore the effects of specific parameter values. In new inversion case, migration is allowed to occur until the population reaches migration-selection-drift balance. New chromosomal inversions are introduced at a mutation rate, μ, of 10^{-7} per gamete per generation in a population at migration-selection-drift balance. The inversion is tracked and generation time is recorded until when the inversion either goes extinct or replaces all the other adapted genotypes (i.e., selected loci in non-inverted chromosomes). In the standing inversion variation case, the starting frequency of the standing inversion variation for each simulation is selected at random from the expected distribution of neutral segregating inversions under the same demographic settings (generated by forward simulation with over 10,000 generations). The inversions are again tracked until either their loss or establishment. To determine whether new inversions or standing inversion variation is more likely to contribute to rapid adaptation, a population allowed for both new inversions and standing inversion variation is generated (as described above). The proportion of simulations in which each of the two sources of adaptive variation are either lost or established is quantified. 10,000 replicates were run for each set of parameter values.

3.4 Results

3.4.1 Selective advantage of a new inversion.

In the simplest scenario of a single inversion mutation that captures locally coadaptive alleles of two loci in a diploid population (Fig. 3.1), gametes *AB*, *Ab*, *aB* and *ab* have the genotype frequencies x_1,x_2,x_3,x_4. The change of gametic frequencies will be $\Delta x_i = (1-m)(x_i\frac{W_i}{\bar{W}} - \rho_i) - x_i$ (Li and Nei, 1974), where ρ_i is the change in frequency of x_i attributable to recombination with either locally coadapted or maladaptive alleles. It is defined as $\rho_i = (-1)^{|i-2.5|+0.5}\frac{w_{1,4}rD}{\bar{W}}$ (Lewontin and Kojima, 1960), where $w_{1,4}$ is

the fitness of double heterozygotes *A*/*a* *B*/*b* (Table 3.1) and *D* is the coefficient of LD, $x_1x_4 - x_2x_3$. At migration-selection balance, where $\Delta x_i = 0$, $\hat{\rho}_i$ can be expressed as $\hat{\rho}_i = \hat{x}_i \left(\frac{\hat{W}_i}{\hat{\bar{W}}} - \frac{1}{1-m} \right)$.

When an inversion captures coadapted alleles, denoted as *AB** (Fig. 3.1), its frequency will increase in the next generation as a function of $\lambda = (1-m)\frac{\bar{W}_I}{\hat{\bar{W}}}$ (Kirkpatrick and Barton, 2006), where $\bar{W}_I$ is the average fitness of individuals with an inverted chromosome and $\hat{\bar{W}}$ is the average fitness of the population at equilibrium, which is determined by the frequencies of the four gametes and the fitness of each genotype (Table 3.1). The initial increase in the frequency of a new inversion, λ, is therefore proportional to the decrease of the frequency of the coadapted genotype attributed to recombination scaled by gene flow,

$$\lambda = (1-m)\frac{\bar{W}_I}{\hat{\bar{W}}} = (1-m)\frac{\hat{W}_1}{\sum \hat{x}_i \hat{W}_i} = 1 + (1-m)\frac{\hat{\rho}_1}{\hat{x}_1} \tag{3.1}$$

where $\hat{W}_i = \sum_j \hat{x}_j w_{i,j}$. The inversion will always be favored by selection (i.e., $\hat{\rho}_1$ will be positive) as long as s is large enough such that locally adapted alleles *A* and *B* can withstand the swamping effect of migration of maladapted alleles *a* and *b* (i.e., $s > m/(1-m)$ for $s \ll 1$; there is no simple approximation if *s* is larger). Adaptation occurs by the increase in the frequency of the inversion, which will reduce the effective gene flow rate and elevate the mean fitness of the population.

3.4.2 Probability of establishment of a new adaptive inversion.

The probability of establishment of a new adaptive inversion (f_{NI}) is determined by its selective advantage in the first few generations. Using a branching process approach, classical work showed that it is approximately twice its initial selective advantage weighted by the reproductive variance (i.e., $2(\lambda - 1)/\lambda$, Haldane, 1927; Kimura, 1957). Hence, a

new adaptive inversion will be established in the population with the probability

$$f_{NI} = \frac{2(1-m)\hat{\rho}_1}{(1-m)\hat{\rho}_1 + \hat{x}_1} \tag{3.2}$$

assuming the number of offspring per parent is Poisson distributed (such that the reproductive variance of inversions equals λ) under a Wright-Fisher model. Using numerical approximations to explore the establishment probability of *AB** under different combinations of *m* and *s* (Fig. 3.2A), we show that it is not just the rate of migration (Kirkpatrick and Barton, 2006), but that the allele effect size is also important. The greatest probability of establishment of a new inversion, f_{NI}, occurs when allele effect size is the smallest. This can be intuitively understood by considering that when locally coadapted alleles are captured by an inversion, they never suffer the disadvantage of being found with maladapted immigrant alleles because of suppressed recombination. However, the selective advantage of inversions decreases as the allele effect sizes increase because $\hat{x}_1$, the frequency of the coadapted genotype *AB* on the standard (i.e., non-inverted) chromosome also increases when the allele effect size increases (Eq. 3.1). Overall, the migration rate, *m*, is the primary determinant of the probability of establishment of a new inversion (Fig. 3.2A), having a larger effect on the probability of establishment of the new inversion than the allele effect size. Nevertheless, there is a limit to which migration can facilitate the adaptation from inversions and this limit is determined by the effect size associated with the contained alleles. Specifically, we show that the equilibrium frequency of inversions,$\hat{y}$,decreases at high gene flow rates,

$$\hat{y} = 1 - \frac{m}{(1+s)^n - (1+hs)^n} - m + O[m]^2 \tag{3.3}$$

where *n* is the number of adaptive loci in inversion, which is opposite of the effect of m on the probability of establishment of a new inversion (see also Kirkpatrick and Barton, 2006).

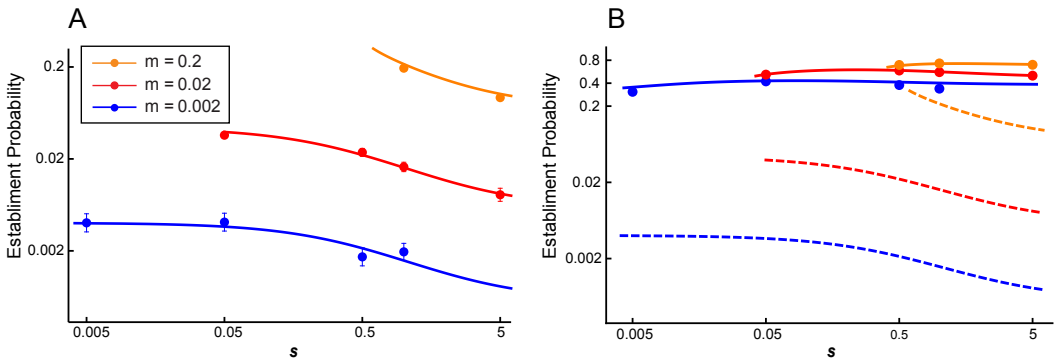

Figure 3.2: Comparison between establishment probability of new mutations versus that from standing inversion variation. (A) Establishment probability of a single new mutation of inversion in the marginal population at migration-selection balance for a population size (2N) of 10,000 under different orders of gene flow rate (m) and allele effect size (s). (B) Establishment probability of standing inversion variations (solid lines) compared with that from a single inversion mutation (dashed lines). Lines are theoretical predictions while solid circles are simulation results with 95% confidence levels shown as error bars.

3.4.3 Probability of establishment of adaptive standing inversion variation.

Now let us consider an adaptive inversion, *AB**, that is segregating in a population at frequency y (Fig. 3.1C). When gene flow starts, with regards to the rate of increase in the frequency of the inversion, Eq. 3.1 still holds, except that the frequencies of genotypes are not in equilibrium. Therefore, the rate of change of the inversion frequency becomes

$$\lambda_t = (1-m)\frac{W_I(t)}{\bar{W}(t)} = (1-m)\frac{W_1(t)}{\sum x_i(t)W_i(t) + y(t)W_I(t)} = 1 + \frac{\Delta x_1(t)}{x_1(t)} + (1-m)\frac{\rho_1(t)}{x_1(t)} \tag{3.4}$$

Somewhat surprisingly, following the influx of maladapted genotypes with the initiation of migration between the populations, our results show that the frequency of the inversion will actually decrease for a few generations ($\lambda_t < 1$). This is because gene flow will initially decrease the frequency of x_1 (i.e., Δx_1 is negative), and with low frequencies of recombinants

(*Ab* or *aB*) $\rho_1(t)$ is small (i.e., there is a small change in the frequency of x_1 attributable to recombination with either locally coadapted or maladaptive alleles at time *t*). However, as the frequency of recombinants increases, the selective advantage of an inversion is realized when divergence occurs with gene flow. Thus, the probability that any single copy of segregating inversion surviving till generation *t*, U_t, can be determined by integration of the changes in λ of each generation using a time heterogeneous branching process (Ohta and Kojima, 1968),$U_t = 1 - Exp\{\lambda_0(\cdots\lambda_{t-3}(Exp\{\lambda_{t-2}(Exp\{-\lambda_{t-1}\}-1)\}-1)\cdots\}$. The probability of establishment of a segregating inversion for a given frequency y is just the probability that at least one copy of the inversion survives stochastic loss, $\Pi_y = 1-(1-U_\infty)^{2Ny}$. Since the segregating inversion can be viewed as neutral mutations prior to the onset of migration (Fig. 3.1C), the probability of observing *k* copies of inversions in a population 2N at the time when gene flow starts can be approximated as $f(k) = C_0\frac{1}{k}$, where $C_0 = 1/\sum_{k=1}^{2N} 1/k$ (Ewens, 2004), assuming no back mutation. Thus, the probability of establishment from standing inversion variation becomes

$$P_{SIV}(N|k>0) = C_0\sum_{k=1}^{2N}\frac{1}{k}\Pi_{k/2N} \tag{3.5}$$

Considering a range of allele effect sizes, we show that the highest probability of establishment of segregating inversion variation occurs when selection is an order higher than the migration rate (Fig. 3.2B). Compared to the establishment probability of a new inversion, the probability of establishment of a segregating inversion (i.e., $k > 0$) depends much more weakly on the migration rate m than in the case of new mutations (i.e., the establishment probability is logarithmically, not linearly, related to *m*; Fig. 3.2B).

3.4.4 Comparison of the probability of adaptation from two sources of inversion variation.

To determine the conditions under which adaptation from new mutations versus standing inversion variation is more probable, we have to consider not only the probability of establishment of the inversion (as discussed in the previous section), but also the availability of inversions. For new inversions, the relevant factors determining the availability of inversions is mutational input, whereas for standing genetic variation, the key parameter is the frequency distribution of segregating inversion variation upon the start of gene flow.

The input of new inversion mutations can be approximated as $\theta = 2N_e\mu_I x_1$, where x_1 is the frequency of coadapted genotype *AB* in the population and μ_I is the mutation rate of inversions per gamete per generation that encompass the region of the chromosome where adaptive loci are located. Therefore, the probability of adaptation from a new inversion mutation within *T* generations is

$$P_{NI} = 1 - Exp\{\theta f_{NI} T\} \tag{3.6}$$

Similar to the establishment probability (f_{NI}) for a new inversion mutation conditioned on the availability of a new mutation, P_{NI} also increases along with *m* (Fig. 3.3A). The ratio of *s* relative to *m*, rather than exact values of *s*, is key to determining P_{NI} given same *m*. While f_{NI} is greatest at low values of *s* (Fig. 3.2A), when the waiting time for a new inversion mutation is taken into account, the probability of adaptation, P_{NI}, is actually improbable at lower range of s/m (Fig. 3.3A). This is because when alleles are under weak selection, the frequency of the adaptive genotype *AB* is so low that it is unlikely for a new inversion mutation to capture it. Instead, the highest probability of adaptation is maximized at a moderate ratio of s/m because of the tradeoff between the rate of establishment and inversion availability (Fig. 3.3A).

Below we derive the probability of adaptation from standing inversion variation by

integrating over the availability of the inversion and its establishment probability,

$$P_{SIV}(N,\mu) = \int_0^1 \rho_I(y)\Pi_y dy \tag{3.7}$$

where the frequency spectrum of segregating inversions in the population at mutation-drift balance before the onset of migration can be derived (see Ewens, 2004; Hermisson and Pennings, 2005) as $\rho_I(y) = \frac{C_0 y^{4N_e\mu_I - 1}(1-y^{1-4N_e\mu_I})}{1-y} \approx 4N_e\mu_I y^{4N_e\mu_I - 1}$, where C_0 is a constant of integration. The shape of $P_{SIV}(N,\mu)$and $P_{SIV}(N|k > 0)$ for different values of m and s are similar (Fig. 3.2B, 3.3A), but $P_{SIV}(N,\mu)$ scales with the availability of inversions, $N_e\mu_I$. This means the chance of observing standing inversion variation at a given time in the population is proportional to the mutation rate. In contrast with the establishment probability, which is always higher for standing inversion variation compared to new inversions, when the availability of the inversion is also considered, the probability of adaptation via standing inversion variation is not necessarily going to be higher than the probability of adaptation by new inversions.

3.4.5 Contribution of standing inversion variation to adaptive divergence.

For rapid adaptation via inversions under a divergence with gene flow model, how important is standing inversion variation relative to new inversion mutations as the likely source? This question can be evaluated by calculating the relative contribution of standing inversion variation to adaptive divergence, as derived by a combination of Eq. 3.6 and 3.7,

$$R_{SIV} = \frac{P_{SIV}}{P_{ADP}} = \frac{P_{SIV}}{P_{SIV} + (1 - P_{SIV})P_{NI}} \tag{3.8}$$

following Hermisson and Pennings (2005). As P_{NI} increases with time, if time allowed for inversions to occur is long enough, P_{NI} will eventually surpass P_{SIV} regardless of the

scenario. However, in this paper we are interested in rapid rescuing effect from inversions after secondary contact, we only simulate the situation when the time allowed for adaptation is short (i.e., 0.2 Ne generations) so that source of inversion variation is highly relevant to the probability of adaptation.

There are several parameter regions where the relative contribution from standing inversion variation is particularly important (Fig. 3.3B), specifically, when s is at the same order of m so that s is too small to withstand the gene flow (see $m = 0.2$, $s = 0.5$ on Fig. 3.3A), or when allele effect sizes are large enough compared to migration ($s \gg m$). These regions, however, all correspond to situations where adaptation via inversions are less to probable to occur. Plotting R_{SIV} against P_{NI} (Fig. 3.3C), we can see that as adaptation via new inversions becomes more probable (higher P_{NI}), contribution from standing inversion variation quickly drops.

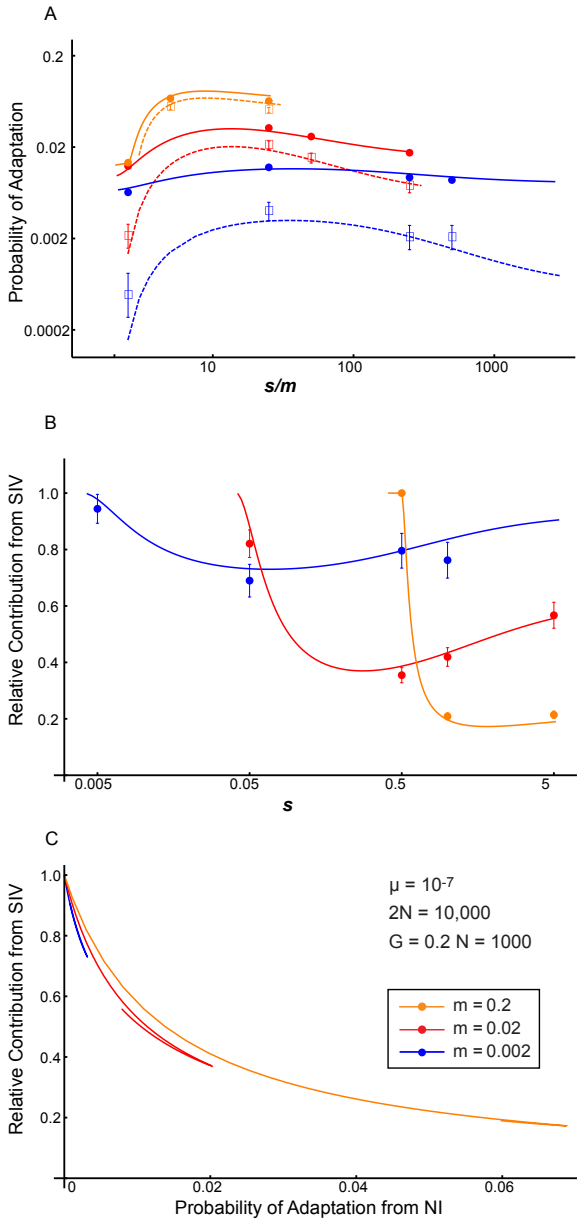

Figure 3.3: Probability of adaptation and relative contribution from standing variation. (A) Comparison between adaptation from new inversions (P_{NI}; dashed lines) and adaptation from both sources (P_{ADP}; solid lines) given new input of mutations persisting for $G = 0.2N_e$ generations after the initiation of maladaptive alleles (see Fig.1) for a population size (2N) of 10,000 under different orders of *m* and *s*. Mutation rate, $\mu = 10^{-7}$. 95% confidence levels of each simulation are showed on error bars of the squares (P_{ADP}) or circles (P_{NI}). Note that the probability of adaptation is plotted against s/m. (B, C) Relative contribution of standing inversion variation to rapid divergence (i.e., within 0.2 N generations). (B) is plotted against *s*, (C) is plotted against P_{ADP}.

3.5 Discussion

Both standing genetic variation and chromosomal inversions have become central foci as mechanisms to facilitate rapid adaptation (Barrett and Schluter, 2008). By developing an analytical model that makes explicit the factors governing the dynamics of rapid adaptation based on inversion variation, we show that when adaptive divergence via inversions with gene flow is more likely, new mutational input (i.e., new inversion variation) becomes a more probable genetic source than standing inversion variation. By considering a broad range of selective values of alleles (instead of assuming weak selection, Kirkpatrick and Barton, 2006), we also use our model and simulation to predict (i) the characteristics of inversions contributing to adaptation (e.g., the selective benefit of alleles and its relationship with migration and number of loci involved) and (ii) conditions when the relative importance of segregating inversion variation as a source of rapid adaptation might be increased.

3.5.1 Implications of results for the genetics of adaptation

Inversions are more likely to facilitate local adaption under higher gene flow rates (Fig 2; see also Kirkpatrick and Barton 2006). Consequently, we can identify how the genes contained within the inversion are likely to be involved in adaptation because of their impact on the effective gene flow rate. The ratio between allele effect size and gene flow, rather than the absolute value of the effect size, determines the likelihood of this scenario. The highest probability of adaptation is maximized at a moderate ratio because of the tradeoff between the rate of establishment and inversion availability (Fig. 3.3A). This finding predicts the genomic profile of adaptation, that is, whether divergence is achieved through multifarious selection on many genes or through linked regions within genetic islands (Nosil and Feder, 2012; Nosil et al., 2009) with inversions involved. The allele effect sizes of selected loci that can benefit the most from being captured in inversions will differ given different levels of gene flow. Under adaptation with strong gene flow (i.e., when pop-

ulation migration rate, $2Nm$, is much larger than 1; Wright (1931)), multifarious selection on many small effect genes cannot resist gene flow effectively, leading to clustering of few genes with larger effects in freely recombining region (Yeaman and Whitlock, 2011) (see Fig. 3.2 for increasing minimum s to withstand gene flow at higher m). In this case, it is beneficial for selected loci with even big allele effect sizes to be captured in inversions. On the other hand, under adaptation with weak gene flow, multifarious selection can be seen more often in freely recombining region. In this case, much smaller effect alleles would be found more often within inversions.

Although it is widely recognized that either increasing the recombination rate between loci, r, or number of adaptive loci, n, will affect the probability of adaptation (Table B.1 and Eq. 2, 3 in Kirkpatrick and Barton 2006), their interaction generates different expectations for the genetics of adaptation, because *r* and *n* are usually negatively correlated for a given length of inversion. In other words, capturing more adaptive loci within the same length of an inversion will continuously increase its selective advantage until the point when all of the fitness-related loci are tightly linked (Fig. B.1). Therefore, the length of an inversion influences its fitness because longer inversions can capture more adaptive loci without tight linkage while shorter inversions are less likely to be advantageous. This result helps to explain the observed size distribution of inversions in natural populations. For example, in *Anopheles gambiae*, while rare chromosomal inversions were found to vary randomly in length (Pombi et al., 2008), common inversions which are more widely spread in the populations tend to be long.

3.5.2 Contribution of standing inversion variation versus new inversions to adaptation

High initial frequency and the immediacy of standing genetic variation are frequently cited as reasons why it is a more probable source for rapid adaptation than new mutations (Barrett and Schluter, 2008). As with adaptation via new point mutation versus standing

genetic variation (see Innan and Kim, 2004; Orr and Unckless, 2008; Przeworski et al., 2005), standing inversion variation also has a significantly higher establishment probability (Fig. 3.2) by virtue of a higher segregating frequency in a population (i.e., they are not as sensitive to stochastic loss by genetic drift compared with new inversions). However, consideration of the establishment probability alone (e.g., Feder et al., 2011; Kirkpatrick and Barton, 2006) is not sufficient for understanding the contribution of standing inversion variation relative to new inversions. As modeled here, the availability of inversions is critical for evaluating whether adaptation is actually probable (Fig. 3.3A). For point mutations, their availability is only determined by the effective population size and mutation rate, whereas it is more complicated for inversions. Therefore, the impact of the immediacy of segregating inversions on the relative contribution of standing inversion variation and new inversions to adaptive divergence varies under different scenarios (i.e., combinations of m and s).

Our results show that higher gene flow rates result in greater contributions of adaptation from standing inversion variation if probability of adaptation (P_{NI}) is controlled for (Fig. 3.3C). This can be understood by considering that an inversion does not confer a fitness difference directly, in contrast with point mutations that facilitate rapid adaptation under sudden environmental change (Orr and Betancourt, 2001; Przeworski et al., 2005), bottlenecks (Hermisson and Pennings, 2005; Orr and Unckless, 2008) or domestication events(Innan and Kim, 2004). For inversions, the sudden influx of maladaptive alleles from migrants gives inversions an advantage over non-inverted chromosomes. Gene flow is not only the driving force of adaptation via inversions (i.e., the level of gene flow determines the selective advantage of inversions), but it also impacts the availability of inversions. Higher gene flow will lower the population size of favorable genotypes, making it less likely that inversion mutations will capture adaptive alleles. Under this scenario, using an available pool of standing inversion variation that already captured good genotypes becomes more important.

We also find the expected contribution of standing inversion variation to adaptive divergence decreases as the conditions become more favorable to adaptation via inversions (i.e., $P_N I$ increases; Fig. 3.3C). This is mainly because the probability for adaptation from standing inversion variation increases slower than that via new inversions when conditions become favorable (compare solid lines with dotted lines in Fig. 3.3A), such that standing inversion variation can only compensate for unfavorable situations (i.e., increase the probability of adaptation relative to new mutation when adaptive divergence is not likely), but not outcompete new inversions under favorable situations (see also Hermisson and Pennings, 2005). The reasons are two-fold. First, the advantage of a higher initial frequency of standing variation levels off when the probability of establishment from a single copy increases with higher gene flow and moderate allele effect size, the very conditions when adaptive divergence via inversions is actually likely. Second, the selective advantage of segregating inversions gradually builds up as the frequency of favorable genotypes drops and recombinants are accumulated (λ changing from negative to positive in Eq. 3.4). In contrast, a new inversion in a population at migration-selection balance realizes its maximum selective advantage, giving it a higher survival rate compared to a pre-existing inversion.

3.5.3 If adaptation occurs from standing inversion variation, what can we infer about the process of adaptation?

Although for the conditions explored here, standing inversion variation is less important overall when conditions are favorable for adaptation via inversions, this finding does not eliminate the potential importance of standing inversion variation. With an understanding of the relevant factors impacting the probability of adaptation from standing inversion variation, we can identify the evolutionary context where standing inversion variation is predicted to contribute to adaptation, as illustrated by the specific scenarios discussed below.

Time until establishment. Whether adaptation can occur rapidly may determine the

likelihood of evolution change (Hermisson and Pennings, 2005; Lynch, 2010). Establishment times are consistently shorter for standing inversion variation as compared to new inversions (compare circles to squares in Fig. 3.4). This discrepancy will be even larger if the waiting time for a new mutation to occur is included. Consequently, given an equal probability of adaptation for new inversions and standing inversions ($P_{NI} = P_{SIV}$), a faster establishment rate (shorter establishment time) of standing inversion variation alone will be highly likely to lead to rapid local adaptations. This was empirically supported by the case of *Drosophila subobscura*, which has an establishment time for standing inversion variation as short as 25 years to reach a similar latitudinal cline of adaptive inversion polymorphism seen in the old world after introduction into New World in the early 1980s (Balanya et al., 2003). These inversions are shown to harbor favorable combinations of alleles (Rego et al., 2010; Santos, 2009).

When and how standing inversion variation is introduced. While our findings hold when inversion variation evolves de novo within a focal population (where mutation rate sets the waiting time for new inversions as well as the chance of having segregating inversions), adaptive inversions introduced from populations located in similar environments could alleviate the recombination load that would accumulate in a population experiencing an influx of maladapted alleles from populations in dissimilar environments. Likewise, introgression from closely-related species is also a possible source of standing inversion variation. For example, the origin of the 2La and 2Rb inversions associated with dry environments in *Anopheles gambiae* (Coluzzi et al., 2002; White et al., 2007b) trace back to an introgression event with *Anopheles arabiensis* (Besansky et al., 2003, 1994). In another example of gene flow of adaptive inversions between populations, northerly distributed *Rhagoletis pomonella* gained inversion polymorphisms from Mexican populations that were strongly associated with the length of overwintering pupal diapauses, which facilitated a host shift (Feder et al., 2003).

Genetic background. In our theoretical model, we assume that segregating inversions

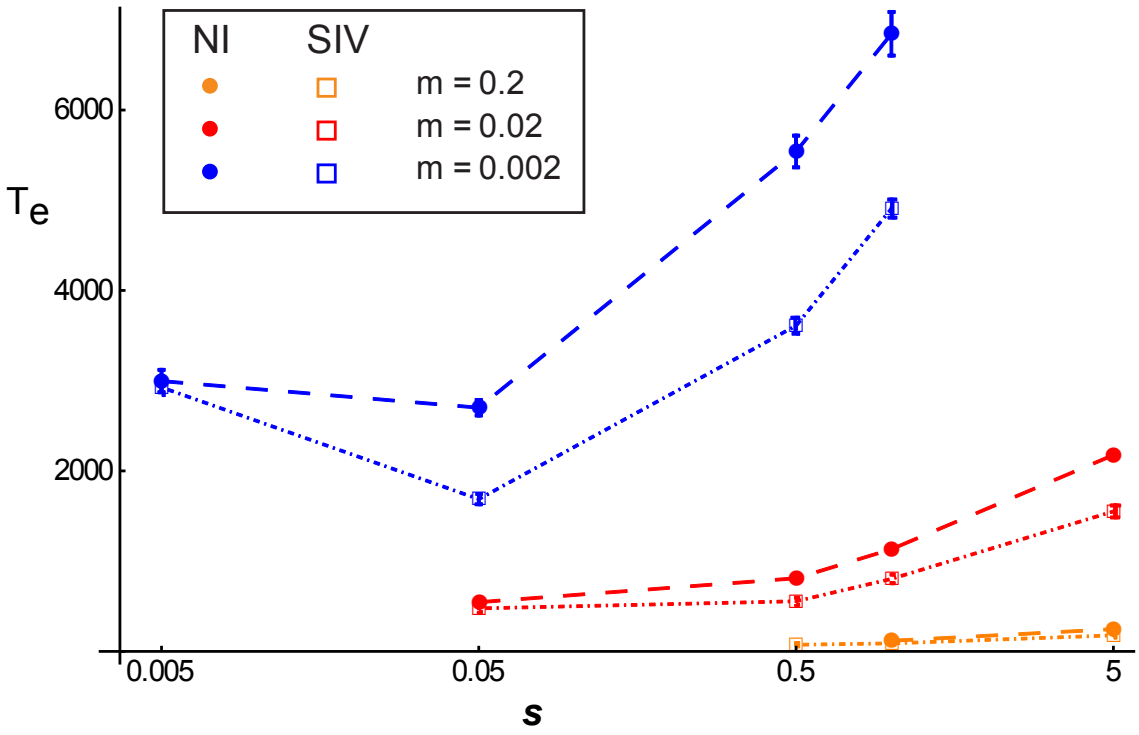

Figure 3.4: Average establishment time of inversions for new mutations versus standing variation calculated from runs with established inversions. Circles are waiting time for new inversions while squares are that for standing variations. Standard errors are shown as bars.

and new inversion mutations have the same fitness-that is, we do not consider the genetic background upon which the inversion occurs. Each new inversion mutation or segregating inversion can have a range of fitness values based on the genes they captured (Nei et al., 1967). Standing inversion variation may, in general, represent a more likely source of adaptive divergence because it will have already been exposed to purifying selection. The pre-filtering process would greatly decrease the frequency of inversions that capture genes with large fitness costs or have a direct fitness cost through meiotic problems. In other words, inversions with a lower fitness cost will segregate at a higher frequency, in contrast

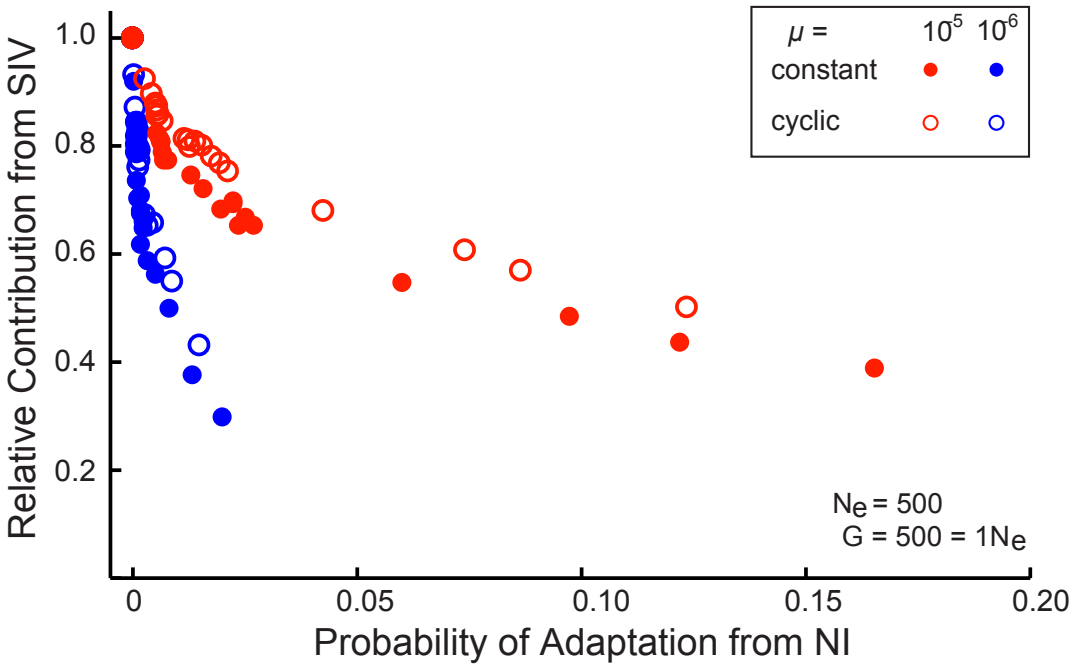

Figure 3.5: The relationship between probability of adaptation from new inversion and relative contribution from standing inversion variation (Ne = 500, G=1Ne) for different parameter settings (*m*, *s*, *r*, *n*) under two demographic scenarios: a constant population (N = Ne = 500) and cyclic population ($N(t) = 2525 + 2475\sin(2\pi(t + 6.5)/10)$). New input of mutations lasts for 500 generations (G = 1N), with two levels of mutation rate, 10^{-6} and 10^{-5}. 5,000 realizations in each scenario were run to observe the impact of different combination of parameters on the proportion of contribution from standing variation to the success of establishment of inversions. Blue colored and red colored circles denote different level of mutation rate, 10^{-6} and 10^{-5}, respectively. Open circles are simulation results from cyclic population while filled circles are from constant population realizations.

to new inversion variants, which have yet to pass through the selection gauntlet.

Demographic histories. Adaptive divergence in empirical populations may of course occur under conditions other than the constant population sizes modeled here. For example,

population sizes may fluctuate, especially in response to shifts in climatic or ecological factors. Such changes in population sizes are relevant because they will not only influence the amount of mutational input but also the relative contribution of new inversions versus standing inversion variation to adaptive divergence (Hermisson and Pennings, 2005; Kimura and Crow., 1970; Orr and Unckless, 2008; Otto and Whitlock, 1997). To compare the relative contributions of new inversions versus standing inversion variation with fluctuating population sizes, we considered a scenario involving cyclic dynamics (such as those observed in mosquito populations) where population sizes differ depending on climatic conditions (e.g., wet/dry or warm/cold). We did forward-time simulations using parameters selected from empirical studies of Anopheles gambiae populations (Manoukis et al., 2008), which are characterized by multiple inversion polymorphisms. We show that when the probability of adaptation becomes larger under different combinations of *m*, *s*, *r*, and *n*, the relative importance of standing inversion variation also decreases with fluctuating population sizes (Fig. 3.5), which is similar to what has been demonstrated under theoretical predictions (Fig. 3.3C). However, population fluctuations affect the steepness of the negative relationship between probability of adaptation and contribution from standing variation. When a population is cyclic, contribution from standing inversion variation is higher (Fig. 3.5). This is contingent on the assumption that the onset of gene flow occurs at the time when population size is increasing, in order to mimic the situation where populations begin to multiply and migrate when the wet season begins. Therefore, if inversions are pre-existing, they have much less chance to be lost by drift. The situation will be reversed when gene flow occurs in a shrinking population. However, the first situation is more probable under the context of secondary contact. In either case, the proportional contribution from standing inversion variation should be boosted or decreased by a factor of N/Ne according to Otto and Whitlock (1997).

CHAPTER 4

Integrative testing of how environments from the past to the present shape genetic structure across landscapes

4.1 Abstract

Tests of the genetic structure of empirical populations typically focus on the correlative relationships between population connectivity and geographic and/or environmental factors in landscape genetics. However, such tests may overlook or misidentify the impact of such factors on genetic structure, especially when connectivity patterns differ between past and present populations because of shifting environmental conditions over time. Here we account for the underlying demographic component of population connectivity associated with a temporarily dynamic landscape in tests of the factors structuring population genetic variation in an Australian lizard, *Lerista lineopunctulata*, from 24 nuclear loci. Correlative tests didn't support significant effect from factors associated with a static contemporary landscape. However, spatially explicit demographic modeling of genetic differentiation shows that changes in environmental conditions (as estimated from paleoclimatic data), and corresponding distributional shifts from the past to present landscape, significantly structure genetic variation. Results from model-based inference (i.e., from an integrative modeling approach that generates spatially explicit expectations that are tested with Approximate

Bayesian Computation) contrasts with those from correlative analyses, highlighting the importance of expanding the landscape genetic perspective to tests the links between pattern and process, revealing how factors shape patterns of genetic variation within species.

4.2 Introduction

Although temporal scale is one of the primary distinguishing factors of landscape genetic and phylogeographic study, such a distinction is not only unnecessary, but also potentially problematic. For example, landscape genetics studies how contemporary habitat suitability and connectivity influence population genetic structures spatially (Manel et al., 2003; Storfer et al., 2007). Phylogeography typically focuses on historical processes that generated the patterns of genetic variation (Avise et al., 1987; Knowles, 2009). There may certainly be cases in which one of the two processes predominates (e.g., Hull et al., 2008; Knowles and Carstens, 2007; Mendez et al., 2010; Perrier et al., 2011; Xu et al., 2009). Yet, because such studies are often pursued under one of the two perspectives, their joint influence can be overlooked, risking the misidentification of factors structuring patterns of genetic variation. As both landscape genetics and phylogeography shift towards the analysis of multilocus data, and specifically as next-generation sequencing technologies become widely applied (e.g., Gompert et al., 2010; Thomson et al., 2010), concerns over molecular markers as a distinguishing factor between landscape genetics and phylogeography (e.g., Wang, 2010) will certainly diminish. Likewise, the greater power and resolution provided by such datasets opens up new possibilities for expanding methodologies that can test causation of, as opposed to seeking associations with, the underlying patterns of genetic variation.

The melding of disciplines is represented in the approach advocated here, which we illustrate with an empirical example-specifically, a test aimed at revealing how geography and the environment shape patterns of genetic variation in a lizard, *Lerista lineopunctulata*.

This lizard is distributed along the southwestern Australian coastal sand plains or dunes (Fig. 4.1) (Cogger, 2000; Wilson and Swan, 2008). Sea level changes in glacial and interglacial periods expanded or contracted suitable costal sand habitats for the species (Hocking et al., 1987; Storr and Harold, 1978). Consequently, it is conceivable that population divergence could reflect the contemporary habitat configuration, which limits migration among the small geographically isolated populations (Excoffier et al., 2009a), or colonization associated with historical shifts in the species distribution (Zellmer and Knowles, 2009), given that a habitat specialist would track climate-induced habitat shifts. We first conduct both individual and population-level correlative tests to identify potential factors structuring genetic variation, including geography, climatic and soil characteristics (see also Edwards et al., 2012). We then move beyond these traditional descriptive landscape genetic analyses (Legendre and Fortin, 2010) with an approach that provides quantitative species-specific predictions that account for the interaction between abiotic and biotic factors (i.e., the environmental factors and the life history characteristics of taxa that mediate the impact of these factors on survival and movement patterns; see Knowles and Alvarado-Serrano 2009; Brown and Knowles 2012). Specifically, we generated a large multilocus dataset to test whether the current genetic structure reflects (i) the geographic configuration of populations, (ii) the contemporary environment, or (iii) the dynamic history of shifting environmental characteristics since the last glacial maximum.

Our work highlights the potential synergy between traditional landscape genetic approaches and model-based inferences by translating hypotheses identified from correlative analyses into a suite of alternative demographic processes that can be formulated as models (see also Brown and Knowles, 2012; Bruggeman et al., 2010; Epperson et al., 2010; Landguth et al., 2010; Morgan et al., 2011; Shirk et al., 2012). Our approach contrasts with the tradition of intuiting qualitative phylogeographic hypotheses from ecological niche models, ENMs (reviewed in Knowles, 2009). Here quantitative information about variation in the habitat suitabilities across space and time is used to inform a spatially explicit

demographic model whose parameters are then used for coalescent simulations. As a consequence, predicted patterns of genetic variation are species-specific, reflecting the interaction between the physical environment and biological parameters (e.g., local population sizes and migration rates) that determines the level and pattern of gene flow across the landscape (see Brown and Knowles, 2012; Knowles, 2009; Morgan et al., 2011). Additionally, we rigorously test these models using approximate Bayesian computation, ABC (Beaumont et al., 2002), and assess the quality of parameter estimates using pseudo-observed datasets, pods (see Bertorelle et al., 2010; Robert et al., 2011).

With reference to the empirical study of *L. lineopunctulata*, we highlight how extrapolating causation from descriptive correlates of genetic variation with the environment and geography would be misleading (see also Meirmans, 2012), but was avoided by applying model-based inference with an expanded repertoire of models (i.e., not only isolation-by-distance, IBD, but also models that include additional environmental factors, and temporal shifts in habitats across the landscape). This approach, iDDC modeling, integrates distributional, demographic, and coalescent models to generate predictions for species-specific patterns of genetic variation. With the intent that the methods proposed here can be generally applied to different biological systems that had experienced non-static demographic history, we include a discussion of not just the promise of iDDC modeling (see approaches described in Neuenschwander et al., 2008; Ray et al., 2005), but also the limitations.

4.3 Methods

4.3.1 Sampling and molecular data

Lerista lineopunctulata tissue samples (N = 89) were field collected or obtained from the Western Australian Museum and Australian Biological Tissue Collections (South Australian Museum) (Table C.1) for full geographic coverage of the species, with multiple individuals sampled from each of the delimited populations (see Edwards et al., 2012,

for details about population assignment). Note that the southern populations, formerly assigned to *L. lineopunctulata*, are considered a separate species under taxonomic revision (Edwards, Doughty and Keogh, unpublished data) and have not been included in this study (see also Edwards et al., 2012). Also note that the number of loci in this study was expanded considerably from 3 to 24 (a prerequisite for testing hypotheses, as opposed to describing genetic variation, as in Edwards et al. 2012).

Anonymous nuclear loci were developed from a Roche 454 sequencing run (procedures similar to Bertozzi et al., 2012), and supplemented with sequences for loci from published primers (see Table C.2). Marker development from a 454 run used one individual of the focal taxon, *L. lineopunctulata*, and one individual of *L. praepedita*. Note that *L. praepedita* was used to identify variable markers while avoiding ascertainment bias that results from using intraspecific screening sets (Carstens and Knowles, 2006, see). Details regarding preparation of DNA samples for developing markers are given in Gompert et al. (2010). This entailed the construction of a reduced representation library for each species from genomic DNA digested with EcoRI and Mse1 enzymes. Unique barcodes were ligated for each species and size-selected fragments in equimolar concentrations were used for the Roche 454 sequencing. The sequences were trimmed and quality filtered using custom perl scripts and assembled using the NGen sequence assembler v2.0 (DNASTAR); settings used in the assembly are provided in Supplemental Table 4.3. Contig consensus sequences were screened against BLAST to ensure loci were not mtDNA or transposable elements and did not belong to known gene families. Primers were designed for amplifying and sequencing fragments between 150-700bp using traditional Sanger-sequencing at the University of Michigan DNA core facility.

A total of 24 nuclear loci were sequenced in both directions in each individual (GenBank KC545970-KC549439), although there were some missing data due to PCR failures (Table C.4). Eighteen loci were identified from the 454 run that produced clear bands, with single copy sequences and contained at least one variable site between the two species (*L.*

lineopunctulata and *L. praepedita*). In addition, we sequenced six loci using published primers (see Table C.2 for references). PCR reactions were run in 20ul volumes with 2l 10x reaction buffer, 0.8-2.5l 50mM MgCl2, 1l 10mM dNTPs, 0.4l Bovine Serum Albumin, 0.8l of each 10M primer, 1U Taq polymerase, 100ng gDNA and the volume made up to 20l with ultra pure H20. PCR cycles were 95C 1min; 30 cycles of 95C 30sec, 59-65C 20sec, 72C 45sec; 72C 4min (see Table C.2 for specific conditions and exceptions). Haplotype phase was determined using PHASE (Scheet and Stephens, 2006).

4.3.2 Species Ecological Niche Modeling (ENMs)

In addition to our collected samples, occurrence data of *L. lineopunctulata* were collated from OZCAM (www.ozcam.org.au) and geo-referenced (see Edwards et al., 2012). Projections of current species distributions based on habitat suitability was estimated from 19 current climate layers from the WorldClim global climate database (www.worldclim.org). The ENMs were generated with MaxEnt v3.3.3k (Phillips et al., 2006) from 10 cross-validation runs, which accurately predicted the species distribution (area-under-the-curve, AUC, value of 0.971). The model was also used to predict the past distribution of the species at the Last Glacier Maximum (LGM) using the same 19 climate layers from the Community Climate System Model derived from PMIP2 database available on WorldClim database (Hijmans et al., 2005).

4.3.3 Tests of associations with genetic structure

The potential impact of environmental factors on genetic structures was tested at both the individual and population level. Distance-based redundancy analysis, or dbRDA (Legendre and Anderson, 1999), was used to test for the relationship between individual pairwise genetic distances and corresponding climatic and soil variables (i.e., the score at the sampling site where the individual was collected), conditioned on geographic distances (i.e., removing the effect of geographic distance separating individuals). dbRDA is a mul-

tivariate technique for testing a distance based matrix (in this study, the matrix of pairwise genetic distances) against rectangular predicting variables, in which the relationship between the principal coordinates of the distance matrix and the variables are then analyzed. For analyses of an association between environmental factors and population-level genetic structure, pairwise FST-values were calculated among populations using Arlequin 3.5 (Excoffier and Lischer, 2010) and the environmental differences among populations were summarized for an isolation by resistance test (detailed below).

Individual pairwise genetic distances were calculated in Arlequin 3.5 with Tajima and Nei's correction (Tajima and Nei, 1984). For this analysis, the multilocus data were condensed into two haplotypes per individual by concatenating one of the two alleles (selected randomly) for each locus across loci. Positions with more than 60% missing data (across individuals) were not included in the calculation of individual pairwise genetic distances. Calculations of environmental distances were conducted on the principal component 1 (PC1) from a principal component analysis (PCA) of the 19 climate layers extracted from ArcGIS10 due to correlation across climatic layers (Hirzel et al., 2002; Manel et al., 2001; Peterson et al., 2011). We performed PCA directly on the climate layers instead of values extracted from sampling points because we want to capture the variation in the environment but avoid any bias in the sampling points. This climate PC1 explained 90% of the variation in the climatic data. We also characterized the spatial variation in soil characteristics. Soil properties were derived from the soil type data in the Atlas of Australian Soils (Northcote et al., 1968) from the Australia Soil Information System (http://www.asris.csiro.au) and interpreted following McKenzie et al. (2000) as 13 measurements of the soil profiles including percentage of clay, thickness, water flow, nutrients (see Table C.5), which were summarized with a PCA. This soil PC1 explained 85% of the total variation in the soil data and was retained for analysis with dbRDA. Pairwise Euclidean geographic distances among all individuals were calculated in ArcGIS 10. Because dbRDA only relates a matrix to rectangular predictors, the geographic Euclidean distance matrix was transformed into

continuous rectangular vectors via principal coordinates analyses using the pcnm function of the Vegan package (Oksanen et al., 2012) in R (Team, 2012). Each of the three potential predictors (geographic distance, climate-PC1, and soil-PC1) were tested separately against genetic distance using the capscale function in the Vegan package, as well as tests of climate-PC1 and soil-PC1 conditioned on geographic distance (i.e., partitioning out the effect of geographic distances).

For the analyses of isolation by resistance (McRae, 2006) used in the population-level tests of association between environmental factors and genetic variation, the average resistance among populations was estimated in Circuitscape v3.5.8 (Shah and McRae, 2008) using habitat suitability score as per-cell conductance. Specifically, for each population, a convex hull (i.e., a polygon) that encompassed the minimum population area from sampling localities was used to define the region from which Circuitscape calculated resistance scores to represent the connectivity among populations. Isolation by resistance was tested using Mantel and partial Mantel tests in IBDWS v3.23 (Jensen et al., 2005) for population-level associations of genetic variation and environmental factors by considering the habitat suitabilities modeled from contemporary climatic variables, as well as the average habitat suitability of the current and past climatic conditions (i.e., an intermediate landscape shown in Fig. 4.2). We chose to use partial Mantel tests because all the explanatory factors are distance-based matrices (Legendre and Fortin, 2010) and the primary interest here is on the change in correlations between the predictor matrix and the genetic distance, and therefore the issues surrounding the interpretation of P-values with partial Mantel tests (Raufaste and Rousset, 2001) and reduced power in detecting relationships compared to dbRDA (Legendre and Fortin, 2010) is not a critical problem as applied here.

4.3.4 Incorporating spatially explicit demographic history into model tests with ABC analyses

We apply the iDDC-modeling approach so that we can examine if correlations between environmental factors or historical shifts in distributions might (or might not) reflect causal relationships with the processes governing population genetic structure. Specifically, with iDDC-modeling a population demographic model is used to make explicit predictions for patterns of genetic variation (Currat and Excoffier, 2004; Sork et al., 2010; Wegmann et al., 2006), where the population demography is informed by the underlying environment (i.e., it takes into account spatial and temporal heterogeneity of the environment in a species-specific manner; see details in Brown and Knowles 2012; Knowles and Alvarado-Serrano 2010). To test whether the current genetic structure results from (a) the geographic configuration of populations, (b) the contemporary environment, and (c) the dynamic history of shifting environmental characteristics associated with the differences between the present and the last glacial maximum, we constructed three corresponding demographic models and used coalescent simulations to predict genetic variations. In contrast with the studies to date utilizing iDDC-modeling, we then use these simulations for identifying the most probable model and estimating parameters using Approximate Bayesian Computation, or ABC (see Beaumont et al., 2002, for an overview of ABC).

The general procedure involves translating the habitat suitability scores from an ENM into spatially explicit population parameters for demographic simulations, which are then used for a spatially explicit coalescent simulation to generate expected patterns of genetic variation (for details about the procedures see Brown and Knowles, 2012; Knowles and Alvarado-Serrano, 2010). This flow of information provides direct links between process and pattern. Specifically for this study, we statistically downscaled the maps from the ENMs for the current and past climatic conditions to 0.1 decimal degree ($121 Km^2$ per cell) to have a tractable number of demes for demographic simulation. All spatially explicit demographic simulations were performed in SPLATCHE2 (Currat and Excoffier, 2004), with

population carrying capacities scaled proportionally to the local habitat suitability score (i.e., the relative values per grid cell differed depending on the predicted habitat suitability derived from the ENM generated with MaxEnt). Patterns of genetic diversity were then generated from coalescent simulations based on the specific demographic simulation (i.e., genetic variation differed across the landscape depending on the probability of coalescence and migration across demes)(Currat and Excoffier, 2004; Excoffier et al., 2000). We ran 24 coalescent simulations for each demographic history corresponding to each of the 24 separate loci in the empirical dataset (see Appendix C tables), such that these independent realizations of the coalescent process generated genealogies for simulating sequence data for each locus, where the sampled individuals from the simulated data sets matched those in the empirical data. DNA sequence data were also simulated according to the empirical DNA sampling conditions (e.g., the same gene length and amounts of missing data). Relative mutation rates among loci matched those from empirical estimations (Table C.6).

The three models tested here were selected to test hypotheses motivated by the correlative analyses described above (Fig. 4.2). Specifically, the hypotheses tested were that patterns of genetic variation reflect: (i) genetic drift associated with the geographic configuration of habitats C tested using a model of isolation-by-distance, or IBD, (ii) genetic drift associated not only with the geographic configuration of habitats, but also differences in local population sizes and the amounts of gene flow as defined by the suitabilities of contemporary environment C tested using a model of the contemporary ENM, or cENM, and (iii) genetic drift associated with distributional shifts caused by changes in environmental conditions C tested using a dynamic ENM model, or dENM. These models again differ with respect to input layers used for the demographic simulations (see Fig. 4.2). Note that the cENM model considers the impact of habitat heterogeneity on patterns of genetic variation, whereas the IBD model only considers the influence of geographic distance, but both of these models are static models in that the layer informing the demographic model does not change over time. In contrast, the dENM also considers how a shifting distri-

bution, and the accompanying colonization process, impacts patterns of genetic variation (Fig. 4.2). Temporal variation in habitat suitability was modeled in a step-wise fashion (i.e., using the habitat suitability scores from three period specific ENMs) (see also Brown and Knowles 2012). Specifics regarding the simulation details for all the models are given in the Supplemental methods. Note that each generation during the demographic simulation, m proportion of the population migrates out of the local deme; migration occurs to the adjacent four cells (north, south, west, east) and the allocation to different directions are defined by the friction score (see Supplemental methods); after exchange of individuals, populations grow logistically at the rate of 1 regulated by the carrying capacity inferred from habitat suitability. ENM maps and the settings for demographic modeling in Splatche2 are deposited in Dryad (to be added when we get the manuscript ID).

Model selection and parameter estimation were conducted using Approximate Bayesian Computation with ABCestimator in ABCtoolbox (Wegmann and Excoffier, 2010). We performed 1,000,000 simulations for each model under a standard ABC rejection sampling approach (Beaumont et al., 2002; Tavare et al., 1997). In addition to comparisons of the performances of different models, we also estimated four critical demographic/mutation parameters: maximum carrying capacity (Kmax), migration rate (m), ancestral population size prior to expansion (NAnc) and average mutation rate (μ), because each model would have different estimates of these parameters that generate simulated data closest to empirical ones. The ABC inference was based on a total of 34 summary statistics calculated within, between, and across all populations using Arlequin (see Table C.7 for the full list of summary statistics). They include segregating sites S for each population and across populations, private segregating sites for each population PrS, the mean number of pairwise genetic differences of each population , and pairwise population F_{ST} (Weir and Cockerham, 1984). In order to remove the effects of interactions between summary statistics, as well as reduce "the curse of dimensionality" (i.e., when too many statistics are included, the distance between the simulated and empirical values systematically increases, reducing the

accuracy of parameter estimates and making it more difficult to distinguish among models), partial least squares components (PLSs, Boulesteix and Strimmer, 2007) were extracted from all predictor variables. This treatment extracts orthogonal components from data with high dimensionality while maximizing the covariance of summary statistics and the parameters of interests (Wegmann and Excoffier, 2010; Wegmann et al., 2009). PLS were calculated in the "PLS" package (Mevik and Wehrens, 2007) with boxcox treatment (Box and Cox, 1964) in R for the first 10,000 runs for each model. The Root Mean Squared Error (RMSE) prediction of each parameter was examined before deciding upon the number of PLS components to be used (see Fig. C.1).

Five thousand simulations (0.5%) that were closest to the empirical observation were retained from each model for model selection. Post sampling regression adjustments were applied using the ABC-GLM function (Leuenberger and Wegmann, 2010) to obtain posterior distributions of the parameters, which assumes the accepted PLSs are produced by a General Linear Model from the parameters. We use Bayes factors for model selection, which is the ratio between marginal densities of two models. The higher the ratio is, the high the support for the first model is. Under the GLM model, the likelihood of the empirical data (i.e., the observation) can be evaluated and compared with the likelihoods of other retained simulations. The fraction of simulations that have a smaller likelihood than the empirical data was shown as P-value to check if the model is capable of generating the observed data. Very small P-values indicate that a model is highly unlikely (Wegmann and Excoffier, 2010). The coefficient of variation (R^2) of each parameter explained by the 6 used PLSs was computed as an indicator of the power of estimation (Neuenschwander et al., 2008). After selecting the highest supported model, we validated the accuracy of parameter estimation in the most supported model. 1000 pseudo observations were generated from prior distributions of the parameters. If the estimation of the parameters is unbiased, posterior quantiles of the parameters from pseudo runs should be uniformly distributed in [0,1] (Cook et al., 2006; Wegmann and Excoffier, 2010). The posterior quantiles of true

parameters for each pseudo run were also calculated based on the posterior distribution of the regression adjusted 5000 simulations closest to the pseudo observation. Average Root Mean Squared Error (RMSE) of the mode estimates for parameters of pseudo observations was calculated to check for the accuracy of estimation.

4.4 Results

The anonymous nuclear markers were all variable (see Table C.6 for summaries of molecular variation), but differed in the mutation rates (θ_π ranges from 0.009 to 0.1). These per-locus differences were incorporated into all correlative tests and simulations used to test hypotheses about the link between patterns and process with ABC (see below).

4.4.1 Associations between patterns of genetic divergence and environmental factors

A significant association between geographic distance and genetic differentiation was detected with both individual-level and population-level analyses (i.e., results from dbRDA and FST-analyses, respectively). Specifically, tests of isolation-by-distance with dbRDA explained 59% of the genetic variation among individuals (Table 4.1). A strong geographic signal was also evident from the regression of linearized F_{ST}-values against pairwise Euclidean distances between populations (Fig. 4.3a).

Contemporary climatic differences are also significantly associated with patterns of genetic divergence; however, when conditioned on the geographic distances between individuals (i.e., controlling for the effects of geographic isolation), the effects are not significant (Table 4.1). For example, although PC1 of the climatic variables was significant when tested alone, when conditioned on the geographic distance separating individuals, it was not, and the proportion of genetic variation explained decreased from 9% to 2% (Table 4.1). SoilPC1 was not significant irrespective of conditioning on geography. For tests

Table 4.1: Tests of an association between genetic distances with geographic distance and/or environmental differences (as captured by two sets of environmental predictors, climate-PC1 and soil-PC1) among sampling sites of individuals using dbRDA (see Fig. 4.1 for a map of sampling sites). Results are given for each geographic and environmental variable separately (i.e., the marginal tests), as well as conditioned on the effects of geographic distance (i.e., the relationship between the predictor and the response matrix controlling for geographic distance as a covariate) (see text for details). Shown are the multivariate *F*-statistics, associated *P*-values, and the percentage of variance explained by each variable; significant *P*-values are shown in bold.

	Marginal Tests			*Conditional Tests*		
Variable	F	P-value	*% variance*	*F*	*P-value*	% variance
distance	2.876	***0.005***	58.554			
climate-PC1	8.430	***0.010***	9.120	1.173	0.230	2.061
soil-PC1	1.981	0.160	2.304	0.736	0.550	1.293

of associations between population-level divergences and environmental differences separating populations as measured by an analysis of isolation-by-resistance (McRae, 2006) from habitat suitability scores for per-cell conductance, a significant association is detected (Fig. 4.3b, c). However, when controlling for geography with a partial Mantel tests (Table 4.2), the genetic differentiation actually shows an inverse relationship, as measured by the resistance from the current ENM alone, where genetic differentiation was greater for lower resistance (rather than a positive relationship between genetic differentiation and levels of resistance). We discuss this enigmatic pattern below, but further analyses suggest it could reflect the confounding influence of past environmental conditions. For example, when pairwise F_{ST}-values are regressed against pairwise resistance-scores calculated from a composite ENM map (i.e., the average habitat suitability scores from the ENMs of past and current climatic conditions; see Fig. 4.2), correlation coefficients are much higher than when only the current climate is considered (r = 0.84 versus r = 0.46; see Table 4.2), and remain significant if controlling for the effect from current climatic conditions (Table 4.2) using a partial mantel test. However, the highest correlation is with the average pairwise Euclidean geographic distances separating populations (Table 4.2), and the effect of the averaged habitat suitability over time (i.e., from past and current ENM) is

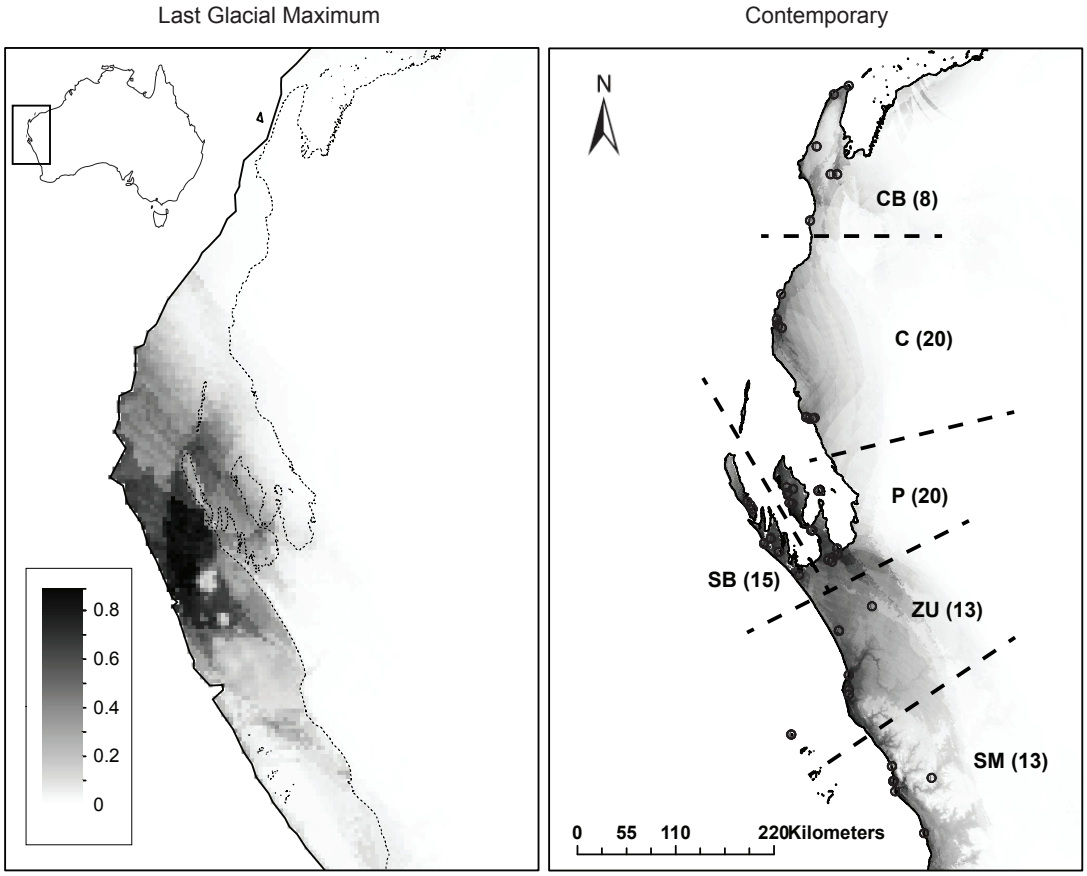

Figure 4.1: Predicted contemporary and past distribution of *Lerista lineopunctulata* in southwest Australia (see inset for location in continent) based on climatic and paleoclimatic variables, respectively (see text for details). Habitat suitability scores are shown as ranging from the lowest (lightest) to the highest (darkest) suitability. Dashed lines separate populations (as determined from barriers associated with breaks in suitable habitat; see Edwards et al. 2012) and population names along with sample sizes (in parentheses) are shown with dots that mark sampling sites. In contrast to the linearly distribution of suitable habitat along the coast today, refugial areas for the species 21kya were more circumscribed and extended westward of current populations SB and P (dashed outline marks the current coast line), given the emergence of vast areas of coastal sand habitats during glacial maximum (Hocking et al. 1987;Mory et al. 2003).

not significant after controlling for the influence of Euclidean geographic distances. Yet, it would seem highly unlikely that the Euclidean geographic distance among individuals or populations reflect dispersal patterns given the shape of the coastline and the distribution of species (i.e., the animals would have to traverse inhospitable habitat, including the

Table 4.2: Results of isolation by resistance as calculated using Mantel and partial Mantel tests (with geography and the current ENM as covariates) between the pairwise FST-values with geographic distances and resistance matrices (i.e., rescaled geographic distances according to the suitability of habitats) separating populations (see also Fig 4.3). Two resistance matrices are tested: the first is calculated from current habitat suitability score, and the second from the average of past and current suitability. Correlation coefficients (r) and the P-values from 1000 permutation tests are shown. In partial Mantel tests, covariates are listed on the second row; significant tests are shown in bold.

	Mantel Tests		Partial Mantel Tests			
			geography		current ENM	
Matrices	*r*	*P-value*	*r*	*P-value*	*r*	*P-value*
average pairwise Euclidean distance	0.868	0.005	-	-	-	-
resistance-values calculated from a map of habitat suitabilities from the current ENM	0.460	0.050	-0.729	0.007	-	-
resistance-values calculated from a composite map of habitat suitabilities from current and past ENMs	0.839	0.010	0.024	0.499	0.892	0.008

ocean, especially for populations P and SB; Fig. 4.1). This raises the question of whether the association of geography and genetic divergence actually arose under an isolation-by-distance model? Secondly, would a model of the population demography produce patterns of genetic variation that are likely to have arisen under isolation-by-distance? Lastly, even though the isolation by resistance takes into account possible paths (McRae, 2006), it does not take into account the demographic consequences of moving through the habitat. Consequently, could environmental factors (either present or past) actually impact patterns of genetic variation, but go undetected with correlative tests? The answers to these questions, which again are motivated by the aforementioned correlative analyses, are discussed in the following section.

4.4.2 Tests of the links between pattern and process

For the ABC analyses, we selected the first six PLSs for calculating the distance between simulations and the empirical observation because RMSE of the four parameters in four models does not decrease significantly with additional PLSs (see Fig. C.1). Based

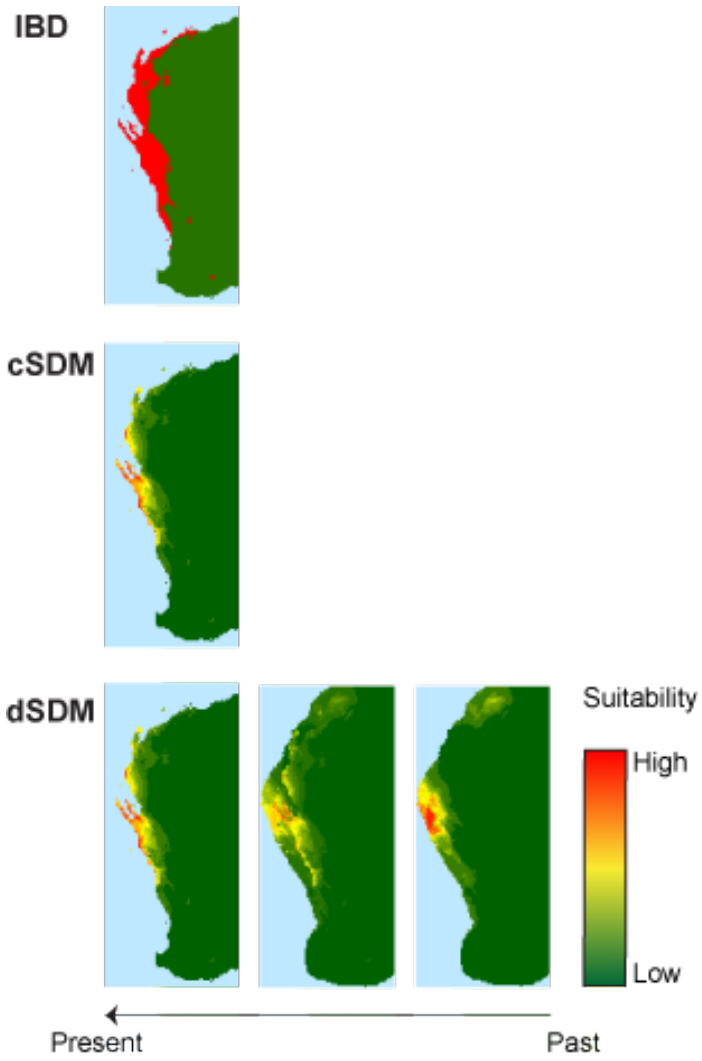

Figure 4.2: Schematic of the three spatially explicit models used in the demographic simulations to evaluate how environmental factors, as well as changing environmental conditions associated with the Pleistocene glaciation, might be causality related to patterns of genetic variation. For each model, variation in the underlying environmental components used for the demographic simulations is shown (see Knowles and Alvarado-Serrano 2010). The respective models are: (i) isolation-by-distance, or IBD, (ii) contemporary ENM, or cENM, and (iii) dynamic ENM, or dENM (as described in detail in the text); shown for each model is the spatially-explicit layer that formed the basis for the demographic simulations. Note that both the IBD and cENM models are static in the sense that the habitat suitability scores used for the demographic modeling were the same across generations, whereas with the dENM model is dynamic with habitat suitability scores changing over time from the last glacial maximum to the present in a step-wise fashion, as shown (see supplemental material for details). After each forward-time demographic simulation, coalescent simulations were run for sampled individuals backward in time.

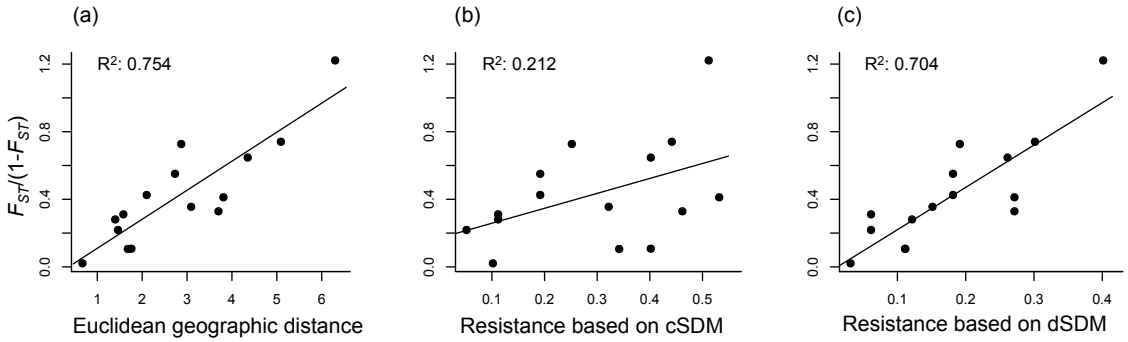

Figure 4.3: Plots of linearized F_{ST} against (a) pairwise population Eulidean geographic distance, and pairwise population resistance calculated from (b) contemporary habitat suitability, and (c) the composite suitability of past and present habitats (see text for details). Fitted line of the points and its R^2 are also shown.

on the marginal density for each model calculated from the 5000 closest simulations for each model, the dENM model, the model of the colonization history under dynamic ENMs (Fig. 4.2, best explains the patterns of genetic divergence observed within *L. lineopunctulata*. The two static models that only consider aspects of the current landscape, the IBD and cENM, have significantly lower marginal densities. For example, even though the cENM has much higher support than the IBD model, the difference in Bayes factors between the cENM and dENM is more than 300 (a substantial difference;(suggested by Jeffreys, 1961). Moreover, the dENM model has a high P-value suggesting a significant correspondence between the observed empirical data and the simulated data under this models, whereas the P-values for the cENM and IBD model are close to zero (Table 4.3). Based on these model comparisons, we may conclude that 1) demographic models that include dispersal regulated by habitat suitability produce models that explain the genetic divergence patterns within *L. lineopunctulata* (e.g., comparing the cENM and dENM to the unlikely IBD model); and 2) a demographic model that involves habitat shifts is much more likely to explain intraspecific divergence within *L. lineopunctulata* than static landscape models (e.g., comparing the dENM to the cENM) (Table 4.3).

Given the dENM best explains the data, analyses were conducted to validate the ac-

Table 4.3: Properties of models and the prior and posterior distributions of estimated parameters. Bayes factor is the ratio between the highest marginal density among models and that of each model. Kmax, carrying capacity of the deme with highest suitability; *m*, migration rate per deme per generation; μ, average mutation rate; NAnc, ancestral population size before expansion from the refugia. The logarithms of all priors are uniformly distributed and have the same prior ranges across models. R^2, the coefficient of determination between a parameter and the 6 used PLS components, shows the power of estimating certain parameters. HPDI 50 and 90 are the interval of 50% and 90% parameter regions with the highest posterior density respectively.

Models	Marginal Density (p-value)	Bayes Factor	Parameters	Prior [min, max]	R^2	Posterior		
						mode	HPDI 50	HPDI 90
IBD	2.14×10^{-14} (0.0002)	9.12×10^{8}	$log_{10}(K_{max})$	[3, 5.3]	0.080	3.465	[3.256,3.697]	[3.000,4.301]
			$log_{10}(m)$	[-4, -0.3]	0.012	-3.290	[-3.701,-2.767]	[-4.000,-1.832]
			$log_{10}(\mu)$	[-8, -6]	0.223	-6.040	[-6.061,-6.000]	[-6.162,-6.000]
			$log_{10}(N_{Anc})$	[3, 5]	0.546	3.848	[3.768,3.929]	[3.626,4.071]
cENM	5.82×10^{-8} (0.0216)	334.72	$log_{10}(K_{max})$	[3, 5.3]	0.145	3.000	[3.000,3.116]	[3.000,3.279]
			$log_{10}(m)$	[-4, -0.3]	0.045	-3.851	[-4.000,-3.589]	[-4.000,-3.028]
			$log_{10}(\mu)$	[-8, -6]	0.910	-6.000	[-6.061,-6.000]	[-6.061,-6.000]
			$log_{10}(N_{Anc})$	[3, 5]	0.744	3.929	[3.849,4.030]	[3.667,4.192]
dENM	1.95×10^{-5} (0.1514)	-	$log_{10}(K_{max})$	[3, 5.3]	0.046	4.975	[4.487,5.230]	[3.604,5.300]
			$log_{10}(m)$	[-4, -0.3]	0.541	-1.571	[-1.870,-1.309]	[-2.243,-0.898]
			$log_{10}(\mu)$	[-8, -6]	0.915	-6.242	[-6.343,-6.142]	[-6.505,-6.020]
			$log_{10}(N_{Anc})$	[3, 5]	0.737	4.414	[4.293,4.535]	[4.131,4.717]

curacy of the dENM. The estimation accuracy of the four parameters differs significantly (Table 4.3, Fig. 4.4). The posterior probability of maximum carrying capacity (Kmax) is much flatter than that for the other three parameters (notice the density of the highest peak; Fig. 4.4) and there is limited power to estimate carrying capacity, as indicated by a R^2 = 0.046 (Table 4.3) and RMSE plot (Supplemental Fig. 4.1). Testing of estimation bias of the parameters shows that the posterior distribution of K is too narrow and that of μ is too wide (Fig. 4.5, histograms of the posterior quantiles significantly deviate from a uniform distribution after Bonferoni correction for multiple testing, $P - value < 0.01$). The other two parameters are more or less uniformly distributed so that migration rates (*m*) and ancestral population size (NAnc) before expansion are the better estimated parameters from the set of four parameters. The ancestral population of the species is estimated to be about 26,000 (Fig. 4.4), and the mode of the migration rate is about 0.0.027 per 10 years per deme (121 km2), that is, about 3% of the population per deme emigrates in 10 years.

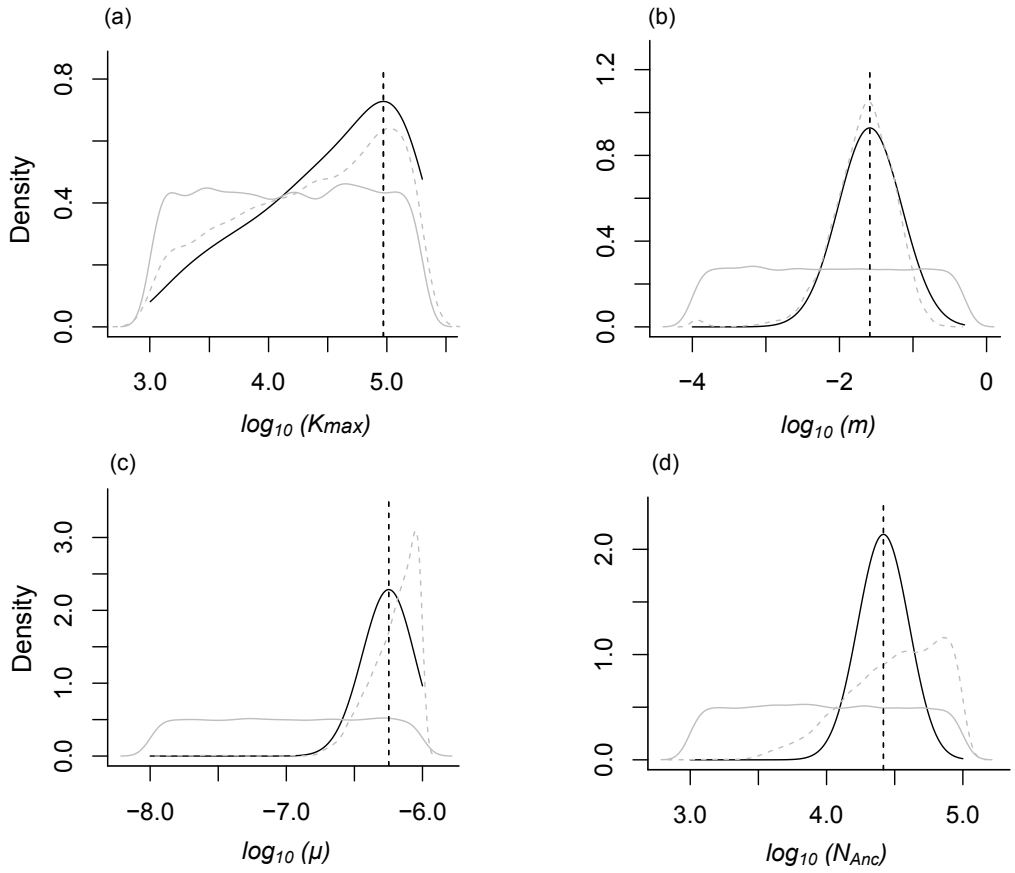

Figure 4.4: Posterior distribution (shown as dark line) and mode (i.e., the vertical line) of parameter estimates for the most probable model - the dENM - based on an GLM regression adjustment of the 5000 closest simulations (see text for detail). The distribution of retained simulations (shown as dashed line) and the prior (shown as grey line) are given to highlight: (i) the improvement the GLM procedure introduced on the parameter estimates (i.e., comparing the dashed and solid dark lines), and (ii) that the data contained information relevant to estimating the parameters (i.e., contrast the solid dark and grey lines).

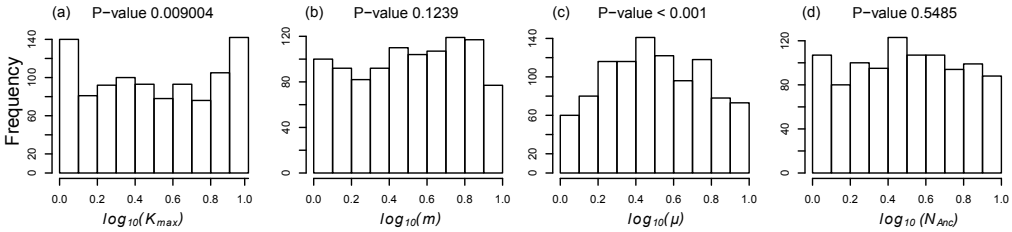

Figure 4.5: Distribution of posterior quantiles of parameters for the most probable model - the dENM - for evaluating potential bias in the parameter estimates, as measured by a departure from a uniform distribution using a Kolmogorov-Smirnov test; analyses are based on 1000 pseudo-observations. Estimation of m and N_{Anc} seem to be unbiased while posterior distribution of K is too narrow and that of μ is too wide

4.5 Discussion

Detecting spatial structure and identifying correlates of patterns of spatial structure are themselves arguably important endeavors and have received enormous attention in landscape genetics (Guillot et al., 2009; Storfer et al., 2007, 2010). What has yet to be fully explored, and remains under-developed, are statistical frameworks for exploring the links between such patterns and the processes that capture biological phenomena critical to addressing issues such as how the environment shapes patterns of genetic variation within species (Balkenhol and Landguth, 2011; Cushman and Landguth, 2010; Shirk et al., 2012) through the modeling of expected patterns of genetic variation.

Our study highlights the need for expanding the traditional perspective and foci of landscape genetics (as discussed below), while also presenting one approach for establishing and testing the links between pattern and process. As such, the study represents a promising new direction for expanding landscape genetics studies. Nevertheless, we also recognize aspects of such complex models that would greatly benefit from further attention. These are discussed with the intention of motivating future development, but also of drawing attention to aspects of the analyses that should be interpreted cautiously.

4.5.1 Importance of exploring the links between genetic patterns and process in landscape genetics

Different demographic processes may lead to the same genetic patterns (Csillery et al., 2010). As a consequence, an explicit model that can generate explicit patterns of genetic variation under different scenarios is critical (Knowles, 2008, 2009). For example, patterns of shared polymorphism may not necessarily reflect recent hybridization (Green et al., 2010), but instead be a case of incomplete lineage sorting, especially if ancient population structure contributes to longer coalescent times (Eriksson and Manica, 2012). Likewise, although geographic barriers might generate substantial genetic differentiation among pop-

ulations (Knowles, 2001), such a pattern might be generated without geographic isolation through the colonization process associated with climate-induced habitat shifts (Knowles and Alvarado-Serrano, 2010, e.g.,). With explicit modeling, a spectrum of factors capable of producing the observed genetic patterns can be explored, thereby cautioning against interpretations without considering alternative processes.

As this study demonstrates, the essential importance of modeling the link between genetic pattern and process also extends to interpretations of the role of the environment in shaping patterns of genetic variation. This is perhaps exemplified best by contrasting the conclusions that might have been drawn from the descriptive association between genetic and geographic distance with the actual likelihood of an isolation-by-distance as the most probable model explaining the data. Specifically, a highly significant association is detected with both individual and population level analyses between genetic and geographic distances (see Fig. 4.3 and Tables 1 and 2). Moreover, the effects of environmental variables, whether measured from current or past climatic variables, become inconsequential after controlling for the effect of geography, reinforcing that geography may indeed be the primary determinant of patterns of genetic variation, rather than aspects of the environment. Yet, the ABC tests clearly show that genetic drift associated with geographic distance alone (i.e., the IBD model; Fig. 4.2), is significantly less likely than models that consider varying aspects of population connectivity as impacted by environmental variables (i.e., the cENM and dENM are both more probable; see Bayes factors in Table 4.3), with the dENM that takes into account shifting habitat suitabilites over time as the most probable. So what explains this apparent contradiction between the conclusions that might be drawn from the descriptive patterns of genetic variation (see also Edwards et al., 2012) versus the models of the actual processes involved? Could such discrepancies reflect problems with the modeling procedure, such as biases in parameter estimates? These issues and their relevance in terms of the biological implications for *L. lineopunctulata* are discussed below.

4.5.2 Demographic modeling as a tool for evaluating and interpreting genetic correlations

Many statistical analysis tools have been developed to examine the correlation between genetic variation and geographic and/or environmental factors (e.g., Adriaensen et al., 2003; ter Braak and Verdonschot, 1995; Epperson and Li, 1996; Lee and Mitchell-Olds, 2011; Legendre and Anderson, 1999; Legendre et al., 2002; Mantel, 1967; McRae, 2006; Smouse et al., 1986; Wang et al., 2012). Although these approaches differ with respect to their statistical power to detect important factors (Legendre and Fortin, 2010), none actually models the underlying processes generating the patterns (Balkenhol et al., 2009; Meirmans, 2012). Although a virtue in some respects (e.g., such approaches are generally broadly applicable and are not particularly computationally demanding), there are also inherent limitations with respect to (i) evaluating how such factors might produce patterns of genetic variation, or (ii) distinguishing among alternative hypotheses about the putative factors underlying patterns of genetic variation.

The merit of model-based inferences has become widely accepted in studies of genetic data (Knowles, 2009; Knowles and Carstens, 2007), especially with increased knowledge about the high variance of mutational and coalescent processes (Hudson, 2002). As computational constraints are overcome algorithmically and with improved computing resources, the incorporation of biological realities has become feasible. For example, methods that model genetic diversity and divergence at the same time and regress against environmental factors (Faubet and Gaggiotti, 2008; Foll and Gaggiotti, 2006) can be used to evaluate which environmental factors (if any) might influence genetic divergence; although such models cannot control for spatial autocorrelation among factors (in contrast to the approach used here). Moreover, the flexibility and versatility of tools for evaluating and interpreting models (Itan et al., 2009; Jaquiry et al., 2011; Neuenschwander et al., 2008, e.g., with ABC:) can expand the repertoire of biological models that might be considered. This includes the incorporation of factors that are typically overlooked in descriptive correlative approaches

(e.g., dbRDA, PCA, MDS), such as changes in population size and/or distribution, because of difficulties with their incorporation.

This later point, we argue, may underlie some of the apparent discrepancies in the relative importance of geography compared to environmental factors in the descriptive versus model-based approach applied here. Specifically, even rescaled distances that incorporate aspects of the environment for testing for an association between environmental differences and genetic variation make a number of implicit simplifying assumptions. For example, even though a method like McRaes (2006) isolation-by-resistance considers multiple possible paths (as opposed to the least-cost path (Adriaensen et al., 2003), and gives a weighted average of the connectivity between the populations, this approach is only valid when landscape does not change over time. However, when a habitat is less stable over time, the level of population connectivity changes depending on the impact of habitat shifts on dispersal dynamics and population sizes (see Brown and Knowles 2012). These demographic consequences that are a direct extension of the underlying environment would necessarily impact patterns of genetic variation (i.e., changes in migration probabilities and local population sizes would impact the relative probabilities of gene lineage coalescence within demes and the times to coalescence, Excoffier et al., 2009a).

Consequently, when we actually model the demographic process of population movements across a landscape, whether it follows an IBD model where the environment does not impact population demographic patterns or one of the alternative models in which the environment does influence migration rates and deme sizes (e.g., the cENM and dENM; Fig. 4.2), it is perhaps not surprising that the results from the ABC analyses (Table 4.3) and the descriptive correlative analyses (Tables 4.1 and 4.2) do not match up. However, can the results from the model-based tests be trusted justifying the tradeoff between the simplicity of correlative analyses and the complexity of process-based models?

4.5.3 Model interpretation, validation, and implications for the factors structuring genetic variation

Model validation is very important in ABC given that ABC approximates the likelihood of models with summary statistics (Beaumont et al., 2002; Pritchard et al., 1999), unlike full likelihood based models that use all of the data (Beerli and Felsenstein, 2001; Hey, 2004, 2010; Hey and Nielsen, 2007; Kuhner, 2006; Kuhner et al., 1998; Nielsen and Beaumont, 2009). Post-sampling adjustment, such as regression (Beaumont et al., 2002) or GLM (Leuenberger and Wegmann, 2010), can pose problems when the relationship between parameters and summary statistics is extrapolated beyond the region of the observed dataset. Moreover, ABC can always produce posterior distributions even if the model is wrong (Bertorelle et al., 2010).

Given the model complexity, one of the concerns was whether the data would be sufficient to discriminate among probable and relatively improbable models, as well as give unbiased parameter estimates. Nevertheless, the several approaches used to validate the models in this study suggest that the results are generally robust.

Our primary objective is to test alternative demographic models (as opposed to a focus on specific parameter values), therefore, we used a standard rejection sampling scheme (Beaumont et al., 2002). Although the method requires a longer computational time than other methods, such as ABC-MCMC, population Monte Carlo (PMC), and adaptive PMC (Beaumont et al., 2009; Moral et al., 2012; Wegmann et al., 2009), it does not create bias among models since performance of Monte Carlo methods are sensitive to the choice of tolerance level and proposal range (Wegmann et al., 2009). To show the support of the models, comparison of marginal densities of each model, as measured by Bayes factor alone is not enough. Rather, the P-value of observed data under the GLM model also needs to be checked to examine the percentage of the simulated data that match the empirical data. IBD and cENM models can be easily rejected based on the Bayes Factor (Table 4.3). In addition, the dENM has a higher probability of generating simulations with smaller or

equal likelihood than the empirical observation, compared to the cENM (see P-value in Table 4.3). In other words, even though some idiosyncratic combination of parameters can produce datasets that match the data under the cENM, the dENM has much wider parameter regions that generate data close to the observation, which is the prerequisite for accurate parameter estimations. Posterior distributions of the parameters in the two models only differ significantly in the estimation of Kmax. However, Kmax has the least power to be informed by the PLSs in the ABC analyses (see $R^2 < 0.1$ in Table 4.3). The estimation of maximum carrying capacity Kmax and average mutation rate μ show some level of bias in estimation based on the tests of uniformity of posterior quantile distributions from pseudo-observations (Fig. 4.5) in that posterior distribution of K is too narrow and that of μ is too wide. Because both of the two parameters are hyper priors that control the change of a series of local parameters, it might be harder to estimate them accurately (Wegmann and Excoffier, 2010). This contrasts with the two parameters, migration and ancestral population size, that are estimated well with low average RMSE of mode (0.19 and 0.15 respectively) and not biased (Fig. 4.5).

We acknowledge that the models informed by the ENMs may not capture all the potential historical scenarios that might be tested. However, this is a huge improvement over simple generic models that limit biological insights (Bertorelle et al., 2010; Knowles, 2009). Moreover, the class of models generated from the iDDC approach, especially the incorporation of information from past distributions, permits tests of hypotheses that could not otherwise be identified (e.g., the impact of climate-induced distributional shifts on patterns of genetic variation) (see also Hugall et al., 2002; Knowles and Alvarado-Serrano, 2010; Moussalli et al., 2009; Strasburg et al., 2007). There may also certainly be other aspects of ENMs that introduce error into projected species distributions (see Arajo and Guisan, 2006; Graham et al., 2004; Lozier et al., 2009; Phillips et al., 2006; Stockwell and Peterson, 2002), this is an active area of research and the field of ENMs will no doubt see significant advances in the near future. Again, despite these sources of errors, we argue the potential

gains outweigh the negatives (which again we note, should become minimized with the advances in ENMs). With respect to *L. lineopunctulata* specifically, this includes avoiding the misleading conclusions that would have resulted from extrapolating causation from descriptive correlates (i.e., only geography was consistently identified as a primary factor structuring variation; Tables 1 and 2) or considering a limited sphere of models (i.e., IBD was the least probable model, which was only apparent with the inclusion of the additional ENM-based models; Table 4.3).

Lastly, despite the aforementioned caveats regarding the models and estimation of parameters, the iDDC modeling procedure which infuses the coupled ENM and coalescent models (Knowles and Carstens, 2007; Richards et al., 2007) with ABC represents an intriguing new advance beyond past applications (Balkenhol and Landguth, 2011; Brown and Knowles, 2012; Cushman and Landguth, 2010; Heckel et al., 2005; Knowles and Alvarado-Serrano, 2010; Morgan et al., 2011; Shirk et al., 2012). Moreover, what this study highlights is the synergy between more traditional landscape genetic approaches and these model-based inferences for addressing the critical issue in model-based inference C how to identify models to be tested (Knowles, 2009). Our study shows the intriguing possibilities of using the descriptive approaches from landscape genetics, which detect associations between genetic and environmental factors, for developing suites of alternative hypotheses that can be translated into models for testing with ABC.

4.5.4 Biological implications and the importance of integrating historical and contemporary environments

With limited information from the lack of ecological study of *L. lineopunctulata*, this study can provide important biological insights. It is noteworthy that because of the ABC framework, we can evaluate the probability of fairly contrasting views on the population demography of this lizard species. Specifically, with endemism along the coast (Fig. 4.1), the relatively high F_{ST}-values (i.e., values above 0.095, Table C.8, except for the compari-

son between the two historically stable regions, P and SB; (see also Carnaval et al., 2009) could be explained by different combinations of parameter. The high divergence level could reflect the lack of migration with small population sizes (the expected pattern under an IBD model), restricted migration due to barriers associated with the contemporary habitat configuration, or colonization associated with a shifting species distribution, as a habitat specialist would track climate-induced habitat shifts. All are plausible hypotheses for *L. lineopunctulata*, an abundant subterranean lizard restricted to sandplain and dune habitats of coastal southwestern Australia, a region subject to pronounced climate shifts during the Pleistocene (Fig. 4.1). The genetic data suggests that *L. lineopunctulata* exhibits fairly strong habitat specialization such that (i) not only is the IBD model unlikely compared to those incorporating an environmental component, but (ii) that the species most likely tracked shifts in their habitat as climate changed from the last glacial maximum (i.e., the dENM model is more probable than the cENM). Because the dynamic model that accounts for shifting species distributions (dENM) is more probable than a static model of the contemporary landscape (the cENM), the population parameter estimates from the ABC analyses also suggest that *L. lineopunctulata* has higher ancestral population size ($\sim$26,000) and much higher migration rate ($\sim$0.03 per 10 years) than if only the contemporary landscape had been considered (contrast estimates for dENM and cENM in Table 4.3). This could have ramifications for developing effective conservation management plans, supporting initiatives for preserving the processes contributing to genetic divergence (Moritz and Faith, 1998).

As a recognized biological hotspot (Cincotta et al., 2000; Myers et al., 2000), our findings provide some valuable perspective on not only the factors promoting divergence within the focal species, but perhaps also those promoting diversification. A combination of an expanded sandplain habitat caused by late-Quaternary sea level fluctuations, local geological activity, and climate-induced distributional shifts are postulated to have driven diversification of the southwest Australian herpetofaua (Edwards, 2007; Hopper and Gioia, 2004;

Melville et al., 2008; Rabosky et al., 2004; Storr and Harold, 1978, 1980). Yet, the lack of detailed spatially and temporally explicit hypotheses have made it difficult to generalize how the SW Australian fauna would have been impacted by past geologic and climatic factors. Within the geographic area of study are a number of other endemic lizard species, many of which are co-distributed with *L. lineopunctulata*, but also show a variety of ecological preferences, despite occupying similar habitats (Cogger, 2000). This raises the question of whether this lizard community has responded similarly to past climatic events, or whether species-specific responses have predominated (Edwards et al., 2012). It may be that *L. lineopunctulata* has a higher dispersal ability compared to other Lerista species, for example, which lack both forelimbs and hind limbs (Bush, 2007; Cogger, 2000; Wilson and Swan, 2008). The iDDC approach applied here could be expanded into a comparative analysis, where species-specific characteristics (e.g., differing degrees of habitat specialization or vagility) can be taken into account when testing sets of biologically informed models for landscape genetic study.

CHAPTER 5

Conclusions and Future Directions

This thesis provides a synthetic framework for studying ecological genomics, which considers selective processes (such as adaptation to new niches) and neutral processes (such as population size changes due to environmental shifts) simultaneously. Conventionally, studies that look for targets of selection on a genome assume a simple demographic model without validations from the species' ecological or phylogeographic histories (Beaumont and Balding, 2004). I demonstrate that one cannot reliably identify selection unless realistic demographic histories are inferred for the species or even a specific genomic region (Chapter 2). The thesis also extends predictions on the sources of adaptive alleles and mechanisms for maintaining adaptive changes (Chapter 2 and 3). The findings are particularly important for understanding mosaic genomic evolution in the early stages of speciation where accumulation of divergence is dampened by gene flow (Nosil and Feder, 2012). The work described in the thesis represents a flexible framework for researchers to translate dynamic phylogeographic hypotheses into testable coalescent models by integrating all available information of the species, such as distribution records, habitat preference, paleo-climate models, and competition between species (see iDDC modeling in Chapter 4). The approach will be useful for examining the influence of historical events on current population genetic structures (He et al., 2013). In general, with the amount of information as well as inherent heterogeneity of genomic data, this thesis contributes to the ongoing paradigm shift from studying separate evolutionary processes towards a holistic analysis

of the interactions of selective and neutral processes under a rigorous statistical framework (Excoffier et al., 2013; Ronen et al., 2013; Yang et al., 2012). Below, I summarize how my work provides new knowledge on the following questions. In addition, I discuss limitations of my approaches as well as future directions.

- How do we infer selection when underlying demographic histories vary for different genomic regions? Why is it important to identify selection within inversions?
- What is the most important source of adaptive variation: new mutations, standing variation or adaptive introgression? Can we distinguish the genetic signature left by different sources?
- How do we model and test spatially-explicit demographic histories? What are the signals of range expansion and contraction in genetic data?

5.1 Detecting selection with the consideration of varying demographic histories

Study of selection is a classic, yet active, theme in evolutionary biology (Nielsen et al., 2007). With the widespread availability of genomic data, an approach that detects putatively selected loci by comparison to background levels of differentiation (termed F_{ST} outlier analyses) has gained popularity (Beaumont and Balding, 2004). Specifically, such approaches are frequently applied when populations inhabiting different environments are under divergent selection (Nosil et al., 2009). It is to be expected that between ecologically heterogeneous populations, loci that presumably contribute to adaptive divergence will exhibit levels of differentiation much higher than the background level of differentiation generated by genetic drift of geographic isolation. Therefore, generating correct expectations for background levels of differentiation is critical to a reliable detection of "causative" alleles. Existing methods, however, either have an inherent assumption for a simple demo-

graphic model which might not capture the characteristics of the species' history (Beaumont and Nichols, 1996) or treat the genome as a homogenous pool evolving neutrally under a single common demographic history (Excoffier et al., 2009b; Foll and Gaggiotti, 2008). These problems may contribute to a preponderance of false positives that may provide misleading results about the target of selection and the prevalence of loci involved in adaptive divergence. In Chapter 2, I designed new approaches to identify selection within inversion regions, which usually evolve under very different demographic histories compared to the collinear regions, with the intent to lower false positive rate as well as increase detection power. The innovation of my method is three fold. First, separate demographic histories are inferred for inversion and collinear regions to generate region-specific background levels of differentiation. Second, genetic measures other than differentiations, such as haplotype and genetic diversity, are also included in the analyses to increase the power of distinguishing adaptive alleles from neutral ones. Third, discriminant functions are used to predict selection instead of an outlier approach by using training datasets from neutral versus selection simulations. By modeling genomic regions with varying demographic histories, my approach allows for identifying targets of selection through discriminant function analyses trained with a suite of genetic measures, where traditional FST outlier analysis is not accurate. Growing evidence supports the prevalent role of inversions in facilitating adaptive divergence by reducing recombination among co-adaptive alleles (see details in Chapter 1.1) (Kirkpatrick, 2011). It is, therefore, important to identify adaptive alleles within inversions. Yet, inversions, acting as a supergene, represent hindrance to detection of selection using conventional methods (Thompson and Jiggins, 2014). On one hand, the functional importance of inversions is usually easily detected through natural clines (Balanya et al., 2003), alternative phenotypes (Joron et al., 2011) or experiments (Gray et al., 2009). On the other hand, due to the high linkage disequilibrium within inversions, the background differentiation is elevated across all genes within inversions as compared to other parts of the genome (Cheng et al., 2012). Thus, adaptive alleles cannot be easily

distinguished from neutral ones. In addition, coalescent patterns of neutral genes within inversions are different from counterparts in collinear regions in that recombination is free within the same karyotypes but suppressed to a different degree among different karyotypes depending on the distance of marker's location to the breakpoints (Guerrero et al., 2012). Our study is the first attempt to identify selection signature in an empirical system using inversion-specific coalescent expectations (Chapter 2). My study also highlights the need for a more advanced coalescent simulator for inversions. Peischl et al. (2013) developed a coalescent simulator for inversions restricted to specific scenarios with limited number of individuals. In my future research, I will extend the simulator to be able to incorporate arbitrary histories and generate SNPs or other markers for genetic diversity comparisons. This will facilitate the theoretical predictions and empirical testing on inversions.

5.2 The source of adaptive variation

When species explore new niches, the adaptation usually involves genetic changes. Facing the challenges of new environments, species can utilize different sources of genetic variation - that is, new mutations, standing variation (existing rare variants), or "borrowed" alleles from other species (i.e., adaptive introgression). It is not clear, however, how different sources of adaptive variation determine the rate and success of such adaptations under different scenarios (Barrett and Schluter, 2008).

Chapters 2 and 3 provide new insights on the conditions under which each of these sources is more likely to contribute to adaptive changes. Previous studies (Hermisson and Pennings, 2005) and my work (Chapter 3) have shown that adaptation from standing variation, if present, is faster and more likely compared to that from new mutations because of no waiting time and higher initial frequency of standing variation. However, standing variation does not dominate the source of adaptation in all scenarios. In Chapter 3, I showed that, counterintuitively, new mutations become a more important source when the selection

for such adaptation is higher. This is because the strength of selection on the adaptive alleles is more important for their survival than the initial frequencies. Thus, this work implies that additional benefits must be associated with standing variation compared to new mutations when empirical examples of adaptation from standing variation are observed. One advantage of standing variation is that such variation has been pretested by selection in the genetic background of the species. This contrasts with new mutations, which may interact negatively on the genetic background of the individual where the mutation arises, making them less likely to contribute to adaptation.

Adaptive introgression, formerly overlooked in animals, is gaining more attention, especially with new molecular and analytical methods that are able to distinguish the origins of adaptive alleles (Hedrick, 2013). Not surprisingly, adaptive introgression often occurs when one species is expanding into the range of another locally-adapted species. Environmental adaptations often require multiple changes in one gene or several genes. The waiting time for multiple new mutations is exceptionally long. Pre-existence of such standing variations is also rare unless the species had experienced similar selection in the past. In this case, ”borrowing” alleles from the locally-adapted species that have accumulated multiple genes for the specific adaptation becomes more likely.

As one of the most severe malarial vectors, *Anopheles gambiae*, has a large effective population size and short generation times, so that new mutations are not limiting in the species. This raises the question of why chromosomal inversions that capture locally-adapted genotypes appear to be shared among species (i.e., why is adaptive introgression so prevalent in the group)? My work suggests that a key to understanding this apparent conundrum is what appears to be high gene flow among populations. As I showed in Chapter 2, geographic structuring in the species' collinear regions is minimal. This demographic setting makes it difficult for co-adaptive genotypes to be retained - that is, locally adaptive alleles will be swamped by an influx of non-adaptive alleles under different environmental/ecological conditions. The introgression of inversions from the arid-adapted *An.*

arabiensis to *An. gambiae* (Besansky et al., 2003) brought not only adaptive alleles, but also the genetic mechanism to keep them together and reduce the loss via the mixing with non-adaptive combinations of alleles.

Thanks to years of karyotyping studies and molecular sequencing of inversion breakpoints (Coluzzi et al., 1979; Sharakhov et al., 2006), we are able to determine the origin of inversions in *An. gambiae*. However, for many adaptive loci, it is usually hard to distinguish whether they are retention of ancestral polymorphism, introgression or convergent evolution of new mutations. The heterogenic nature of genomic evolution of many non-model organisms is yet to be discovered with the increasing amount of genomic data. As a future direction, new statistical tests or coalescent models can provide answers to these questions that is more broadly applicable to any species for which there is some basic knowledge about the phylogenetic history of the taxa.

5.3 Spatially-explicit demographic inference

Past climatic cycles result in range expansion, contraction, and recolonization in species that track specific environmental niches (Excoffier et al., 2009a). Whether different genomic regions preserve signatures of these events depends on their evolutionary rates. Fast evolving markers may only display patterns of recent migration-drift equilibrium in populations, while slow evolving markers retain signatures of the deep past. With the inclusion of a spectrum of markers across the genome, we can infer very complex demographic histories with greater accuracy. Given these vast and complex data, how can we generate testable hypotheses to specifically estimate the influence of historical events on the current population genetic pattern? In Chapter 4, I developed iDDC (integrative Distributional Demographic Coalescent) modeling (see details in Chapter 1.3), a flexible framework to generate more realistic evolutionary models by integrating all available information of the species, such as distribution records, habitat preference, paleoclimate models, and competition between

species. In addition, instead of using full likelihood models (e.g., IM program; Hey, 2010), I conduct model selections through Approximate Bayesian Computation (ABC) (Beaumont et al., 2002). Analytic formula of the likelihood function is infeasible for complex models, while ABC methods approximate the likelihood through simulations.

Therefore, the biggest advantage of iDDC modeling is its flexibility, with the intent of generating species-specific phylogeographic models. The application is very wide in phylogeography, population history and conservation genetics. The approach highlights the importance to have a good knowledge of the ecology of the taxa in order to build reliable distributional/connectivity models. We demonstrated the success of this methodology to identify the importance of p temporarily dynamic landscape during Pleistocene in structuring the genetic variation of an Australian lizard, Lerista lineopunctulata, in Chapter 4.

As future work, with the spatial and temporal explicit features of iDDC modeling, I will apply this method to identify important determinants of disease vectors' spatial demographic history. As *Anopheles* mosquitoes are highly associated with human settlements and agriculture fields, it is of interest to test the contribution of environmental factors versus anthropogenic forces in shaping the current genetic patterns of *Anopheles* species. Researchers have long suspected that the genus experienced pronounced demographic and range expansion from late Pleistocene to Holocene (Ayala and Coluzzi, 2005). However, it is not clear whether rapid increase in human land use (agriculture and pasture) around 4000 years ago in Africa (and earlier in Asia) resulted in similar demographic/range expansion time and magnitude across different *Anopheles* species. Using iDDC modeling, range expansion through environmental and/or land-use changes will be built as spatially-explicit coalescent models, compared with summary statistics from empirical genetic data and selected in an ABC framework.

Several important questions related to *Anopheles* vector ecology can be answered by a comparative phylogegraphy study of *Anopheles* species using iDDC: 1) whether the contribution of human land use change is correlated with the species' affinity for feeding on

human blood (Zahar, 1990)? 2) Whether the stability of the species environmental niche (i.e., if there is big range shifts or contraction from the last glacial maximum) influences the current prevalence of the species (Carnaval et al., 2009)? 3) The results can also inform which environmental factors or land use intensities are good predictors for the projection of future distributions of *Anopheles* species under global warming scenarios (Sinka et al., 2010). Thus, by studying past demographic histories in a comparative and spatially and temporally-explicit framework, we can not only understand the vectors' habitat adaptation histories but also predict their regional epidemiological significance in the future. Preliminary results showed that climatically suitable areas did not go through dramatic expansion or contraction based on niche modeling reconstruction in *Anopheles gambiae* and *Anopheles dirus* complex during the last glacial maximum, while the land use model predicted that suitable areas dramatically increased from 3000 years ago (Fig. 5.1), with the magnitude high in Africa than Asia. A comparison of range expansion signals in empirical genetic data will elucidate whether the species tracked the environmental or land-use changes.

To conclude, the thesis aimed to identify processes that generated patterns with the inclusion of biological realism. The thesis took advantage of the increasing availability of large genomic data, advancement in statistical methods as well as integrated information from cytology, ecology, paleo-climatology, and anthropology. As I stated in the thesis introduction, in the era of ecological genomics, this work is only the beginning of a new methodological burst.

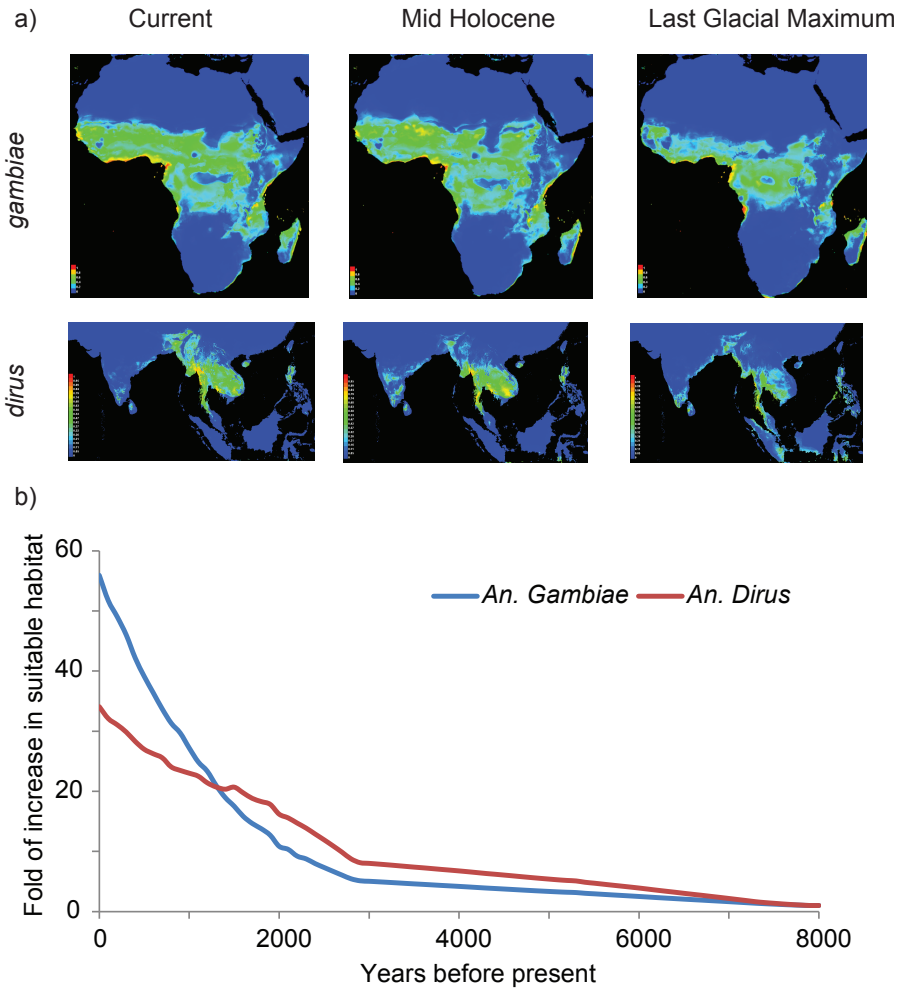

Figure 5.1: Habitat reconstruction of *Anopheles* species based on Ecological Niche Modeling (ENM) and Human land use changes. (a) Current and future predictions of species' distribution using MaxEnt model (Phillips et al. 2008) based on habitat suitability. (b) folds of increase in land use (i.e., pristine habitats converted into agriculture and pasture lands) in climatically suitable regions for *Anopheles gambiae* and *An. dirus*.

APPENDIX A

Supplementary material of Chapter 2

A.1 Supplementary results

Population genetic structures on collinear and inverted regions 667 *Anopheles gambiae* complex mosquito samples were collected at six sites in Cameroon (Fig. A.1, coordinates listed in Table A.1). They represent gradient changes in habitats (from forest, wet to dry savanna), which boast the highest diversity of polymorphic inversions within the species (previous collection data from PopI database: https://grassi2.ucdavis.edu/). Molecular identification showed that proportions of *An. arabiensis* increase from south to north. Only 6 An. colluzzi were found (Coetzee et al., 2013, previously known as the M form of *An. gambiae*; [[) in Mbe, while the rest were *An. gambiae* (previously known as the S form of *An. gambiae*). We selected 259 *An. gambiae* individuals (40-60 individuals with good DNA qualities per population except for Bankim, Table A.1) and 8 An. arabiensis, prepared individually-barcoded double digest Radseq libraries (Peterson et al., 2012), and generated two lanes of 100bp paired-end sequencing reads on Illumina HiSeq2000 platform.

After filtering for unmapped or ambiguously mapped reads and loci with low coverage per sample or low presence across samples, a total of 25,966 loci were mapped onto Chromosome 2, 3 and X. *An. gambiae* has a nucleotide diversity (π) of 0.01024 4.0E-5, while π of *An. arabiensis* is 0.00888 4.7E-5. One SNP per locus with at least 1000bp between them were selected for PCA analysis to detect population genetic structure of the species.

Table A.1: Collection sites, coordinates and sampling sizes of Anopheles gambiae.

Population	Latitude	Longitude	Sample size
MBE	7.78	13.55	89
NGOUNDERE	7.48	13.55	74
MEIGANGA	6.55	14.26	60
MBAKAOU	6.37	12.76	63
BANKIM	6.05	11.40	4
BAFOUSSAM	5.48	10.59	106

When PCA were performed on all individuals in collinear regions, we found three clusters of individuals consistently across all chromosomes: 1) one big cluster of individuals from all populations; 2) one cluster of individuals from subset of Mbakaou (blue dots in A.2); 3) *An. gambiae* as a separate cluster (red dots in Fig. A.2). When the latter two clusters of individuals were excluded, PCA showed no apparent geographic structures among any populations (Fig. A.3) and DAPC analysis had highest support for one group (K = 1) of all individuals. It is intriguing to find a very distinctive population within molecularly-identified *An. gambiae* individuals in Mbakaou. However, since our goal is to identify inversion associated selected regions rather than population specific effects, we excluded these individuals for the following analysis. Contrary to collinear regions, when individuals are clustered by SNPs from 2La (2L: 20524058-42165532) or 2Rb (2R: 19023925-26758676) regions, three clusters of individuals can be seen from first PCs (explaining 20.3% and 11.9% of the total variance respectively, Fig. A.4a, d). DAPC results supported K=3 as the most likely number of genetic clusters (Fig. A.4b, e). When compared to molecular karyotyping results, the three clusters match nicely to inverted homokaryotypes (I/I hereafter), heterokaryotypes (I/S) and standard homokaryotypes (S/S) (Fig. A.4c, f).

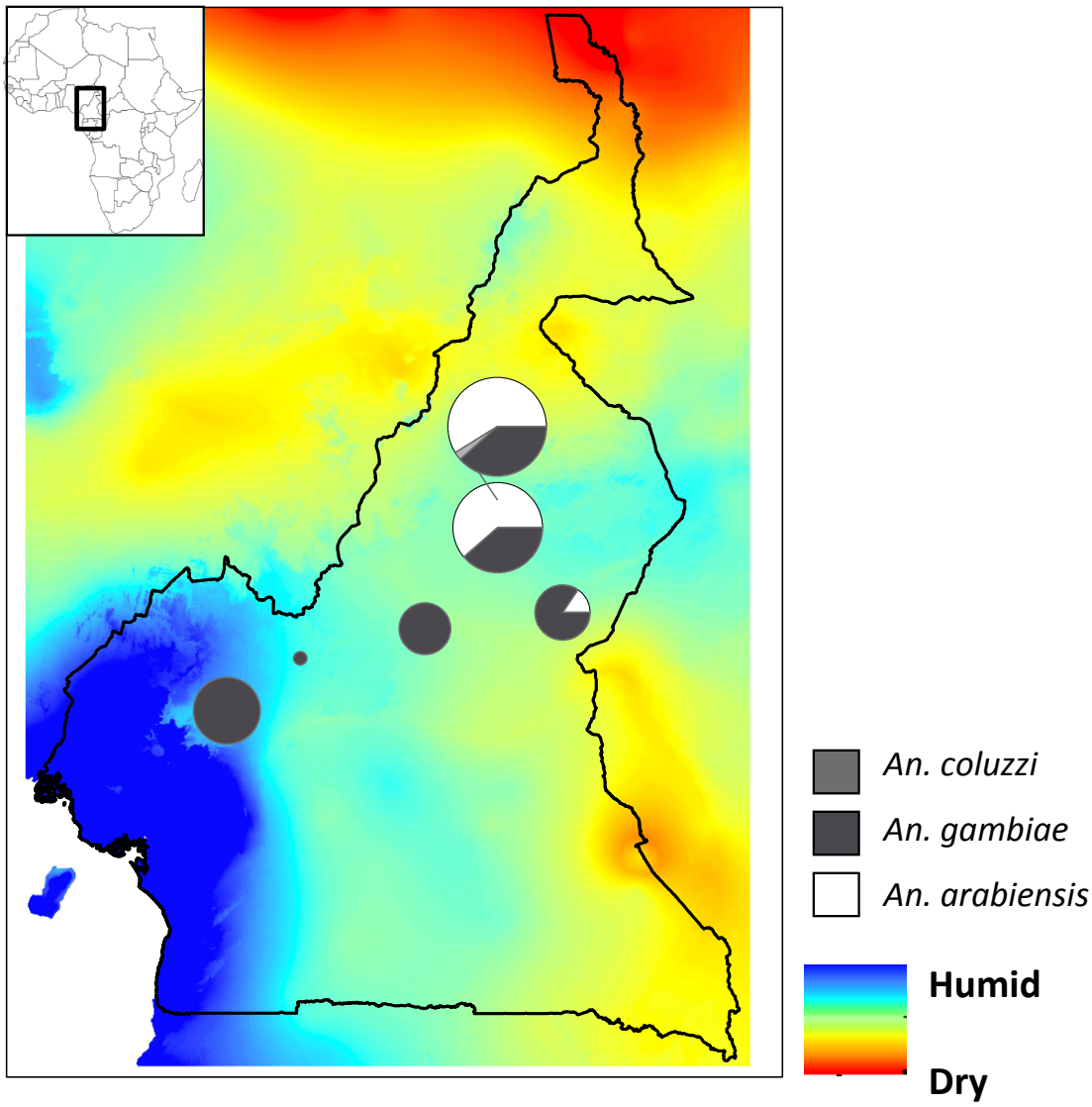

Figure A.1: Sampling locations and species composition of *Anopheles gambiae* species complex. The area of each pie chart correspond to the sample size. Map color from blue to red stands for humid to dry areas.

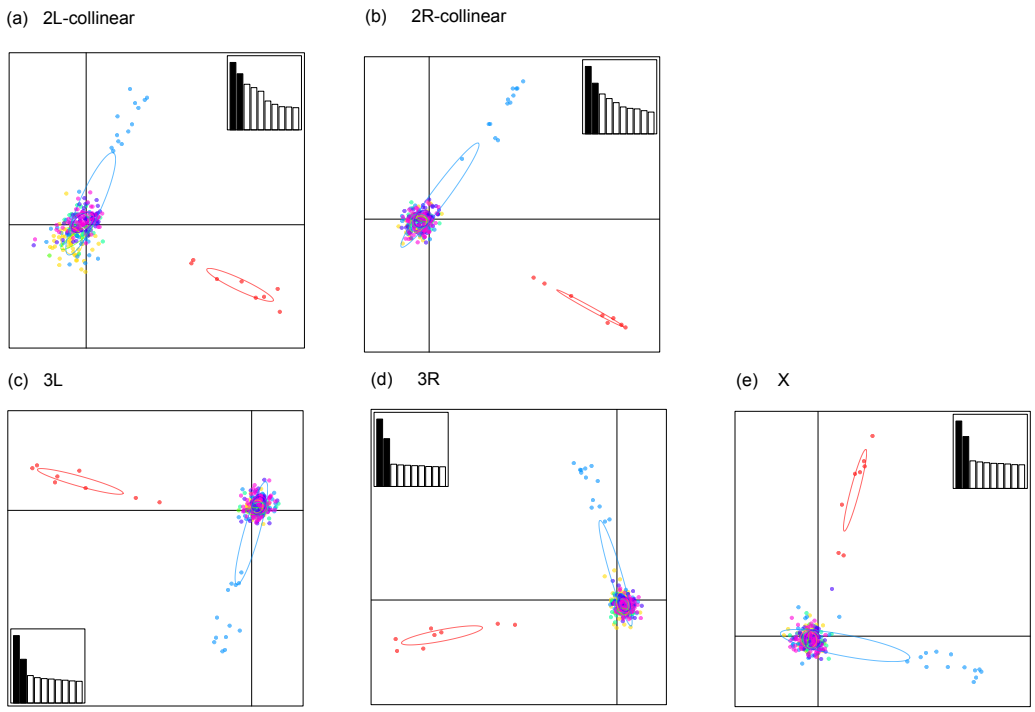

Figure A.2: Principle component analyses using SNPs in different collinear genomic regions. Color of the dots represent different populations. Red dots are individuals of *An. arabiensis*.

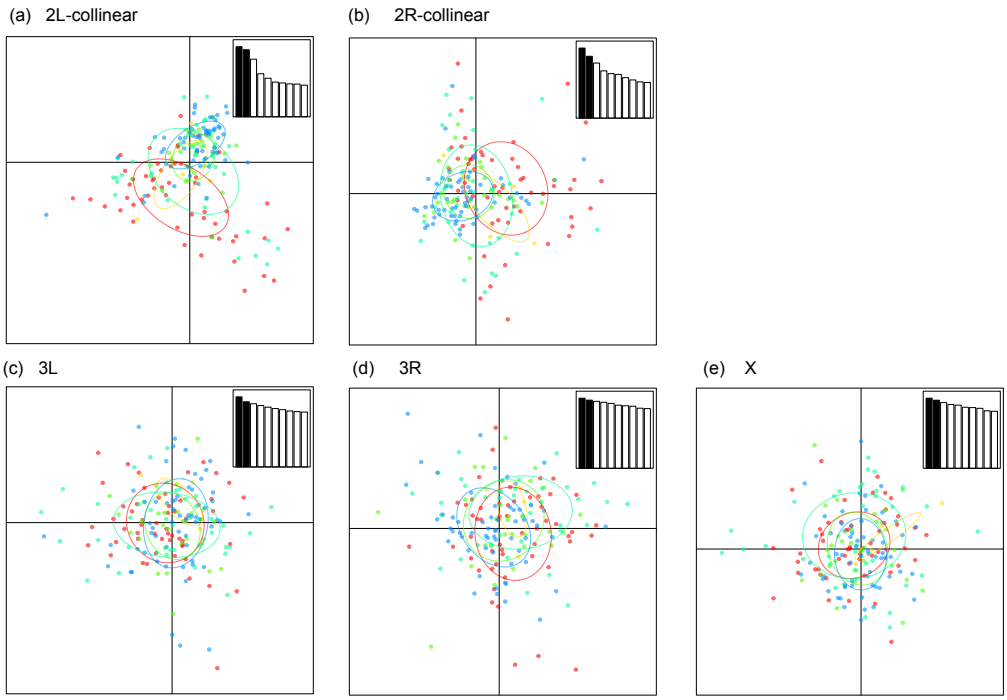

Figure A.3: Principle component analyses of *An. gambiae* using SNPs in different collinear genomic regions.

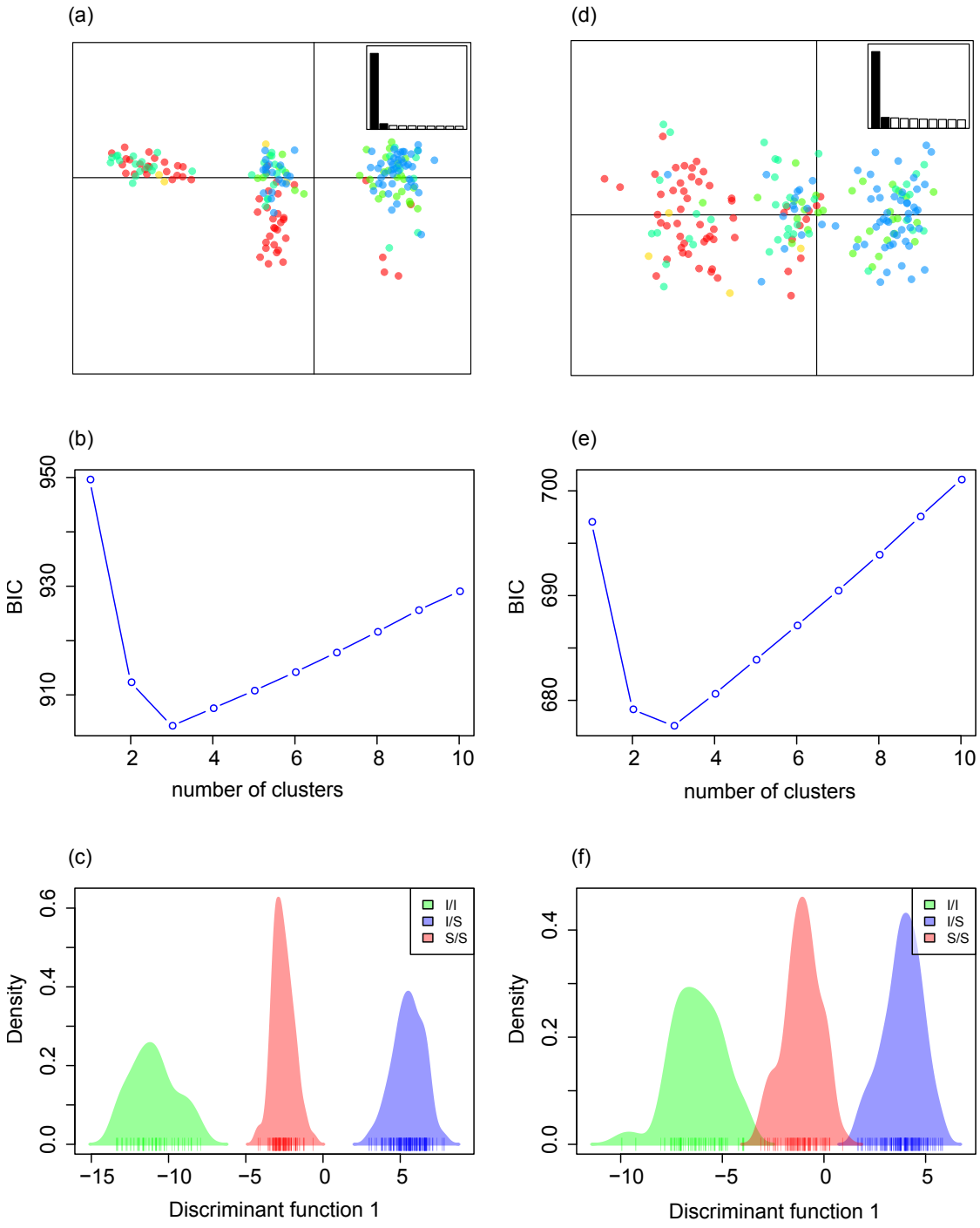

Figure A.4: Principle component analyses of *An. gambiae* using SNPs from 2La and 2Rb. Left and Right panels are the result for 2La and 2Rb, respectively. Top panel is the result for PCA clustering of individuals from different populations. Middle panel finds the best number of clusters based on BIC scores. The bottom panel shows how divergent each cluster is from each other on the discriminant function space.

APPENDIX B

Supplementary material of Chapter 3

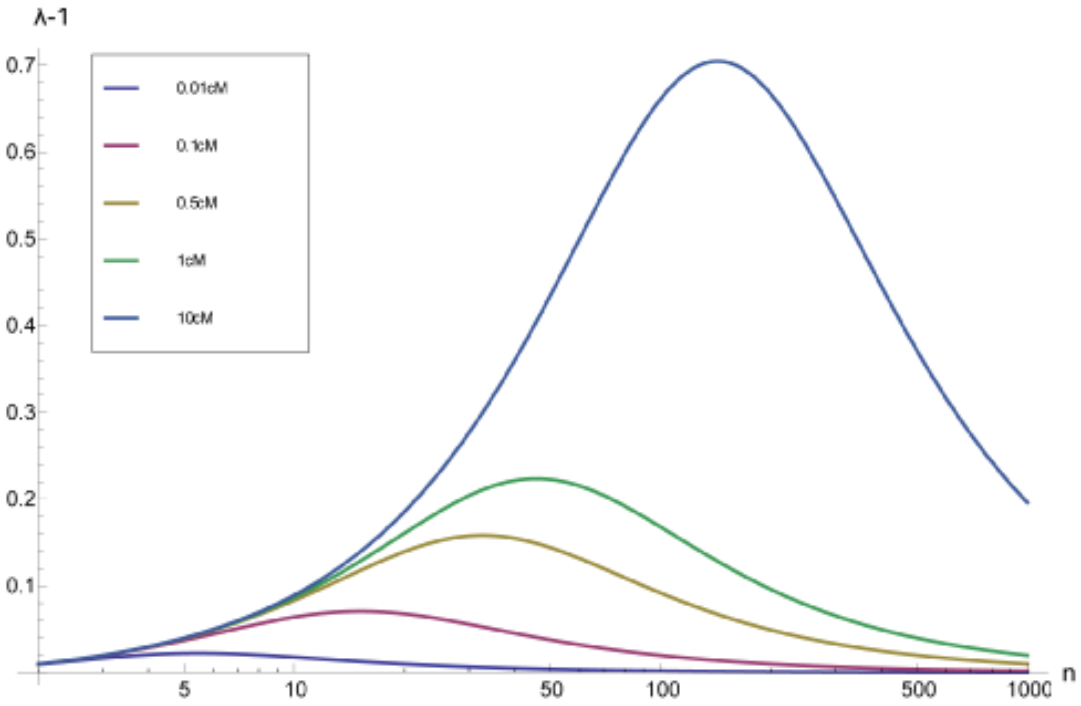

Figure B.1: Relationship between number of adaptive alleles, n, and selective advantage of inversion, $\lambda - 1$ based on Eq. 3 in Kirkpatrick and Barton (2006). Each line represents a set genetic length (from 0.01 cM to 10 cM). Alleles are assumed to be evenly spaced between each other, so that the recombination rate between each pair of loci, *r*, can be estimated using Kosambi's map function.

Table B.1: Summary of the probability of adaptation and the relative contribution of standing inversion variation to rapid adaptive divergence for different parameter settings (migration rate, *m*; beneficial selection, *s*; recombination rate, *r*; number of adapted loci, *n*) under two demographic scenarios: a constant population (N = Ne = 500) and cyclic population ($N(t) = 2525+2475\sin(2\pi(t+6.5)/10)$), which were chosen based on empirical studies of mosquito populations (Manoukis et al. 2008). New input of mutations lasts for 500 generations (G = 1N), with two levels of mutation rate, $10{-}6$ and 10^{-5}. 5,000 realizations in each scenario were run to observe the impact of different combination of parameters on the proportion of contribution from standing variation to the success of establishment of inversions.

	2Nu=0.001											
	s = 0.1						s=1					
	constant			cyclic			constant			cyclic		
	PNI	PADP	RSIV	PNI	PADP	RSIV	PNI	PADP	RSIV	PNI	PADP	RSIV
m = 0.002												
r = 0.01												
n = 2	0.0004	0.0026	0.8477	0.0000	0.0032	1.0000	0.0000	0.0021	1.0000	0.0000	0.0028	1.0000
n = 4	0.0010	0.0033	0.7023	0.0007	0.0039	0.8187	0.0000	0.0020	1.0000	0.0007	0.0036	0.8029
r = 0.1												
n = 2	0.0016	0.0049	0.6814	0.0008	0.0051	0.8450	0.0006	0.0030	0.8003	0.0010	0.0041	0.7622
n = 4	0.0008	0.0052	0.8471	0.0014	0.0069	0.7946	0.0008	0.0030	0.7373	0.0002	0.0035	0.9328
m = 0.02												
r = 0.01												
n = 2	0.0016	0.0049	0.6759	0.0014	0.0061	0.7738	0.0002	0.0024	0.9195	0.0008	0.0037	0.7898
n = 4	0.0024	0.0067	0.6480	0.0012	0.0069	0.8290	0.0016	0.0041	0.6189	0.0005	0.0037	0.8722
r = 0.1												
n = 2	0.0050	0.0113	0.5627	0.0044	0.0126	0.6571	0.0018	0.0061	0.7094	0.0026	0.0079	0.6742
n = 4	0.0080	0.0157	0.4987	0.0071	0.0172	0.5917	0.0032	0.0077	0.5882	0.0031	0.0088	0.6542
m = 0.2												
r = 0.01												
n = 2	0.0000	0.0018	1.0000	0.0000	0.0021	1.0000	0.0008	0.0044	0.8216	0.0012	0.0059	0.7982
n = 4	0.0000	0.0026	1.0000	0.0000	0.0031	1.0000	0.0022	0.0064	0.6597	0.0014	0.0068	0.7916
r = 0.1												
n = 2	0.0000	0.0019	1.0000	0.0000	0.0020	1.0000	0.0132	0.0209	0.3773	0.0086	0.0188	0.5493
n = 4	0.0000	0.0045	1.0000	0.0000	0.0049	1.0000	0.0200	0.0281	0.2995	0.0148	0.0255	0.4307

Table B.2: Table B.1 continued

	2Nu=0.01											
	s = 0.1						s=1					
	constant			cyclic			constant			cyclic		
	PNI	PADP	RSIV	PNI	PADP	RSIV	PNI	PADP	RSIV	PNI	PADP	RSIV
m = 0.002												
r = 0.01												
n = 2	0.0058	0.0270	0.8152	0.0052	0.0335	0.8667	0.0052	0.0253	0.8237	0.0054	0.0326	0.8579
n = 4	0.0062	0.0280	0.8096	0.0056	0.0343	0.8599	0.0068	0.0258	0.7732	0.0026	0.0296	0.9247
r = 0.1												
n = 2	0.0130	0.0441	0.7468	0.0114	0.0525	0.8136	0.0062	0.0275	0.8061	0.0052	0.0352	0.8732
n = 4	0.0250	0.0647	0.6682	0.0154	0.0665	0.8010	0.0062	0.0279	0.8094	0.0066	0.0368	0.8457
m = 0.02												
r = 0.01												
n = 2	0.0156	0.0481	0.7215	0.0126	0.0540	0.7997	0.0076	0.0290	0.7750	0.0050	0.0353	0.8785
n = 4	0.0268	0.0664	0.6535	0.0212	0.0739	0.7536	0.0066	0.0268	0.7884	0.0040	0.0333	0.8968
r = 0.1												
n = 2	0.0600	0.1141	0.5483	0.0424	0.1142	0.6811	0.0222	0.0622	0.6934	0.0176	0.0688	0.7802
n = 4	0.0974	0.1620	0.4836	0.0738	0.1615	0.6075	0.0224	0.0639	0.6988	0.0192	0.0716	0.7696
m = 0.2												
r = 0.01												
n = 2	0.0000	0.0159	1.0000	0.0000	0.0197	1.0000	0.0196	0.0531	0.6828	0.0124	0.0564	0.8112
n = 4	0.0000	0.0245	1.0000	0.0000	0.0294	1.0000	0.0236	0.0585	0.6532	0.0138	0.0617	0.8079
r = 0.1												
n = 2	0.0000	0.0172	1.0000	0.0000	0.0190	1.0000	0.1220	0.1860	0.4366	0.0862	0.1724	0.5706
n = 4	0.0000	0.0397	1.0000	0.0000	0.0448	1.0000	0.1654	0.2321	0.3878	0.1234	0.2133	0.5032

APPENDIX C

Supplementary material of Chapter 4

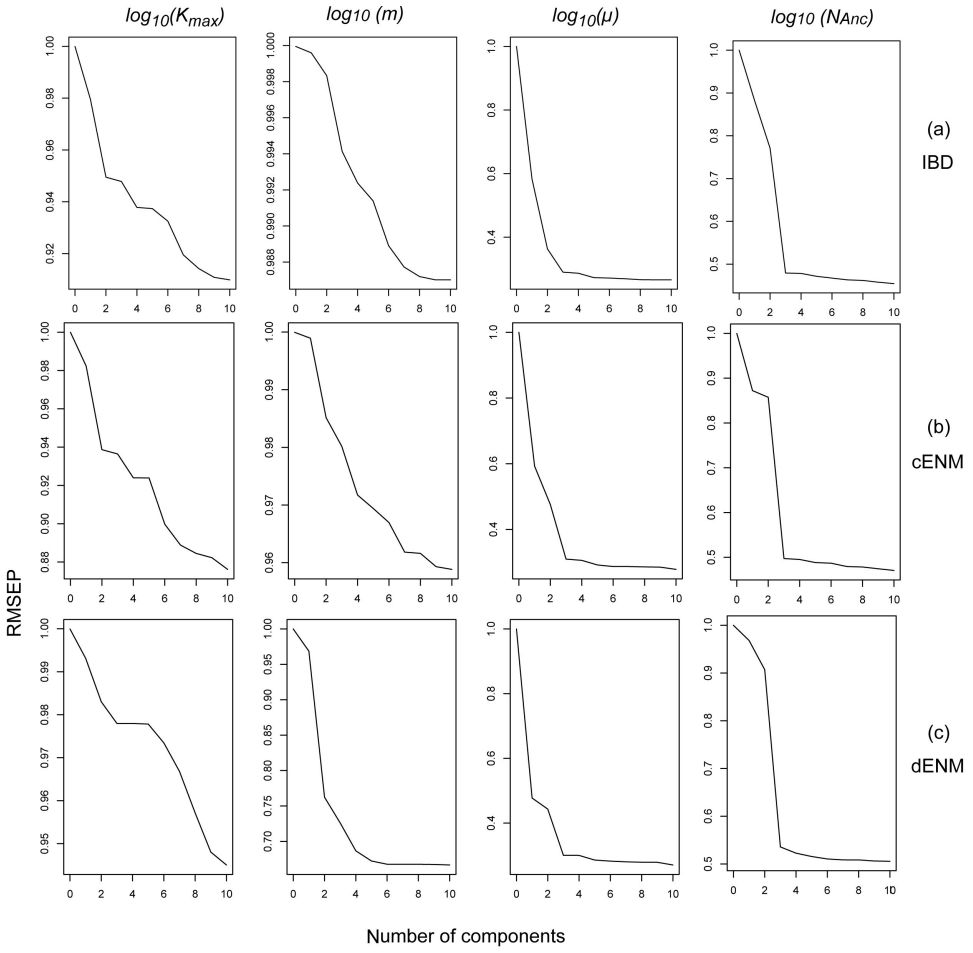

Figure C.1: Root Mean Square Error (RMSE) of parameter estimation against number of PLSs included under four demographic models: a) IBD, b) cENM, and c) dENM.

Table C.1: Geographic locations of sampled individuals and their assigned population (see Fig. C.1 for distributional details).

ID#	LAT	LONG	POP	ID#	LAT	LONG	POP	ID#	LAT	LONG	POP
LL40	-21.7833	114.1833	CB	LL33	-25.8388	113.6064	P	LL74	-26.4666	113.4667	SB
LL90	-21.8833	114.0167	CB	LL98	-25.8541	113.8625	P	LL75	-26.4666	113.4667	SB
LL118	-22.4025	113.8433	CB	LL148	-25.85417	113.8625	P	LL107	-26.7	113.6667	SB
LL19	-22.6833	114.05	CB	LL37	-25.8752	113.5503	P	LL73	-27.0116	114.4	ZU
LL21	-22.6833	114.05	CB	LL149	-25.97538	113.57049	P	LL28	-27.2591	114.0675	ZU
LL69	-22.6833	113.9833	CB	LL71	-25.9833	113.6	P	LL29	-27.2591	114.0675	ZU
LL151	-23.1466	113.7764	CB	LL143	-26.25	113.8	P	LL30	-27.2591	114.0675	ZU
LL150	-23.1477	113.7761	CB	LL144	-26.25	113.8	P	LL31	-27.2591	114.0675	ZU
LL15	-23.8833	113.4833	C	LL145	-26.26667	113.78333	P	LL46	-27.2591	114.0675	ZU
LL135	-24.1383	113.4458	C	LL146	-26.26667	113.78333	P	LL60	-27.7	114.1667	ZU
LL34	-24.1383	113.4458	C	LL12	-26.4333	114.0667	P	LL61	-27.7	114.1667	ZU
LL49	-24.1383	113.4458	C	LL39	-26.4333	114.0667	P	LL08	-27.85	114.1667	ZU
LL128	-24.1833	113.45	C	LL45	-26.4894	114.0556	P	LL41	-27.8977	114.1669	ZU
LL134	-24.1833	113.45	C	LL47	-26.4894	114.0556	P	LL42	-27.8977	114.1669	ZU
LL137	-24.1833	113.45	C	LL52	-26.5225	114.0025	P	LL66	-28.3	113.5833	ZU
LL133	-24.193	113.4553	C	LL53	-26.5225	114.0025	P	LL67	-28.3	113.5833	ZU
LL38	-24.193	113.4553	C	LL44	-26.5463	113.9633	P	LL22	-28.6166	114.6	SM
LL48	-24.223	113.4914	C	LL54	-26.5463	113.9633	P	LL23	-28.6166	114.6	SM
LL50	-24.223	113.4914	C	LL72	-26.5666	114	P	LL10	-28.7333	115	SM
LL142	-25.1155	113.7292	C	LL70	-25.9333	113.1667	SB	LL03	-28.7666	114.6167	SM
LL64	-25.1155	113.7292	C	LL126	-26	113.2	SB	LL04	-28.7666	114.6167	SM
LL65	-25.1155	113.7292	C	LL05	-26.0333	113.2	SB	LL14	-28.7666	114.6167	SM
LL35	-25.1258	113.8228	C	LL06	-26.0333	113.2	SB	LL82	-28.7666	114.6167	SM
LL63	-25.1258	113.8228	C	LL07	-26.0333	113.2	SB	LL83	-28.7666	114.6167	SM
LL141	-25.1316	113.7681	C	LL87	-26.3333	113.3833	SB	LL84	-28.7666	114.6167	SM
LL25	-25.1316	113.7681	C	LL76	-26.3833	113.3167	SB	LL85	-28.7666	114.6167	SM
LL36	-25.1316	113.7681	C	LL77	-26.3833	113.3167	SB	LL86	-28.7666	114.6167	SM
LL62	-25.1341	113.8056	C	LL78	-26.3833	113.3167	SB	LL13	-28.8666	114.6333	SM
LL43	-25.8205	113.5392	P	LL79	-26.3833	113.3167	SB	LL100	-29.2866	114.9244	SM
LL32	-25.8388	113.6064	P	LL89	-26.3833	113.3167	SB				

Table C.2: List of nuclear loci sequenced in this study. Primers used for amplification and the PCR conditions (i.e., annealing temperature, TA (C), and magnesium chloride concentrations, $MgCl_2$ (mM)) are provided along with the type of marker; anonymous nuclear loci developed from this study are listed as anonymous and PCRs using touchdown amplification are marked as TD (see Edwards 2007).

Locus	Type	Forward Primer (5' - 3')	Reverse Primer (5' - 3')	Reference	TA (C)	MgCl2 (mM)
ATP	intron	CGTGAGGGHAAYGATTTHTACCATGAGATG	TCTGTCCATAAACTAGCG	(Jarman et al. 2002)	59	2
BACH1	exon	GATTTGAHCCYTTRCTTCAGTTTGC	ACCTCACATTCYTGTTCYCTRGC	(Townsend et al. 2008)	(see Townsend et al. 2008)	
GAPD	intron	ACCTTTAATGCGGGTGCTGGCATTGC	CATCAAGTCCACAACACGGTTGCTGTA	(Dolman and Phillips 2004)	63	2
L17	anonymous	TGTCCCCGTCRGTCAATAAA	AAGGAGAGCCAAGACCTGAAC	This study	65	2
L25	anonymous	GCTCTGGAATTAGCATTATCWCTTG	GGTGGGAAACATTTCTTTGTTG	This study	65	2
L37	anonymous	GTGTGCSAAGAAGAAAGGAGGM	TTTGCGAGTGCGACAGWTT	This study	65	2
L74	anonymous	GTGGATGGGGGTTATGTTTG	TGGTCAGCATTTGCMCTCAC	This study	65	2
L101	anonymous	GGCACACACAGCACATTTTT	AGAAGAAGAARAAAYCCCAAGGT	This study	65	2
L110	anonymous	TTGTGTGGGGATGCTGACT	KGCGGAGAGGAAAAATGG	This study	65	2
L115	anonymous	GGGGAACCTGTCTATCCTACAA	AGCGGAACCACTGCAAAA	This study	65	2
L145	anonymous	YRAGGACCARCAAATCATCAAC	GCCAGCAAGGGCTAYMAA	This study	65	2
L169	anonymous	CAAAGAAAAAGACAAGGGGAGA	AGGTGACTGAAAGGCTGAGAAG	This study	65	2
L218	anonymous	GCAAACCCAGAATGTCCTAATC	TGCAAGCAAGGGTACAAGG	This study	65	2
L269	anonymous	CACCCAGCCCAAGAAATG	TTCATCAGACACAACAGAAGTGG	This study	TD	2.5
L272	anonymous	GAAAGACCCCCAAGAAAGAMAG	GACACACCAAGAGAAGGCATAAA	This study	65	2
L308	anonymous	TTGTGGTGTCCAGTGMGGAA	TGGGTGAAGGGAGGAATG	This study	63	2.5
L323	anonymous	CAGCAYAGAGGACACAAAGGT	AAYGWACRGAGGGAACTAACAAG	This study	65	2
L426	anonymous	TCAACTGCCTTCCAAAATAACC	TCTTCCATACAATCCTACCCATCT	This study	65	2
L907	anonymous	CAGATGATAGCCAGAAATAAGCAC	TCACAGAAATCCAAACCTACCT	This study	59	2.5
L926	anonymous	ACCCCTTTCCTTCTCACCTT	CCACCTTTGCTTSCTTCCTC	This study	63	0.8
L1088	anonymous	CAAAAGGTTYGTGAGGCAAGA	GCCAGAGGATTGGAGGATAG	This study	65	2
NTF	exon	ATGTCCATCTTGTTTTATGTGATATTT	ACRAGTTTRTTGTTYTCTGAAGTC	(Townsend et al. 2008)	(see Townsend et al. 2008)	
PRLR	exon	GACARYGARGACCAGCAACTRATGCC	GACYTTGTGRACTTCYACRTAATCCAT	(Townsend et al. 2008)	(see Townsend et al. 2008)	
PTPN	exon	AGTTGCCTTGTWGAAGGRGATGC	CTRGCAATKGACATYGGYAATAC	(Townsend et al. 2008)	(see Townsend et al. 2008)	

Table C.3: Settings for NGen sequence assembler (DNASTAR) used for the 454 dataset in the discovery of polymorphic loci.

Categories	Parameters	Settings
Repeat	Max Mer Gap	10
Repeat	Match Size	17
Repeat	Min End Flag Len	25
Repeat	Min Flag Length	50
Repeat	Min Mer Match	2
Quality	End Region	5
Quality	Maximum uncalled bases	2
Quality	Minimum Average High Quality	22
Quality	Minimum Average Low Quality	20
Quality	Minimum End Basepair Quality	15
Quality	NTrimWinLength	7
Quality	Window Length	30
Alignment	Fixed Coverage	20
Alignment	Default Quality	15
Alignment	Gap Penalty	75
Alignment	Genome Length	10000000
Alignment	HaploidSNP	FALSE
Alignment	HaploidThreshold	0
Alignment	LowCoverageThreshold	0
Alignment	Match Score	10
Alignment	Match Window Length	50
Alignment	Match Repeat Percent	150
Alignment	Max Gap	15
Alignment	Max Usable Count	25
Alignment	Match Size	19
Alignment	Match Spacing	20
Alignment	Minimum Match Percent	85
Alignment	Mismatch Penalty	15
Alignment	Skip Realign	FALSE
Alignment	SNP Match Percentage	90
Alignment	SNP Passes	2
Alignment	Split False Joins	FALSE
Alignment	Split Template Contigs	FALSE
Alignment	Template Default Quality	500
Alignment	Use Repeat Handling	TRUE

Table C.4: Length of each locus and sampling per populations for each locus.

		Populations							
Gene	Length	CB	C	P	SB	ZU	SM	Total	Genbank Accession No.
ATP	490	16	40	40	28	24	26	174	KC545970 - KC546143
BACH1	1218	16	40	32	24	18	22	152	KC546144 - KC546295
GAPD	630	12	30	22	19	12	20	115	KC546296 - KC546411
L17	178	16	40	40	28	24	26	174	KC546412 - KC546585
L25	276	16	38	40	28	26	26	174	KC546586 - KC546759
L37	234	14	28	16	16	24	22	120	KC546760 - KC546879
L74	265	14	34	34	24	22	26	154	KC546880 - KC547035
L101	367	6	12	12	14	8	22	74	KC547036 - KC547109
L110	378	14	24	18	14	16	16	102	KC547110 - KC547211
L115	543	14	38	40	26	22	24	164	KC547212 - KC547375
L145	259	14	26	20	24	14	14	112	KC547376 - KC547487
L169	587	14	36	36	18	14	18	136	KC547488 - KC547623
L218	180	6	20	18	6	18	26	94	KC547624 - KC547717
L269	302	14	36	26	20	22	20	138	KC547718 - KC547855
L272	212	12	38	42	28	20	16	156	KC547856 - KC548011
L308	369	14	34	28	22	22	22	142	KC548012 - KC548153
L323	341	16	36	38	26	18	26	160	KC548154 - KC548313
L426	154	16	34	28	26	24	26	154	KC548314 - KC548467
L907	222	16	40	40	24	18	24	162	KC548468 - KC548629
L926	169	16	38	36	28	26	26	170	KC548630 - KC548799
L1088	179	16	40	42	28	26	26	178	KC548800 - KC548977
NTF3	656	14	30	32	20	20	20	136	KC548978 - KC549113
PRLR	557	14	38	34	28	22	26	162	KC549114 - KC549275
PTPN12	865	16	40	34	26	22	26	164	KC549276 - KC549439

Table C.5: Soil properties used in the construction of soil layers for the PCA analyses (for detailed description see McKenzie et al. 2000.

Variable	Type	Description
Aclay50	integer	median A horizon clay %
Bclay50	integer	median B horizon clay %
Athick50	numeric	median A horizon thickness (m)
Bthick50	numeric	median B horizon thickness (m)
Solumthick50	numeric	median solum thickness (m)
Astruct50	integer	median A horizon grade of pedality
Bstruct50	integer	median B horizon grade of pedality
A Ks	integer	A horizon log10 (saturated hydraulic conductivity mm/hr) - 50th percentile
AKserror	integer	log10(A Ks) error
B Ks	integer	B horizon log10 (saturated hydraulic conductivity mm/hr) - 50th percentile
BKserror	integer	log10(B Ks) error
Calcrete	Boolean	absence (0) or presence (1) of calcrete in or below the profile
Nutrients	integer	nutrient status low (1), moderate (2) and high (3)

Table C.6: Molecular indices calculated per locus and presented for each population separately (see Fig. 4.1 for distributional information), as well as across all populations, including heterozygozity (H) and the standard deviation (H_{sd}), the number of segregating sites (S), the number of haplotypes (K) and nucleotide diversity (π).

				K							π						
Gene	H	H_{sd}	S	CB	C	P	SB	ZU	SM	species	CB	C	P	SB	ZU	SM	species
ATP	0.955	0.065	49	9	15	21	17	24	13	55	0.009	0.010	0.015	0.011	0.009	0.005	0.100
BACH1	0.762	0.017	40	7	10	7	8	4	5	35	0.001	0.002	0.001	0.001	0.001	0.001	0.033
GAPD	0.914	0.057	60	6	9	10	9	6	6	40	0.018	0.008	0.017	0.005	0.008	0.002	0.095
L101	0.956	0.101	28	4	7	8	6	6	5	29	0.012	0.020	0.017	0.009	0.012	0.010	0.076
L1088	0.500	0.042	3	3	3	3	5	5	8	3	0.007	0.005	0.004	0.004	0.002	0.007	0.017
L110	0.961	0.102	32	5	8	11	4	6	6	34	0.009	0.016	0.016	0.004	0.017	0.010	0.085
L115	0.730	0.033	5	5	6	5	4	6	3	7	0.003	0.003	0.003	0.002	0.003	0.002	0.009
L145	0.944	0.110	32	4	11	11	8	2	5	34	0.004	0.026	0.025	0.020	0.010	0.018	0.124
L169	0.975	0.043	29	6	9	14	10	8	10	49	0.003	0.005	0.007	0.013	0.010	0.006	0.049
L17	0.404	0.017	5	6	4	5	4	2	1	7	0.003	0.004	0.003	0.002	0.001	0.000	0.028
L218	0.625	0.043	20	6	4	3	2	5	3	13	0.055	0.005	0.004	0.003	0.006	0.002	0.111
L25	0.744	0.059	24	8	10	15	2	10	5	31	0.060	0.012	0.027	0.000	0.013	0.004	0.087
L269	0.944	0.080	20	1	14	14	13	8	4	38	0.000	0.018	0.014	0.009	0.014	0.004	0.066
L272	0.652	0.040	6	6	3	5	3	6	4	9	0.007	0.005	0.004	0.004	0.004	0.003	0.028
L308	0.954	0.050	22	5	6	11	8	6	9	39	0.004	0.003	0.003	0.006	0.027	0.029	0.060
L323	0.627	0.022	12	3	4	8	9	6	2	21	0.001	0.001	0.004	0.005	0.002	0.000	0.035
L37	0.938	0.087	22	8	15	4	5	3	6	35	0.041	0.021	0.004	0.020	0.004	0.011	0.094
L426	0.615	0.040	7	3	4	4	4	5	2	8	0.003	0.004	0.002	0.005	0.005	0.001	0.045
L74	0.767	0.030	10	3	7	5	3	2	5	17	0.003	0.009	0.010	0.004	0.001	0.005	0.038
L907	0.749	0.033	11	12	12	23	16	16	17	16	0.003	0.014	0.002	0.006	0.008	0.004	0.050
L926	0.741	0.111	6	12	7	5	5	9	3	12	0.082	0.055	0.024	0.052	0.032	0.001	0.036
NTF3	0.816	0.024	26	3	7	10	10	6	2	22	0.002	0.003	0.003	0.004	0.002	0.001	0.040
PRLR	0.932	0.033	48	5	7	17	9	18	11	52	0.002	0.002	0.005	0.005	0.006	0.003	0.086
PTPN12	0.951	0.023	40	7	10	18	8	15	8	59	0.002	0.004	0.004	0.002	0.002	0.002	0.046
Total	1.000	0.055	520	28	25	24	42	40	16	175	0.003	0.006	0.004	0.005	0.005	0.003	0.061

Table C.7: List of summary statistics used in ABC analyses.

Summary Statistics	Description	Number
STOT	segregating sites in the species	1
S	segregating sites in each population	6
PrS	private segregating sites	6
p	pairwise genetic differences within each population	6
FST	pairwise FSTvalues among populations	15
total		34

Table C.8: Pairwise F_{st} of the six populations ordered from north to south (lower triangle) and the significance (upper triangle). Haplotype distances between individuals are calculated using Tajima and Nei's correction. Significance of each F_{st} is assessed with 1023 permutations. $+$, significant ($P < 0.05$); $-$, non-significant. Note that F_{st} between P and SB is the only non-significant one.

	CB	C	P	SB	ZU	SM
CB		$+$	$+$	$+$	$+$	$+$
C	0.298		$+$	$+$	$+$	$+$
P	0.247	0.095		$-$	$+$	$+$
SB	0.291	0.097	0.020		$+$	$+$
ZU	0.425	0.262	0.178	0.237		$+$
SM	0.550	0.392	0.355	0.421	0.219	

C.1 Supplemental methods:

Conditions used for simulation under each of the models tested with ABC:

(i) IBD model: simulations are started from the current distribution. Migration and population growth are allowed each generation, but there is no suitability differences among demes (i.e., the carrying capacity remains the same across grid cells). In other words, the population densities and the number of emigrants from a specific deme that migrated into the surrounding cells were uniform across surrounding cells (i.e., they did not differ according to cell-specific suitabilities).

(ii) cENM model: expansion of populations start at current distributions as with the IBD model. However, the suitability of each cell follows the current ENM. Consequently, the local carrying capacity of a cell is proportional to its suitability score, as is the number of emigrants.

(iii) dENM model: populations start expansion at the LGM refugia areas (i.e., areas with highest suitability) predicted from ENM based on paleoclimatic data from the LGM 21thousand years ago. Maps of population carrying capacities change overtime (see Brown and Knowles, 2012), with the demographic simulations informed during the first and last 1/3 of the generations from the ENM from the past and present, respectively, and a composite map (i.e., the average habitat suitabilities from the past and present ENMs) informing the demographic simulations for the intervening generations.

Given that the species lives in coastal sand plain or dunes, it is very unlikely for the species to migrate to majority of the inland area. Thus, we coded areas with predicted suitability less than 0.01 to be inhabitable in all models. The number of generations for expansion and migrations in the models were as follows: 7000 years before the present for the IBD, cENM, and the last third of the dENM. The dENM model was run for a total of 21,000 years before the present to account for the shifting distributions associated with the last glacial maximum (as detailed above). We used one generation per 10 years as a generation time to reduce the length of the time forward simulations (i.e., the demographic simu-

lations conducted in SPLATCHE2; Currat and Excoffier, 2004). The maximum coalescent time (a parameter in the program Splatch2 that specifies when gene lineage coalescence across patches must occur) was set to be much greater than 4N for all models to avoid the forced coalescence to a single common ancestor.

We note that any biological interpretation of absolute parameter values of the mutation rate and migration rate would therefore need to be scaled according to realistic generational times for *L. lineopuntulata*, although this has not been studied empirically.

BIBLIOGRAPHY

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 64:233–247.

Aeschbacher, S. and R. Brger, 2014. The effect of linkage on establishment and survival of locally beneficial mutations. Genetics 197:317–336.

Antao, T., A. Lopes, R. J. Lopes, A. Beja-Pereira, and G. Luikart, 2008. Lositan: A workbench to detect molecular adaptation based on a f(st)-outlier method. BMC Bioinformatics 9.

Arajo, M. B. and A. Guisan, 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography 33:1677–1688.

Auton, A. and G. McVean, 2007. Recombination rate estimation in the presence of hotspots. Genome research 17:1219–1227.

Avise, J. C., J. Arnold, R. M. Ball, E. Bermingham, T. Lamb, J. E. Neigel, C. A. Reeb, and N. C. Saunders, 1987. Intraspecific phylogeography: the mitochondrial dna bridge between population genetics and systematics. Annual Review of Ecology and Systematics 18:489–522.

Ayala, F. J. and M. Coluzzi, 2005. Chromosome speciation: Humans, drosophila, and mosquitoes. Proceedings of the National Academy of Sciences of the United States of America 102:6535–6542.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, and E. A. Johnson, 2008. Rapid snp discovery and genetic mapping using sequenced rad markers. PloS one 3:e3376.

Balanya, J., L. Serra, G. W. Gilchrist, R. B. Huey, M. Pascual, F. Mestres, and E. Sole, 2003. Evolutionary pace of chromosomal polymorphism in colonizing populations of drosophila subobscura: An evolutionary time series. Evolution 57:1837–1845.

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

Balkenhol, N. and E. L. Landguth, 2011. Simulation modelling in landscape genetics: on the need to go further. Molecular Ecology 20:667–670.

Barrett, R. D. H. and D. Schluter, 2008. Adaptation from standing genetic variation. Trends in Ecology & Evolution 23:38–44.

Barton, N., 1995. A general model for the evolution of recombination. Genetical research 65:123–144.

Bayoh, M. N., C. J. Thomas, and S. W. Lindsay, 2001. Mapping distributions of chromosomal forms of anopheles gambiae in west africa using climate data. Medical and Veterinary Entomology 15:267–274.

Beaumont, M. A., 2005. Adaptation and speciation: what can F_{ST} tell us? Trends in Ecology & Evolution 20:435–440.

Beaumont, M. A. and D. J. Balding, 2004. Identifying adaptive genetic divergence among populations from genome scans. Molecular Ecology 13:969–980.

Beaumont, M. A., J.-M. Cornuet, J. Marin, and C. P. Robert, 2009. Adaptive approximate bayesian computation. Biometrika 86:983–990.

Beaumont, M. A. and R. A. Nichols, 1996. Evaluating loci for use in the genetic analysis of population structure. Proceedings of the Royal Society of London. Series B: Biological Sciences 263:1619–1626.

Beaumont, M. A., W. Y. Zhang, and D. J. Balding, 2002. Approximate bayesian computation in population genetics. Genetics 162:2025–2035.

Beerli, P. and J. Felsenstein, 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proceedings of the National Academy of Sciences of the United States of America 98:4563–4568.

Bertorelle, G., A. Benazzo, and S. Mona, 2010. Abc as a flexible framework to estimate demography over space and time: some cons, many pros. Molecular Ecology 19:2609–2625.

Bertozzi, T., K. L. Sanders, M. J. Sistrom, and M. G. Gardner, 2012. Anonymous nuclear loci in non-model organisms: making the most of high-throughput genome surveys. Bioinformatics 28:1807–1810.

Besansky, N. J., J. Krzywinski, T. Lehmann, F. Simard, M. Kern, O. Mukabayire, D. Fontenille, Y. Toure, and N. F. Sagnon, 2003. Semipermeable species boundaries between anopheles gambiae and anopheles arabiensis: Evidence from multilocus dna sequence variation. Proceedings of the National Academy of Sciences of the United States of America 100:10818–10823.

Besansky, N. J., T. Lehmann, G. T. Fahey, D. Fontenille, L. E. O. Braack, W. A. Hawley, and F. H. Collins, 1997. Patterns of mitochondrial variation within and between african malaria vectors, anopheles gambiae and an-arabiensis, suggest extensive gene flow. Genetics 147:1817–1828.

Besansky, N. J., J. R. Powell, A. Caccone, D. M. Hamm, J. A. Scott, and F. H. Collins, 1994. Molecular phylogeny of the anopheles gambiae complex suggests genetic introgression between principal malaria vectors. Proceedings of the National Academy of Sciences of the United States of America 91:6885–6888.

Bhutkar, A., W. M. Gelbart, and T. F. Smith, 2007. Inferring genome-scale rearrangement phylogeny and ancestral gene order: a drosophila case study. Genome Biology 8.

Bonhomme, M., C. Chevalet, B. Servin, S. Boitard, J. Abdallah, S. Blott, and M. San-Cristobal, 2010. Detecting selection in population trees: The lewontin and krakauer test extended. Genetics 186:241–U406.

Boulesteix, A.-L. and K. Strimmer, 2007. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics 8:32–44.

Box, G. E. P. and D. R. Cox, 1964. An analysis of transformations. "Journal of the Royal Statistical Society. Series B, Statistical methodology" 26:211–252.

ter Braak, C. J. F. and P. F. M. Verdonschot, 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquatic Sciences 57:255–289.

Bradshaw, H. D., S. M. Wilbert, K. G. Otto, and D. W. Schemske, 1995. Genetic-mapping of floral traits associated with reproductive isolation in monkeyflowers (mimulus). Nature 376:762–765.

Brown, J. L. and L. L. Knowles, 2012. Spatially explicit models of dynamic histories: examination of the genetic consequences of pleistocene glaciation and recent climate change on the american pika. Molecular Ecology 21:3757–3775.

Bruggeman, D. J., T. Wiegand, and N. Fernandez, 2010. The relative effects of habitat loss and fragmentation on population genetic variation in the red-cockaded woodpecker (picoides borealis). Molecular Ecology 19:3679–3691.

Bush, A. B. G., 2007. Extratropical influences on the el nino-southern oscillation through the late quaternary. Journal of Climate 20:788–800.

Carlson, C. S., D. J. Thomas, M. A. Eberle, J. E. Swanson, R. J. Livingston, M. J. Rieder, and D. A. Nickerson, 2005. Genomic regions exhibiting positive selection identified from dense genotype data. Genome research 15:1553–1565.

Carnaval, A. C., M. J. Hickerson, C. F. B. Haddad, M. T. Rodrigues, and C. Moritz, 2009. Stability predicts genetic diversity in the brazilian atlantic forest hotspot. Science 323:785–789.

Carstens, B. and L. Knowles, 2006. Variable nuclear markers for melanoplus oregonensis identified from the screening of a genomic library. Molecular Ecology Notes 6:683–685.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko, 2013. Stacks: an analysis tool set for population genomics. Molecular ecology 22:3124–3140.

Catchen, J. M., A. Amores, P. Hohenlohe, W. Cresko, and J. H. Postlethwait, 2011. Stacks: Building and genotyping loci de novo from short-read sequences. G3: Genes, Genomes, Genetics 1:171–182.

Cheng, C., B. J. White, C. Kamdem, K. Mockaitis, C. Costantini, M. W. Hahn, and N. J. Besansky, 2012. Ecological genomics of anopheles gambiae along a latitudinal cline: A population-resequencing approach. Genetics 190:1417–1432.

Cincotta, R. P., J. Wisnewski, and R. Engelman, 2000. Human population in the biodiversity hotspots. Nature 404:990–992.

Coetzee, M., R. H. Hunt, R. Wilkerson, A. Della Torre, M. B. Coulibaly, and N. J. Besansky, 2013. Anopheles coluzzii and anopheles amharicus, new members of the anopheles gambiae complex. Zootaxa 3619:246–274.

Cogger, H., 2000. Reptiles and amphibians of Australia, 5th edn. Reed New Holland, Sydney.

Colosimo, P. F., K. E. Hosemann, S. Balabhadra, G. Villarreal, M. Dickson, J. Grimwood, J. Schmutz, R. M. Myers, D. Schluter, and D. M. Kingsley, 2005. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. Science 307:1928–1933.

Coluzzi, M., V. Petrarca, and M. A. d. Deco, 1985. Chromosomal inversion intergradation and incipient speciation in anopheles gambiae. Bolletino di zoologia 52:45 – 63.

Coluzzi, M., A. Sabatini, V. Petrarca, and M. A. Dideco, 1979. Chromosomal differentiation and adaptation to human environments in the anopheles gambiae complex. Transactions of the Royal Society of Tropical Medicine and Hygiene 73:483–497.

Coluzzi, M., A. Sabatini, A. della Torre, M. A. Di Deco, and V. Petrarca, 2002. A polytene chromosome analysis of the anopheles gambiae species complex. Science 298:1415–1418.

Cook, S. R., A. Gelman, and D. B. Rubin, 2006. Validation of software for bayesian models using posterior quantiles. "Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America" 15:675–692.

Costantini, C., D. Ayala, W. Guelbeogo, M. Pombi, C. Some, I. Bassole, K. Ose, J.-M. Fotsing, N. Sagnon, D. Fontenille, N. Besansky, and F. Simard, 2009. Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in anopheles gambiae. BMC Ecology 9:16.

Csillery, K., M. G. B. Blum, O. E. Gaggiotti, and O. Francois, 2010. Approximate bayesian computation (abc) in practice. Trends in Ecology & Evolution 25:410–418.

Currat, M. and L. Excoffier, 2004. Modern humans did not admix with neanderthals during their range expansion into europe. Plos Biology 2:2264–2274.

Cushman, S. A. and E. L. Landguth, 2010. Scale dependent inference in landscape genetics. Landscape Ecology 25:967–979.

Czeher, C., R. Labbo, G. Vieville, I. Arzika, H. Bogreau, C. Rogier, L. Diancourt, S. Brisse, F. Ariey, and J. B. Duchemin, 2010. Population genetic structure of anopheles gambiae and anopheles arabiensis in niger. Journal of Medical Entomology 47:355–366.

De Maio, N., C. Schlotterer, and C. Kosiol, 2013. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. Molecular biology and evolution 30:2249–2262.

Delcher, A. L., S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S. L. Salzberg, 1999. Alignment of whole genomes. Nucleic Acids Research 27:2369–2376.

Delcher, A. L., A. Phillippy, J. Carlton, and S. L. Salzberg, 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic acids research 30:2478–2483.

Dolman, G. and B. Phillips, 2004. Single copy nuclear dna markers characterized for comparative phylogeography in australian wet tropics rainforest skinks. Molecular Ecology Notes 4:185–187.

Donnelly, M. J., J. Pinto, R. Girod, N. J. Besansky, and T. Lehmann, 2004. Revisiting the role of introgression vs shared ancestral polymorphisms as key processes shaping genetic diversity in the recently separated sibling species of the anopheles gambiae complex. Heredity 92:61–68.

Donnelly, M. J., F. Simard, and T. Lehmann, 2002. Evolutionary studies of malaria vectors. Trends in Parasitology 18:75–80.

Edwards, D. L., 2007. Biogeography and speciation of a direct developing frog from the coastal arid zone of western australia. Molecular Phylogenetics and Evolution 45:494–505.

Edwards, D. L., J. S. Keogh, and L. L. Knowles, 2012. Effects of vicariant barriers, habitat stability, population isolation and environmental features on species divergence in the south-western australian coastal reptile community. Molecular Ecology 21:3809–3822.

Ekblom, R. and J. Galindo, 2010. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107:1–15.

Epperson, B. K. and T. Q. Li, 1996. Measurement of genetic structure within populations using moran's spatial autocorrelation statistics. Proceedings of the National Academy of Sciences of the United States of America 93:10528–10532.

Epperson, B. K., B. H. McRae, K. Scribner, S. A. Cushman, M. S. Rosenberg, M.-J. Fortin, P. M. A. James, M. Murphy, S. Manel, P. Legendre, and M. R. T. Dale, 2010. Utility of computer simulations in landscape genetics. Molecular Ecology 19:3549–3564.

Eriksson, A. and A. Manica, 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proceedings of the National Academy of Sciences of the United States of America 109:13956–13960.

Ewens, W. J., 2004. Mathematical population genetics. I. Theoretical Introduction. Interdisciplinary Applied Mathematics. Springer, New York.

Ewing, G. and J. Hermisson, 2010. Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26:2064–2065.

Excoffier, L., I. Dupanloup, E. Huerta-Snchez, V. C. Sousa, and M. Foll, 2013. Robust demographic inference from genomic and snp data. PLoS genetics 9:e1003905.

Excoffier, L., M. Foll, and R. J. Petit, 2009a. Genetic consequences of range expansions. Annual Review of Ecology Evolution and Systematics 40:481–501.

Excoffier, L., T. Hofer, and M. Foll, 2009b. Detecting loci under selection in a hierarchically structured population. Heredity 103:285–298.

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 10:564–567.

Excoffier, L., J. Novembre, and S. Schneider, 2000. Simcoal: A general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. Journal of Heredity 91:506–509.

Falush, D., M. Stephens, and J. K. Pritchard, 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587.

Fanello, C., F. Santolamazza, and A. Della Torre, 2002. Simultaneous identification of species and molecular forms of the anopheles gambiae complex by pcr-rflp. Medical and veterinary entomology 16:461–464.

Faubet, P. and O. E. Gaggiotti, 2008. A new bayesian method to identify the environmental factors that influence recent migration. Genetics 178:1491–1504.

Feder, J. L., S. H. Berlocher, J. B. Roethele, H. Dambroski, J. J. Smith, W. L. Perry, V. Gavrilovic, K. E. Filchak, J. Rull, and M. Aluja, 2003. Allopatric genetic origins for sympatric host-plant shifts and race formation in rhagoletis. Proceedings of the National Academy of Sciences of the United States of America 100:10314–10319.

Feder, J. L., R. Gejji, T. H. Q. Powell, and P. Nosil, 2011. Adaptive chromosomal divergence driven by mixed geographic mode of evolution. Evolution 65:2157–2170.

Fisher, R. A., 1930. The genetical theory of natural selection. Oxford Univ. Press, Oxford, U.K.

Foll, M. and O. Gaggiotti, 2006. Identifying the environmental factors that determine the genetic structure of populations. Genetics 174:875–891.

———, 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. Genetics 180:977–993.

George, P., M. V. Sharakhova, and I. V. Sharakhov, 2010. High-resolution cytogenetic map for the african malaria vector anopheles gambiae. Insect molecular biology 19:675–682.

Gillies, M. and M. Coetzee, 1987. A supplement to the anophelinae of africa south of the sahara. Publications of the South African Institute for Medical Research 55:1–143.

Gillies, M. T. and B. d. Meillon, 1968. The Anophelinae of Africa South Or the Sahara (Ethiopian Zoogeographical Region). South African Institute of Medical Research Johannesburg.

Gompert, Z., M. L. Forister, J. A. Fordyce, C. C. Nice, R. J. Williamson, and C. Alex Buerkle, 2010. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of lycaeides butterflies. Molecular Ecology 19:2455–2473.

Graham, C. H., S. R. Ron, J. C. Santos, C. J. Schneider, and C. Moritz, 2004. Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. Evolution 58:1781–1793.

Gray, E. M., K. A. C. Rocca, C. Costantini, and N. J. Besansky, 2009. Inversion 2la is associated with enhanced desiccation resistance in anopheles gambiae. Malaria Journal 8.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. W. Zhai, M. H. Y. Fritz, N. F. Hansen, E. Y. Durand, A. S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prufer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Hober, B. Hoffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, and S. Paabo, 2010. A draft sequence of the neandertal genome. Science 328:710–722.

Guerrero, R. F., F. Rousset, and M. Kirkpatrick, 2012. Coalescent patterns for chromosomal inversions in divergent populations. Philosophical Transactions of the Royal Society B: Biological Sciences 367:430–438.

Guillot, G., R. Leblois, A. Coulon, and A. C. Frantz, 2009. Statistical methods in spatial genetics. Molecular Ecology 18:4734–4756.

Gunther, T. and G. Coop, 2013. Robust identification of local adaptation from allele frequencies. Genetics 195:205–20.

Haldane, J., 1927. A mathematical theory of natural and artificial selection. part v. selection and mutation. Proceedings of the Cambridge Philosophical Society 23:838–844.

Harr, B., 2006. Genomic islands of differentiation between house mouse subspecies. Genome Research 16:730–737.

He, Q., D. L. Edwards, and L. L. Knowles, 2013. Integrative testing of how environments from the past to the present shape genetic structure across landscapes. Evolution 67:3386–3402.

Heckel, G., R. Burri, S. Fink, J. F. Desmet, and L. Excoffier, 2005. Genetic structure and colonization processes in european populations of the common vole, microtus arvalis. Evolution 59:2231–2242.

Hedrick, P. W., 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Molecular ecology 22:4606–4618.

Hermisson, J. and P. S. Pennings, 2005. Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. Genetics 169:2335–2352.

Hey, J., 2004. Fpg - a computer program for forward population genetic simulation .

———, 2010. Isolation with migration models for more than two populations. Molecular Biology and Evolution 27:905–920.

Hey, J. and R. Nielsen, 2007. Integration within the felsenstein equation for improved markov chain monte carlo methods in population genetics. Proceedings of the National Academy of Sciences of the United States of America 104:2785–2790.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25:1965–1978.

Hirzel, A. H., J. Hausser, D. Chessel, and N. Perrin, 2002. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? Ecology 83:2027–2036.

Hocking, R., H. Moors, and W. Van de Graaff, 1987. Geology of the Carnarvon Basin, Western Australia. State Print. Division, Perth.

Hohenlohe, P. A., J. Catchen, and W. A. Cresko, 2012. Population genomic analysis of model and nonmodel organisms using sequenced RAD tags, Pp. 235–260. Springer.

Holt, R. A., G. M. Subramanian, A. Halpern, G. G. Sutton, R. Charlab, D. R. Nusskern, P. Wincker, A. G. Clark, J. M. C. Ribeiro, R. Wides, S. L. Salzberg, B. Loftus, M. Yandell, W. H. Majoros, D. B. Rusch, Z. W. Lai, C. L. Kraft, J. F. Abril, V. Anthouard, P. Arensburger, P. W. Atkinson, H. Baden, V. de Berardinis, D. Baldwin, V. Benes, J. Biedler, C. Blass, R. Bolanos, D. Boscus, M. Barnstead, S. Cai, A. Center, K. Chatuverdi, G. K. Christophides, M. A. Chrystal, M. Clamp, A. Cravchik, V. Curwen, A. Dana, A. Delcher, I. Dew, C. A. Evans, M. Flanigan, A. Grundschober-Freimoser, L. Friedli, Z. P. Gu, P. Guan, R. Guigo, M. E. Hillenmeyer, S. L. Hladun, J. R. Hogan, Y. S. Hong, J. Hoover, O. Jaillon, Z. X. Ke, C. Kodira, E. Kokoza, A. Koutsos, I. Letunic, A. Levitsky, Y. Liang, J. J. Lin, N. F. Lobo, J. R. Lopez, J. A. Malek, T. C. McIntosh, S. Meister, J. Miller, C. Mobarry, E. Mongin, S. D. Murphy, D. A. O'Brochta, C. Pfannkoch, R. Qi, M. A. Regier, K. Remington, H. G. Shao, M. V. Sharakhova, C. D. Sitter, J. Shetty, T. J. Smith, R. Strong, J. T. Sun, D. Thomasova, L. Q. Ton, P. Topalis, Z. J. Tu, M. F. Unger, B. Walenz, A. H. Wang, J. Wang, M. Wang, X. L. Wang, K. J. Woodford, J. R. Wortman, M. Wu, A. Yao, E. M. Zdobnov, H. Y. Zhang, Q. Zhao, et al., 2002. The genome sequence of the malaria mosquito anopheles gambiae. Science 298:129–+.

Hopper, S. D. and P. Gioia, 2004. The southwest australian floristic region: Evolution and conservation of a global hot spot of biodiversity. Annual Review of Ecology Evolution and Systematics 35:623–650.

Hudson, R. R., 2002. Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Hugall, A., C. Moritz, A. Moussalli, and J. Stanisic, 2002. Reconciling paleodistribution models and comparative phylogeography in the wet tropics rainforest land snail gnarosophia bellendenkerensis (brazier 1875). Proceedings of the National Academy of Sciences of the United States of America 99:6112–6117.

Hull, J. M., A. C. Hull, B. N. Sacks, J. P. Smith, and H. B. Ernest, 2008. Landscape characteristics influence morphological and genetic differentiation in a widespread raptor (buteo jamaicensis). Molecular Ecology 17:810–824.

Innan, H. and Y. Kim, 2004. Pattern of polymorphism after strong artificial selection in a domestication event. Proceedings of the National Academy of Sciences of the United States of America 101:10667–10672.

Itan, Y., A. Powell, M. A. Beaumont, J. Burger, and M. G. Thomas, 2009. The origins of lactase persistence in europe. Plos Computational Biology 5.

Jaquiry, J., T. Broquet, A. H. Hirzel, J. Yearsley, and N. Perrin, 2011. Inferring landscape effects on dispersal from genetic distances: how far can we go? Molecular Ecology 20:692–705.

Jarman, S. N., R. D. Ward, and N. G. Elliott, 2002. Oligonucleotide primers for pcr amplification of coelomate introns. Marine Biotechnology 4:347–355.

Jeffreys, H., 1961. Theory of Probability, 3rd edition. Clarendon Press, Oxford.

Jensen, J., A. Bohonak, and S. Kelley, 2005. Isolation by distance, web service. BMC Genetics 6:13.

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

Jombart, T., S. Devillard, and F. Balloux, 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC genetics 11:94.

Joron, M., L. Frezal, R. T. Jones, N. L. Chamberlain, S. F. Lee, C. R. Haag, A. Whibley, M. Becuwe, S. W. Baxter, L. Ferguson, P. A. Wilkinson, C. Salazar, C. Davidson, R. Clark, M. A. Quail, H. Beasley, R. Glithero, C. Lloyd, S. Sims, M. C. Jones, J. Rogers, C. D. Jiggins, and R. H. ffrench Constant, 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature 477:203–206.

Kamali, M., A. Xia, Z. Tu, and I. V. Sharakhov, 2012. A new chromosomal phylogeny supports the repeated origin of vectorial capacity in malaria mosquitoes of the anopheles gambiae complex. PLoS pathogens 8:e1002960.

Karasov, T., P. W. Messer, and D. A. Petrov, 2010. Evidence that adaptation in drosophila is not limited by mutation at single sites. PLoS Genet 6:e1000924.

Keightley, P. D., U. Trivedi, M. Thomson, F. Oliver, S. Kumar, and M. Blaxter, 2009. Analysis of the genome sequences of three drosophila melanogaster spontaneous mutation accumulation lines. Genome research P. gr. 091231.109.

Kim, Y. and R. Nielsen, 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics 167:1513–1524.

Kimura, M., 1957. Some problems of stochastic processes in genetics. The Annals of Mathematical Statistics 28:882–901.

———, 1983. The neutral theory of molecular evolution. Cambridge Univ. Press, Cambridge, U.K.

Kimura, M. and J. F. Crow., 1970. An introduction to population genetics theory. Harper & Row, New York.

Kingman, J. F. C., 1982. The coalescent. Stochastic processes and their applications 13:235–248.

Kirkpatrick, M., 2011. How and why chromosome inversions evolve. Plos Biology 8:5.

Kirkpatrick, M. and N. Barton, 2006. Chromosome inversions, local adaptation and speciation. Genetics 173:419–434.

Knowles, L. L., 2001. Did the pleistocene glaciations promote divergence? tests of explicit refugial models in montane grasshopprers. Molecular Ecology 10:691–701.

———, 2008. Why does a method that fails continue to be used? Evolution 62:2713–2717.

———, 2009. Statistical phylogeography. Annual Review of Ecology Evolution and Systematics 40:593–612.

Knowles, L. L. and D. F. Alvarado-Serrano, 2010. Exploring the population genetic consequences of the colonization process with spatio-temporally explicit models: insights from coupled ecological, demographic and genetic models in montane grasshoppers. Molecular Ecology 19:3727–3745.

Knowles, L. L. and B. C. Carstens, 2007. Estimating a geographically explicit model of population divergence. Evolution 61:477–493.

Kuhner, M. K., 2006. Lamarc 2.0: maximum likelihood and bayesian estimation of population parameters. Bioinformatics 22:768–770.

Kuhner, M. K., J. Yamato, and J. Felsenstein, 1998. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics 149:429–434.

Landguth, E. L., S. A. Cushman, M. A. Murphy, and G. Luikart, 2010. Relationships between migration rates and landscape resistance assessed using individual-based simulations. Molecular Ecology Resources 10:854–862.

Lang, M., S. Murat, A. G. Clark, G. Gouppil, C. Blais, L. M. Matzkin, m. Guittard, T. Yoshiyama-Yanagawa, H. Kataoka, and R. Niwa, 2012. Mutations in the neverland gene turned drosophila pachea into an obligate specialist species. Science 337:1658–1661.

Lanzaro, G. C., Y. T. Toure, J. Carnahan, L. B. Zheng, G. Dolo, S. Traore, V. Petrarca, K. D. Vernick, and C. E. Taylor, 1998. Complexities in the genetic structure of anopheles gambiae populations in west africa as revealed by microsatellite dna analysis. Proceedings of the National Academy of Sciences of the United States of America 95:14260–14265.

Lee, C. R. and T. Mitchell-Olds, 2011. Quantifying effects of environmental and geographical factors on patterns of genetic differentiation. Molecular Ecology 20:4631–4642.

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

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 25:601–615.

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 10:831–844.

Lehmann, T., N. J. Besansky, W. A. Hawley, T. G. Fahey, L. Kamau, and F. H. Collins, 1997. Microgeographic structure of anopheles gambiae in western kenya based on mtdna and microsatellite loci. Molecular Ecology 6:243–253.

Lehmann, T., C. R. Blackston, N. J. Besansky, A. A. Escalante, F. H. Collins, and W. A. Hawley, 2000. The rift valley complex as a barrier to gene flow for anopheles gambiae in kenya: The mtdna perspective. Journal of Heredity 91:165–168.

Lehmann, T., W. A. Hawley, H. Grebert, and F. H. Collins, 1998. The effective population size of anopheles gambiae in kenya: Implications for population structure. Molecular Biology and Evolution 15:264–276.

Lehmann, T., W. A. Hawley, H. Grebert, M. Danga, F. Atieli, and F. H. Collins, 1999. The rift valley complex as a barrier to gene flow for anopheles gambiae in kenya. Journal of Heredity 90:613–621.

Leuenberger, C. and D. Wegmann, 2010. Bayesian computation and model selection without likelihoods. Genetics 184:243–252.

Lewontin, R. C. and K. Kojima, 1960. The evolutionary dynamics of complex polymorphisms. Evolution 14:458–472.

Li, H. and R. Durbin, 2009. Fast and accurate short read alignment with burrowswheeler transform. Bioinformatics 25:1754–1760.

Li, W. H. and M. Nei, 1974. Stable linkage disequilibrium without epistasis in subdivided population. Theoretical population biology 6:173–183.

Lobo, N. F., D. M. Sangare, A. A. Regier, K. R. Reidenbach, D. A. Bretz, M. V. Sharakhova, S. J. Emrich, S. F. Traore, C. Costantini, N. J. Besansky, and F. H. Collins, 2010. Breakpoint structure of the anopheles gambiae 2rb chromosomal inversion. Malaria Journal 9.

Lotterhos, K. E. and M. C. Whitlock, 2014. Evaluation of demographic history and neutral parameterization on the performance of fst outlier tests. Molecular ecology 23:2178–2192.

Lowry, D. B. and J. H. Willis, 2010. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. Plos Biology 8:14.

Lozier, J. D., P. Aniello, and M. J. Hickerson, 2009. Predicting the distribution of sasquatch in western north america: anything goes with ecological niche modelling. Journal of Biogeography 36:1623–1627.

Lynch, M., 2010. Scaling expectations for the time to establishment of complex adaptations. Proceedings of the National Academy of Sciences of the United States of America 107:16577–16582.

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

Manel, S., H. C. Williams, and S. J. Ormerod, 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38:921–931.

Manoukis, N. C., J. R. Powell, M. B. Toure, A. Sacko, F. E. Edillo, M. B. Coulibaly, S. F. Traore, C. E. Taylor, and N. J. Besansky, 2008. A test of the chromosomal theory of ecotypic speciation in anopheles gambiae. Proceedings of the National Academy of Sciences of the United States of America 105:2940–2945.

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

McKenzie, N. J., D. W. Jacquier, L. J. Ashton, and H. P. Cresswell, 2000. Estimation of soil properties using the Atlas of Australian Soils. Technical Report 11/00, CSIRO Land and Water, Canberra.

McRae, B. H., 2006. Isolation by resistance. Evolution 60:1551–1561.

Meirmans, P. G., 2012. The trouble with isolation by distance. Molecular Ecology 21:2839–2846.

Melville, J., L. P. Shoo, and P. Doughty, 2008. Phylogenetic relationships of the heath dragons (rankinia adelaidensis and r. parviceps) from the south-western australian biodiversity hotspot. Australian Journal of Zoology 56:159–171.

Mendez, M., H. C. Rosenbaum, A. Subramaniam, C. Yackulic, and P. Bordino, 2010. Isolation by environmental distance in mobile marine species: molecular ecology of franciscana dolphins at their southern range. Molecular Ecology 19:2212–2228.

Mevik, B. H. and R. Wehrens, 2007. The pls package: principal component and partial least squares regression in r. Journal of Statistical Software 18:1–24.

Moral, P. D., A. Doucet, and A. Jasra, 2012. An adaptive sequential monte carlo method for approximate bayesian computation. Stat Comput 22:1009–1020.

Morgan, K., S. M. O'Loughlin, B. Chen, Y. M. Linton, D. Thongwat, P. Somboon, M. Y. Fong, R. Butlin, R. Verity, A. Prakash, P. T. Htun, T. Hlaing, S. Nambanya, D. Socheat, T. H. Dinh, and C. Walton, 2011. Comparative phylogeography reveals a shared impact of pleistocene environmental change in shaping genetic diversity within nine anopheles mosquito species across the indo-burma biodiversity hotspot. Molecular Ecology 20:4533–4549.

Moritz, C. and D. P. Faith, 1998. Comparative phylogeography and the identification of genetically divergent areas for conservation. Molecular Ecology 7:419–429.

Moussalli, A., C. Moritz, S. E. Williams, and A. C. Carnaval, 2009. Variable responses of skinks to a common history of rainforest fluctuation: concordance between phylogeography and palaeo-distribution models. Molecular Ecology 18:483–499.

Myers, N., R. A. Mittermeier, C. G. Mittermeier, G. A. B. da Fonseca, and J. Kent, 2000. Biodiversity hotspots for conservation priorities. Nature 403:853–858.

Nadeau, N. J., A. Whibley, R. T. Jones, J. W. Davey, K. K. Dasmahapatra, S. W. Baxter, M. A. Quail, M. Joron, M. L. Blaxter, and J. Mallet, 2012. Genomic islands of divergence in hybridizing heliconius butterflies identified by large-scale targeted sequencing. Philosophical Transactions of the Royal Society B: Biological Sciences 367:343–353.

Navarro, A. and N. H. Barton, 2003. Accumulating postzygotic isolation genes in parapatry: A new twist on chromosomal speciation. Evolution 57:447–459.

Navarro, A., E. Betran, A. Barbadilla, and A. Ruiz, 1997. Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. Genetics 146:695–709.

Neafsey, D., M. Lawniczak, D. Park, S. Redmond, M. Coulibaly, S. Traore, N. Sagnon, C. Costantini, C. Johnson, and R. Wiegand, 2010. Snp genotyping defines complex gene-flow boundaries among african malaria vector mosquitoes. Science 330:514–517.

Nei, M., K.-I. Kojima, and H. E. Schaffer, 1967. Frequency changes of new inversions in populations under mutation-selection equilibria. Genetics 57:741–750.

Neuenschwander, S., C. R. Largiader, N. Ray, M. Currat, P. Vonlanthen, and L. Excoffier, 2008. Colonization history of the swiss rhine basin by the bullhead (cottus gobio): inference under a bayesian spatially explicit framework. Molecular Ecology 17:757–772.

Nielsen, R. and M. A. Beaumont, 2009. Statistical inferences in phylogeography. Molecular Ecology 18:1034–1047.

Nielsen, R., I. Hellmann, M. Hubisz, C. Bustamante, and A. G. Clark, 2007. Recent and ongoing selection in the human genome. Nature Reviews Genetics 8:857–868.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante, 2005. Genomic scans for selective sweeps using snp data. Genome Research 15:1566–1575.

Noor, M. A. F., K. L. Grams, L. A. Bertucci, and J. Reiland, 2001. Chromosomal inversions and the reproductive isolation of species. Proceedings of the National Academy of Sciences of the United States of America 98:12084–12088.

Northcote, K. H., G. G. Beckmann, E. Bettenay, H. M. Churchward, D. C. Van Dijk, G. M. Dimmock, G. D. Hubble, R. F. Isbell, W. M. McArthur, G. G. Murtha, K. D. Nicolls, T. R. Paton, C. H. Thompson, A. A. Webb, and M. J. Wright, 1968. Atlas of Australian Soils, Sheets 1 to 10. With explanatory data. CSIRO Aust. and Melbourne University Press, Melbourne.

Nosil, P. and J. L. Feder, 2012. Genomic divergence during speciation: causes and consequences introduction. Royal Society Philosophical Transactions Biological Sciences 367:332–342.

Nosil, P., D. J. Funk, and D. Ortiz-Barrientos, 2009. Divergent selection and heterogeneous genomic divergence. Molecular Ecology 18:375–402.

Ohta, T. and K. I. Kojima, 1968. Survival probabilities of new inversions in large populations. Biometrics 24:501–516.

Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. Henry, H. Stevens, and H. Wagner, 2012. vegan: Community ecology package. r package version 2.0-3. http://cran.r-project.org/package=vegan .

Oliveira, E., P. Salgueiro, K. Palsson, J. L. Vicente, A. P. Arez, T. G. Jaenson, A. Caccone, and J. Pinto, 2008. High levels of hybridization between molecular forms of anopheles gambiae from guinea bissau. J Med Entomol 45:1057–63.

Onyabe, D. Y. and J. E. Conn, 2001. Genetic differentiation of the malaria vector anopheles gambiae across nigeria suggests that selection limits gene flow. Heredity 87:647–658.

Orr, H. A., 1998. The population genetics of adaptation: The distribution of factors fixed during adaptive evolution. Evolution 52:935–949.

Orr, H. A. and A. J. Betancourt, 2001. Haldane's sieve and adaptation from the standing genetic variation. Genetics 157:875–884.

Orr, H. A. and R. L. Unckless, 2008. Population extinction and the genetics of adaptation. American Naturalist 172:160–169.

Otto, S. P. and M. C. Whitlock, 1997. The probability of fixation in populations of changing size. Genetics 146:723–733.

Papadopulos, A. S. T., W. J. Baker, D. Crayn, R. K. Butlin, R. G. Kynast, I. Hutton, and V. Savolainen, 2011. Speciation with gene flow on lord howe island. Proceedings of the National Academy of Sciences of the United States of America 108:13188–13193.

Peischl, S., E. Koch, R. Guerrero, and M. Kirkpatrick, 2013. A sequential coalescent algorithm for chromosomal inversions. Heredity 111:200–209.

Perrier, C., R. Guyomard, J. L. Bagliniere, and G. Evanno, 2011. Determinants of hierarchical genetic structure in atlantic salmon populations: environmental factors vs. anthropogenic influences. Molecular Ecology 20:4231–4245.

Peterson, A. T., J. Soberon, R. G. Pearson, R. P. Anderson, E. Martinez-Meyer, M. Nakamura, and M. Araujo, 2011. Ecological niches and geographic distributions. Monographs in Population Biology, 49. Princeton University Press.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra, 2012. Double digest radseq: an inexpensive method for de novo snp discovery and genotyping in model and non-model species. PloS one 7:e37135.

Phillips, S. J., R. P. Anderson, and R. E. Schapire, 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190:231–259.

Pombi, M., B. Caputo, C. Costantini, M. A. Di Deco, M. Coluzzi, A. della Torre, F. Simard, N. J. Besansky, and V. Petrarca, 2008. Chromosomal plasticity and evolutionary potential in the malaria vector anopheles gambiae sensu stricto: insights from three decades of rare paracentric inversions. BMC evolutionary biology 8:309.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman, 1999. Population growth of human y chromosomes: A study of y chromosome microsatellites. Molecular Biology and Evolution 16:1791–1798.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Przeworski, M., G. Coop, and J. D. Wall, 2005. The signature of positive selection on standing genetic variation. Evolution 59:2312–2323.

Rabosky, D. L., K. P. Aplin, S. C. Donnellan, and S. B. Hedges, 2004. Molecular phylogeny of blindsnakes (ramphotyphlops) from western australia and resurrection of ramphotyphlops bicolor (peters, 1857). Australian Journal of Zoology 52:531–548.

Rafajlovic, M., A. Klassmann, A. Eriksson, T. Wiehe, and B. Mehlig, 2014. Demography-adjusted tests of neutrality based on genome-wide snp data. Theoretical Population Biology .

Raufaste, N. and F. Rousset, 2001. Are partial mantel tests adequate? Evolution 55:1703–1705.

Ray, N., M. Currat, P. Berthier, and L. Excoffier, 2005. Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. Genome Research 15:1161–1167.

Rego, C., J. Balanya, I. Fragata, M. Matos, E. L. Rezende, and M. Santos, 2010. Clinal patterns of chromosomal inversion polymorphisms in drosophila subobscura are partly associated with thermal preferences and heat stress resistance. Evolution 64:385–397.

Richards, C. L., B. C. Carstens, and L. L. Knowles, 2007. Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. Journal of Biogeography 34:1833–1845.

Rieseberg, L. H., 2001. Chromosomal rearrangements and speciation. Trends in Ecology & Evolution 16:351–358.

Robert, C. P., J.-M. Cornuet, J.-M. Marin, and N. S. Pillai, 2011. Lack of confidence in approximate bayesian computation model choice. Proceedings of the National Academy of Sciences 108:15112–15117.

Ronen, R., N. Udpa, E. Halperin, and V. Bafna, 2013. Learning natural selection from the site frequency spectrum. Genetics 195:181–193.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander, 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll, R. Gaudet, S. F. Schaffner, E. S. Lander, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, W. Sun, H. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, et al., 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–8.

Santos, M., 2009. Recombination load in a chromosomal inversion polymorphism of drosophila subobscura. Genetics 181:803–809.

Scheet, P. and M. Stephens, 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. American Journal of Human Genetics 78:629–644.

Scoville, A. G. and M. E. Pfrender, 2010. Phenotypic plasticity facilitates recurrent rapid adaptation to introduced predators. Proceedings of the National Academy of Sciences of the United States of America 107:4260–4263.

Shah, V. and B. McRae, 2008. Circuitscape: a tool for landscape ecology. Pp. 62–66. Proceedings of the 7th Python in Science Conference (Scipy 2008).

Sharakhov, I. V., B. J. White, M. V. Sharakhova, J. Kayondo, N. F. Lobo, F. Santolamazza, A. della Torre, F. Simard, F. H. Collins, and N. J. Besansky, 2006. Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (2la) in the anopheles gambiae complex. Proceedings of the National Academy of Sciences of the United States of America 103:6258–6262.

Shirk, A. J., S. A. Cushman, and E. L. Landguth, 2012. Simulating pattern-process relationships to validate landscape genetic models. International Journal of Ecology 2012:8 (Article ID 539109).

Simard, F., D. Ayala, G. Kamdem, M. Pombi, J. Etouna, K. Ose, J.-M. Fotsing, D. Fontenille, N. Besansky, and C. Costantini, 2009. Ecological niche partitioning between anopheles gambiae molecular forms in cameroon: the ecological side of speciation. BMC Ecology 9:17.

Sinka, M. E., Y. Rubio-Palis, S. Manguin, A. P. Patil, W. H. Temperley, P. W. Gething, T. Van Boeckel, C. W. Kabaria, R. E. Harbach, and S. I. Hay, 2010. The dominant anopheles vectors of human malaria in the americas: occurrence data, distribution maps and bionomic prcis. Parasites & vectors 3:72.

Smouse, P. E., J. C. Long, and R. R. Sokal, 1986. Multiple-regression and correlation extensions of the mantel test of matrix correspondence. Systematic Zoology 35:627–632.

Soria-Carrasco, V., Z. Gompert, A. A. Comeault, T. E. Farkas, T. L. Parchman, J. S. Johnston, C. A. Buerkle, J. L. Feder, J. Bast, T. Schwander, et al., 2014. Stick insect genomes reveal natural selections role in parallel speciation. Science 344:738–742.

Sork, V. L., F. W. Davis, R. Westfall, A. Flint, M. Ikegami, H. F. Wang, and D. Grivet, 2010. Gene movement and genetic association with regional climate gradients in california valley oak (quercus lobata ne) in the face of climate change. Molecular Ecology 19:3806–3823.

Stockwell, D. R. B. and A. T. Peterson, 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling 148:1–13.

Storfer, A., M. A. Murphy, J. S. Evans, C. S. Goldberg, S. Robinson, S. F. Spear, R. Dezzani, E. Delmelle, L. Vierling, and L. P. Waits, 2007. Putting the 'landscape' in landscape genetics. Heredity 98:128–142.

Storfer, A., M. A. Murphy, S. F. Spear, R. Holderegger, and L. P. Waits, 2010. Landscape genetics: where are we now? Molecular Ecology 19:3496–3514.

Storr, G. and G. Harold, 1978. Herpetofauna of the shark bay region, western australia. Records of the Western Australian Museum 6:449–467.

———, 1980. Herpetofauna of the zuytdorp coast and hinterland, western australia. Records of the Western Australian Museum 8:359–375.

Strasburg, J. L., M. Kearney, C. Moritz, and A. R. Templeton, 2007. Combining phylogeography with distribution modeling: Multiple pleistocene range expansions in a parthenogenetic gecko from the australian arid zone. Plos One 2:e760.

Tajima, F. and M. Nei, 1984. Estimation of evolutionary distance between nucleotide sequences. Molecular Biology and Evolution 1:269–285.

Tavare, S., D. J. Balding, R. C. Griffiths, and P. Donnelly, 1997. Inferring coalescence times from dna sequence data. Genetics 145:505–518.

Team, R. C., 2012. R: A language and environment for statistical computing.

Temu, E. A. and G. Y. Yan, 2005. Microsatellite and mitochondrial genetic differentiation of anopheles arabiensis (diptera : Culicidae) from western kenya, the great rift valley, and coastal kenya. American Journal of Tropical Medicine and Hygiene 73:726–733.

Thompson, M. and C. Jiggins, 2014. Supergenes and their role in evolution. Heredity .

Thomson, R. C., I. J. Wang, and J. R. Johnson, 2010. Genome-enabled development of dna markers for ecology, evolution and conservation. Molecular Ecology 19:2184–2195.

della Torre, A., C. Costantini, N. J. Besansky, A. Caccone, V. Petrarca, J. R. Powell, and M. Coluzzi, 2002. Speciation within anopheles gambiae - the glass is half full. Science 298:115–117.

della Torre, A., Z. J. Tu, and V. Petrarca, 2005. On the distribution and genetic differentiation of anopheles gambiae s.s. molecular forms. Insect Biochemistry and Molecular Biology 35:755–769.

Turner, M. G., 2005. Landscape ecology: What is the state of the science?, *Annual Review of Ecology Evolution and Systematics*, vol. 36, Pp. 319–344.

Turner, T. L. and M. W. Hahn, 2007. Locus- and population-specific selection and differentiation between incipient species of anopheles gambiae. Molecular Biology and Evolution 24:2132–2138.

Voight, B. F., S. Kudaravalli, X. Q. Wen, and J. K. Pritchard, 2006. A map of recent positive selection in the human genome. PLoS Biology 4:446–458.

Wang, C., S. Zollner, and N. A. Rosenberg, 2012. A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS Genet 8:e1002886.

Wang, E. T., G. Kodama, P. Baidi, and R. K. Moyzis, 2006. Global landscape of recent inferred darwinian selection for homo sapiens. Proceedings of the National Academy of Sciences of the United States of America 103:135–140.

Wang, I. J., 2010. Recognizing the temporal distinctions between landscape genetics and phylogeography. Molecular Ecology 19:2605–2608.

Wegmann, D., M. Currat, and L. Excoffier, 2006. Molecular diversity after a range expansion in heterogeneous environments. Genetics 174:2009–2020.

Wegmann, D. and L. Excoffier, 2010. Bayesian inference of the demographic history of chimpanzees. Molecular Biology and Evolution 27:1425–1435.

Wegmann, D., C. Leuenberger, and L. Excoffier, 2009. Efficient approximate bayesian computation coupled with markov chain monte carlo without likelihood. Genetics 182:1207–1218.

Weir, B. S. and C. C. Cockerham, 1984. Estimating f-statistics for the analysis of population structure. Evolution 38:1358–1370.

White, B. J., C. Cheng, D. Sangar, N. F. Lobo, F. H. Collins, and N. J. Besansky, 2009. The population genomics of trans-specific inversion polymorphisms in anopheles gambiae. Genetics 183:275–288.

White, B. J., C. Cheng, F. Simard, C. Costantini, and N. J. Besansky, 2010. Genetic association of physically unlinked islands of genomic divergence in incipient species of anopheles gambiae. Molecular Ecology 19:925–939.

White, B. J., F. H. Collins, and N. J. Besansky, 2011. Evolution of anopheles gambiae in relation to humans and malaria. Annual Review of Ecology, Evolution, and Systematics 42:111–132.

White, B. J., M. W. Hahn, M. Pombi, B. J. Cassone, N. F. Lobo, F. Simard, and N. J. Besansky, 2007a. Localization of candidate regions maintaining a common polymorphic inversion (2la) in anopheles gambiae. PLoS Genet 3:e217.

White, B. J., F. Santolamazza, L. Kamau, M. Pombi, O. Grushko, K. Mouline, C. Brengues, W. Guelbeogo, M. Coulibaly, J. K. Kayondo, I. Sharakhov, F. Simard, V. Petrarca, A. Della Torre, and N. J. Besansky, 2007b. Molecular karyotyping of the 2la inversion in anopheles gambiae. American Journal of Tropical Medicine and Hygiene 76:334–339.

Wilson, S. and G. Swan, 2008. A complete guide to reptiles of Australia, 2nd Ed. New Holland Publishers, Sydney, Australia.

Wright, S., 1931. Evolution in mendelian populations. Genetics 16:0097–0159.

Xu, J. W., M. Perez-Losada, C. G. Jara, and K. A. Crandall, 2009. Pleistocene glaciation leaves deep signature on the freshwater crab aegla alacalufi in chilean patagonia. Molecular Ecology 18:904–918.

Yang, W.-Y., J. Novembre, E. Eskin, and E. Halperin, 2012. A model-based approach for analysis of spatial structure in genetic data. Nature genetics 44:725–731.

Yawson, A. E., D. Weetman, M. D. Wilson, and M. J. Donnelly, 2007. Ecological zones rather than molecular forms predict genetic differentiation in the malaria vector anopheles gambiae s.s. in ghana. Genetics 175:751–761.

Yeaman, S., 2013. Genomic rearrangements and the evolution of clusters of locally adaptive loci. Proceedings of the National Academy of Sciences 110:E1743–E1751.

Yeaman, S. and M. C. Whitlock, 2011. The genetic architecture of adaptation under migration-selection balance. Evolution 65:1897–1911.

Zahar, A., 1990. Vector bionomics in the epidemiology and control of malaria. part ii. the who european region & the who eastern mediterranean region. volume ii. applied field studies. section iii: vector bionomics, malaria epidemiology and control by geographical areas.(a) the mediterranean basin .

Zellmer, A. J. and L. L. Knowles, 2009. Disentangling the effects of historic vs. contemporary landscape structure on population genetic divergence. Molecular Ecology 18:3593–3602.

Zhong, D. B., E. A. Temu, T. Guda, L. Gouagna, D. Menge, A. Pai, J. Githure, J. C. Beier, and G. Y. Yan, 2006. Dynamics of gene introgression in the african malaria vector anopheles gambiae. Genetics 172:2359–2365.