UNIVERSITY OF CALIFORNIA, SAN DIEGO

Testing Modes of Speciation and the Effects of Demography on Selection in Closely Related Species

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in

Biology

by

Andrea Susan Putnam

Committee in charge:

Professor Peter Andolfatto, Chair
Professor Doris Bachtrog
Professor Ron Burton
Professor Kevin Thornton
Professor Chris Wills

2008

The dissertation of Andrea Susan Putnam is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2008

TABLE OF CONTENTS

Signature Page …… iii

Table of Contents…… iv

List of Figures……... v

List of Tables…… vi

Acknowledgements…… vii

Vita…… ix

Abstract…… x

Part I. Testing Modes of Speciation..... 1

Chapter 1. Discordant divergence times among Z-chromosome regions between two ecologically distinct swallowtail butterfly species…… 2

Chapter 2. A stepwise approximate Bayesian approach to estimating speciation times and ancestral population size in closely related species..… 27

Part II. The effects of demography on selection..... 69

Chapter 3. Influence of demography on detecting positive selection at *Drosophila melanogaster* female reproductive genes..……..…….. 70

LIST OF FIGURES

Figure 1.1 Frequencies of three Z-linked loci……..……...……..... 5
Figure 1.2 Model of allopatric species divergence…….……..... 7
Figure 1.3 Approximate Bayesian joint distribution of a) θ and b) ρ/θ……..... 7
Figure 1.4 Diagram of *Ldh* gene……... 10
Figure 1.5 Posterior distribution of *T*.……..……. 11
Figure 1.6 The polymorphism frequency spectrum…… 13
Figure 1.7 Performance of divergence time estimation……..... 24
Figure 1.8 Approximate Bayesian posterior distribution of θ per locus...…… 25
Figure 1.9 Approximate Bayesian posterior distribution of ρ/θ per locus..... 26
Figure 2.1 A model of allopatric speciation…….……. 64
Figure 2.2 Joint posterior distribution of θ_A……. 65
Figure 2.3 Posterior distribution of per locus *T*……... 66
Figure 2.4 Posterior distribution of a) θ and b) ρ/θ in humans and chimps… 67
Figure 3.1 Tests of neutrality……. 100
Figure 3.2 Distribution of polymorphisms in candidate gene…...…… 101
Figure 3.3 Estimates of the fraction of amino acid divergence…… 102
Figure 3.4 Distribution of polymorphisms in sex biased genes…… 103
Figure 3.5 Distribution of *P* values for Fay and Wu's *H*…… 106

LIST OF TABLES

Table 1.1 Polymorphism and divergence statistics…..……..…….…. 9
Table 1.2 Divergence time, *T*, under allopatry……. 12
Table 1.3 Goodness-of-fit test for one divergence time versus five divergence times…….…….... 13
Table 1.4 Sampling locations for *Papilio*…… 19
Table 1.5 Primer sequences for Z-linked loci…… 21
Table 1.6 Summary of the frequency distribution……. 22
Table 2.1 Summary statistics of observed data used in simulations……. 58
Table 2.2 *P. glaucus* and *P. canadensis* per locus estimates of *T*……... 59
Table 2.3 Human and chimpanzee per locus estimates of *T*……. 59
Table 2.4 Goodness-of-fit test for a model of allopatry…… 59
Table 2.5 Performance of STE…… 60
Table 2.6 Summary statistics of *TH*…… 61
Table 2.7 Chimpanzee and human summary statistics…….. 62
Table 3.1 Tests of neutrality in the candidate and control loci……..... 98
Table 3.2 *P*-values for tests of neutrality…… 98
Table 3.3 Estimates of the fraction of amino acid divergence…… 99
Table 3.4 Cytological positions and sex-biased expression in candidate loci. 104
Table 3.5 Summary statistics of candidate and control loci……... 105
Table 3.6 Polymorphism and divergence of candidate and control loci…… 105

ACKNOWLEDGEMENTS

It is a pleasure to thank the many people who made this thesis possible. It is difficult to overstate my gratitude to my Ph.D. advisor, Peter Andolfatto. No part of this thesis could have been written without his guidance and valuable insights. Over the past four years, Peter has provided me with a tremendous education in population genetics, going to great lengths to help me develop this dissertation. I would also like to thank Doris Bachtrog who has always been encouraging of my work and provided sound advice. I am indebted to Kevin Thornton for his patient help with various applications and especially his statistical expertise. I am also grateful to Ron Burton and Chris Wills for providing a fresh perspective on my thesis and useful comments. The opportunity to work with colleagues like Jeff Jensen and Fedya Kondrashov provided me with a wealth of editorial and technical advice as well as entertainment. Other members the Andolfatto and Bachtrog labs offered encouragement and support, and so I would like to thank Brielle Fischman, Anna Bree, Karen Wong, Mayuri Naidu, Michael Breen, Clinton Edwards, and Tatiana Gurbich.

Finally, I am indebted all my friends and family who made my time at UCSD a pleasure and who reminded me about things going on in the real world: Kristen Marhaver, Phil Fenberg, Boris Igic, Tali Vardi, Steve Smirga, Rachel Borgatti, Angelica Cibrian, Sara Carlson, Shahrina Chowdhury, Reem Hajjar, Sergios

Orestis Kolokotronis, Tim Paape, Ben Evans, Stuart Bogatko, Isaac Mehl, Philip Putnam, Thea Convissar, and Lynne Putnam. Jeremy Boyd deserves a special mention for his companionship and humor.

VITA

2000	Bachelor of Arts New York University
2004	Master of Science Columbia University
2008	Doctor of Philosophy University of California, San Diego

PUBLICATIONS

Putnam AS, Edwards C, Andolfatto P. (*in prep*) Influence of demography on detecting positive selection at *Drosophila melanogaster* female reproductive genes.

Donaldson ZR, Putnam AS, Kondrashov F, Yaohui B, Stoinski TL, Hammock E, Young LJ. (*accepted,* BMC Evolution) Evolution of a behavior-linked microsatellite-containing element in the 5' flanking region of the primate AVPR1A gene.

Putnam AS, Scriber JM, Andolfatto P. (2007) Discordant divergence times among Z chromosome regions between two ecologically distinct swallowtail butterfly species. Evolution 61(4):912-927.

Campbell P, Putnam AS, Bonney C, Bilgin R, Morales JC, Kunz TH, Ruedas LA. (2007) Contrasting patterns of genetic differentiation between endemic and widespread species of fruit bats (Chiroptera: Pteropodidae) in Sulawesi, Indonesia. Molecular Phylogenetics and Evolution, Volume 44, (1):474-482.

ABSTRACT OF THE DISSERTATION

Testing Modes of Speciation and the Effects of Demography on Selection in Closely Related Species

by

Andrea Susan Putnam

Doctor of Philosophy in Biology

University of California, San Diego, 2008

Professor Peter Andolfatto

A fundamental controversy in the study of speciation is whether allopatry is the dominant mode of speciation. Understanding the genetic architecture of reproductive isolation between incipient species may be important in resolving this question. One approach to examining reproductive isolation is to estimate divergence times at multiple loci to determine if gene flow was restricted instantaneously among species (allopatry) or whether the divergence is better explained by stages of reproductive isolation (parapatry). The first two chapters of my dissertation develop a model to test these modes of speciation in sister species and to predict which genes were important in the early stages of speciation. The last chapter of my thesis focuses on genomic scans for these speciation factors to determine the degree of positive selection acting on genes with enriched expression in the female reproductive tract. This class of genes is thought to be the target of early genomic incompatibilities driving reproductive

isolation in many incipient species.

In Chapters 1 and 2, I use two swallowtail butterfly species, *Papilio glaucus* and *Papilio canadensis*, as a system to study the genetic basis of reproductive isolation. The two species make a compelling system because they differ in many ecological traits, like mimicry, that are linked to their sex chromosomes. Yet the species are still capable of forming fertile hybrids. Using markers linked to the Z chromosome, I ask whether a single divergence time estimate (allopatry) fits the polymorphism data better than a complex speciation process (parapatry). I develop an approximate Bayesian coalescent-based method to estimate the ancestral population size of the species and the per locus and joint divergence times. Using this framework, one can also identify candidates for reproductive isolation factors. These loci are expected to have deeper divergence time estimates than randomly selected loci. I establish that allopatric speciation is unlikely in *P. glaucus* and *P. canadensis*, and identify two genes that may be linked to speciation factors.

In Chapter 3, I examine the effects of demography on scans for positive selection. Reproductive isolation between incipient species may arise as a by-product of divergent selection on a small number of phenotypic traits. If these traits have a simple genetic basis, then a relatively small number of genes may be responsible for reproductive isolation. In scans for positive selection, researchers often focus on genes preferentially expressed in the reproductive

tract because interactions between male and female proteins may be the earliest drivers of reproductive isolation. Using a variety of single and multilocus tests for selection in ancestral African *Drosophila melanogaster*, I compare polymorphism in 9 female reproductive genes that are candidates for positive selection to a control set of 137 randomly chosen genes. Though I find evidence for recent positive selection at 2 of the 9 candidate loci in *D. melanogaster* from Zimbabwe, I find no evidence supporting the notion that the candidate loci are more frequent targets of adaptive evolution. These results demonstrate that a previous study identifying elevated rates of positive selection on female reproductive genes as a group may be incorrect because the study examined a population that recently underwent a severe bottleneck. This study highlights the importance of incorporating demographic parameters into scans for positive selection. Taken together, my dissertation describes a method for systematically evaluating the prevalence of allopatric speciation and shows the demographic considerations necessary for identifying candidate genes underlying reproductive isolation.

Part I.

Testing Modes of Speciation

Chapter 1

Discordant divergence times among Z-chromosome regions between two ecologically distinct swallowtail butterfly species

ORIGINAL ARTICLE

doi:10.1111/j.1558-5646.2007.00076.x

DISCORDANT DIVERGENCE TIMES AMONG Z-CHROMOSOME REGIONS BETWEEN TWO ECOLOGICALLY DISTINCT SWALLOWTAIL BUTTERFLY SPECIES

Andrea S. Putnam,[1,2] J. Mark Scriber,[3] and Peter Andolfatto[1]

[1]*Section of Ecology, Behavior, and Evolution, Division of Biological Sciences, University of California, San Diego, La Jolla, California 92093*

[2]*E-mail: asputnam@biomail.ucsd.edu*

[3]*Department of Entomology, Michigan State University, East Lansing, Michigan 48824*

Received July 27, 2006
Accepted December 2, 2006

We investigate multilocus patterns of differentiation between parental populations of two swallowtail butterfly species that differ at a number of ecologically important sex-linked traits. Using a new coalescent-based approach, we show that there is significant heterogeneity in estimated divergence times among five Z-linked markers, rejecting a purely allopatric speciation model. We infer that the Z chromosome is a mosaic of regions that differ in the extent of historical gene flow, potentially due to isolating barriers that prevent the introgression of species-specific traits that result in hybrid incompatibilities. Surprisingly, a candidate region for a strong barrier to introgression, *Ldh*, does not show a significantly deeper divergence time than other markers on the Z chromosome. Our approach can be used to test alternative models of speciation and can potentially assign chronological order to the appearance of factors contributing to reproductive isolation between species.

KEY WORDS: **Approximate Bayesian computation, divergence, hybrid zones, introgression, Lepidoptera, *Papilio*, speciation.**

The understanding of speciation—the evolution of barriers to gene flow between taxa—is central to our current understanding of evolutionary biology (Dobzhansky 1937; Mayr 1942; Coyne and Orr 2004). Evolutionary geneticists are particularly interested in understanding the genetic basis of speciation, namely, how many and which genes are involved, what types of changes to these genes contribute to reproductive isolation, and what population genetic processes led to the fixation of different alleles at these genes (Orr et al. 2004). Historically, the study of the genetic basis of speciation has been hampered by the fact that alleles causing reproductive isolation are not particularly amenable to genetic analysis.

Despite this formidable obstacle, evolutionary geneticists have recently made progress in identifying these barriers to gene flow using two types of approaches. The first is a series of clever genetic mapping experiments designed to pinpoint genomic regions, and in some cases individual genes, causing reproductive isolation (Wittbrodt et al. 1989; True et al. 1996; Ting et al. 1998; Barbash et al. 2003; Presgraves 2003; Presgraves et al. 2003; Tao et al. 2003; Sawamura et al. 2004; Moehring et al. 2006; Turner et al. 2005). Presgraves (2003) estimated that between *Drosophila melanogaster* and its closest known relative, *D. simulans*, intrinsic hybrid inviability alone involves incompatible alleles at almost 200 genes. Hybrid male sterility factors have also been shown to disproportionately accumulate on the X chromosome (True et al. 1996; Tao et al. 2003), as predicted if the alleles causing them are partly recessive and positively selected (Charlesworth et al. 1987). In addition, several individual genes causing reproductive

isolation have been shown to be targets of recurrent adaptive amino acid substitution (Ting et al. 1998; Barbash et al. 2003; Presgraves et al. 2003). Interestingly, these latter two observations suggest a link, if only indirect, between adaptive evolution and the evolution of reproductive isolation.

A second approach is based on statistical analysis of hybrid zones between parapatric, incompletely isolated species. The principle of this approach is that hybrid zones, in which hybrids are less fit than parental populations, represent a conflict between selection against unfit hybrids promoting species divergence and gene flow through dispersal preventing divergence (Slatkin 1973; Endler 1977; Mallet and Barton 1989; Harrison 1990; Barton 2001). The degree to which a genomic region can introgress across a hybrid zone can be related to the strength of selection against it in hybrids relative to dispersal (Barton 2001). In this way, genomic regions causing incompatibilities in hybrids can be mapped as those that have higher levels of differentiation between species (Hagen and Scriber 1989; Rieseberg et al. 1999; Payseur and Nachman 2005; Grahame et al. 2006). Though indirect, this approach is amenable to a wide range of species and presumably has the potential to map a broader range of factors contributing to reproductive isolation. Similar to more direct genetic mapping approaches, statistical analyses of hybrid zones suggest that a large number of loci contribute to reproductive isolation (Barton and Gale 1993). For example, Rieseberg et al. (1999) showed that of 26 genome segments showing significantly reduced introgression across a sunflower hybrid zone, 16 were associated with pollen sterility. These results demonstrate the use of hybrid zones in elucidating the genetic architecture of reproductive barriers between species.

Although both approaches above have proven useful, they suffer from caveats that limit how informative they are about the genetics of speciation. One concern is that hybrid zones are probably not stable over long periods of time. Thus, there may be a large historical component to patterns of differentiation between species, complicating estimates of the strength of selection across a cline. The second is that genes currently contributing to reproductive isolation may not have been involved in the initial speciation process (Coyne and Orr 2004). Whereas the evolution of reproduction isolation is a gradual process, the number of loci required to confer almost complete reproductive isolation between species may be small. One example is Presgraves' (2003) estimate that about 200 genes contribute to hybrid inviability alone in *D. melanogaster*/*D. simulans* hybrids. This implies that the total number of loci contribution to reproductive isolation between species (including, among other things, hybrid sterility, ecological differences, premating isolation, etc.) is likely to be much larger. Orr (1995) showed that a rapid accumulation of incompatibility factors (the "snowball effect") is expected after reproductive isolation is complete between two populations. These theoretical considerations imply that a randomly chosen gene that is currently involved in reproductive isolation between completely isolated species is unlikely to have participated in the speciation process itself. The problem thus becomes trying to distinguish between true "speciation" genes from genic incompatibilities that secondarily strengthen reproductive isolation.

A third, complementary approach is based on coalescent theory. A coalescent approach considers the genealogical properties of samples from parental populations, and can be used to infer population genetic parameters under explicit models of speciation. The simplest of these models considers an allopatric speciation model with no gene flow (Hudson et al. 1987). In the presence of recombination, migration and selection can produce greater heterogeneity in divergence patterns across the genome than expected under a purely allopatric speciation model (Hudson et al. 1987; Palopoli et al. 1996; Wakeley and Hey 1997; Wang et al. 1997; Wu 2001; Machado et al. 2002; Hey and Nielsen 2004; Hey 2005; Bachtrog et al. 2006; Bull et al. 2006). In particular, those parts of the genome that move more freely between species (neutral markers) are expected to diverge more slowly than regions tightly linked to a gene causing reproductive isolation.

By partitioning the genome based on estimated population genetic parameters, such as ancestral and current population size, divergence time and migration rate, we can begin to ask which regions of the genome began to diverge first and/or have the lowest historical migration rates. These regions are more likely to be tightly linked to genes initially causing reproductive isolation rather than neutral parts of the genome or parts of the genome that only recently became associated with reproductive isolation. In addition, we can use estimated population genetic parameters to test explicit models of speciation. Although allopatry is believed to be the dominant mechanism of speciation (Mayr 1942), discordant gene genealogies and divergence time estimates among closely related *Drosophila* species (i.e., *D. pseudoobscura* and relatives; Wang et al. 1997; Machado et al. 2002) and between humans and chimps (Osada and Wu 2005) have rejected models of strict allopatric speciation.

Several Lepidopteran species reveal differential patterns of introgression at multiple loci among ecologically distinct strains or closely related species using various approaches (Lushai et al. 2003; Emelianov et al. 2004; Prowell, et al. 2004; Dopman et al. 2005; Bull et al. 2006; Kronforst et al. 2006). For example, a comparison of three Z-chromosome markers in strains of European corn borer reveals that at one marker haplotypes are not shared (Dopman et al. 2005). This marker is tightly linked to a factor that differentially affects postdiapause developmental time and may contribute to reproductive isolation between strains. Here we develop a new coalescent-based approach and apply it to parental populations of two hybridizing, parapatric species

of Lepidoptera. *Papilio glaucus* and *P. canadensis* are partially reproductively isolated swallowtail butterfly species that form hybrids in a narrow hybrid zone. These species are differentiated by diapause regulation, female-limited mimicry, host-plant preferences, morphological characters, and at least two loci contributing to hybrid inviability (Hagen and Scriber 1989; Hagen et al. 1991; Scriber et al. 1991). Previous surveys of allozymes and mitochondrial DNA (mtDNA) haplotypes revealed a remarkable pattern of differentiation between these two species (Hagen and Scriber 1989; Hagen 1990; Sperling 1993; Bossart and Scriber 1995). In particular, of 21 autosomal allozymes surveyed, most were polymorphic but showed little differentiation between species suggesting high levels of gene flow (Hagen and Scriber 1989). In contrast, mtDNA and three allozymes, including two on the Z chromosome (the Lepidopteran analog of the X in the XY male/XX female system), exhibit strong patterns of differentiation between species, consistent with selection against these markers in hybrids. The two Z-linked allozymes (*Pgd* and *Ldh*) are only loosely linked to each other and show distinct patterns of differentiation across the hybrid zone (Hagen 1990; Fig. 1). These patterns strongly suggest that the genomes of these species are a mosaic of regions that experience differential selection pressures in hybrids, and thus they may also show heterogeneous patterns of differentiation.

We examine patterns of divergence between samples of the parental species for five distinct regions of the Z chromosome and the mtDNA (*COI/COII*). One of the Z-linked regions, *Ldh*, is an allozyme locus that shows particularly strong differentiation between species in transects through the hybrid zone and is thus a candidate for tight linkage to a gene causing reproductive isolation. MtDNA haplotypes also show strong differentiation in transects through the hybrid zone, which may be a consequence of its expected linkage to the W chromosome in Lepidoptera (Andolfatto et al. 2003). This is interesting because female-limited mimicry in *P. glaucus* (a trait that distinguishes species) is partly determined by a W-linked locus (Clarke and Sheppard 1962; Scriber et al. 1996). Here we implement a novel approximate Bayesian approach to estimating speciation time that extends previous approaches (Hudson et al. 1987; Wakeley and Hey 1997; Bachtrog et al. 2006). We use these divergence time estimates to test the strictly allopatric model of speciation, which predicts that each genomic region began to diverge at the same time. Under a model of continuing migration, selection against hybrids, and recombination, we may expect to reject the purely allopatric model. We also relax the strictly allopatric model and estimate locus-specific divergence times. In particular, we expect that our candidate regions, *Ldh* and the mtDNA, should yield deeper divergence time estimates than randomly selected markers. We test this prediction and discuss the implications of our results for mapping speciation genes.

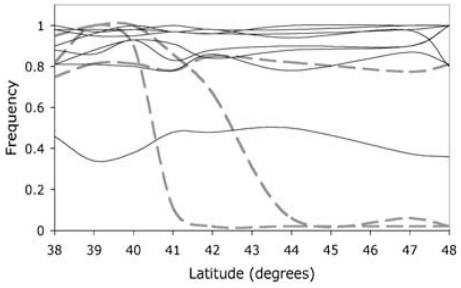


Figure 1. **Frequencies of three Z-linked, *Ldh*, *Pgd*, and *Acp* (grey dashed lines) and eight autosomal (black lines) allozymes across the *P. glaucus/P. canadensis* hybrid zone that corresponds to 41–43°. The y-axis plots the frequency of *P. glaucus*-like allozyme variants. Data replotted from Hagen (1990).**

Materials and Methods

DNA EXTRACTION AND SEQUENCING

Papilio glaucus and *P. canadensis* were collected from a broad geographic sample of their ranges (see online Supplementary Material, Table S1) and identified by hind-wing size and bandwidth. Genomic DNA was isolated using a modified Puregene (Gentra Systems, Minneapolis, MN) protocol. Two legs and the thorax of frozen individuals were ground in 5 μL of 20 mg/mL proteinase K. The homogenate was incubated at 55°C for 12 h, followed by a 2-min incubation at 95°C. The standard Purgene protocol for *Drosophila* DNA purification was followed from this point onward using three times the suggested solution volumes.

A λFIXII genomic DNA genomic library of *P. glaucus* (P. Andolfatto, unpubl. data) was screened for *Lactose Dehydrogenase* (*Ldh*) and *Kettin* (*Ket*) using *D. melanogaster* derived probes and standard library screening protocols (Sambrook and Russell 2001). Positive clones were isolated, sequenced, and intron/exon boundaries were mapped using rapid amplification of cDNA-PCR from total RNA extractions. Primers for *Per* were designed using degenerate primers reported in Regier et. al. (1998). Degenerate primers for *Tpi* were designed using multispecies protein sequence alignments (Logsdon et al. 1995). New primers used in this study are listed in Supplementary Material, Table S2. Primers for *COI/COII* were previously reported in Andolfatto et al. (2003). PCR conditions in a thermocycler (Bio-Rad, Hercules, CA) included an initial denaturing step at 95°C for 2 min followed by 40 cycles of 95°C for 30 sec, 52 °C for 45 sec, 72°C for 2 min with a final 5 min at 75°C.

The PCR product clean-up was performed using Exo/SAP reagents (Fermentas, Hanover, MD). Templates were directly

sequenced on both strands using primers listed in the Supplementary Materials (Table S2) and the BigDye sequencing kit (ver. 3.1, Roche, Nutley, NJ). Sequence reactions were run on an ABI 3730 sequencer (Applied Biosystems, Foster City, CA). Sequencing each gene in a panel of male and female *P. glaucus* confirmed Z-linkage, as chromatograms showed heterozygous sites in all males, and never in females (data not shown). Nucleotide sequences were edited using Sequencher 4.1 software (Gene Codes, Ann Arbor, MI), aligned using ClustalX (Thompson et al. 1997) and manually adjusted (GenBank accession EF115340–EF115363, and EF126370–EF126497).

LEVELS OF POLYMORPHISM, DIVERGENCE, AND RECOMBINATION

To quantify and characterize polymorphism and divergence at these Z-linked loci, we considered all silent sites (all synonymous and noncoding sites) in exons and introns. Sites overlapping insertions and/or deletions were excluded. The tRNA separating *COI* and *COII* in the mtDNA was excluded, as were GT/AG splice sites associated with exon/intron boundaries in the case of nuclear genes. Synonymous and nonsynonymous sites were characterized using DnaSP version 3 (Rozas and Rozas 1999).

Levels of intragenic recombination are an important population genetic parameter, particularly in the context of testing population genetic models (Hudson 1983; Wang et al. 1997; Przeworski et al. 2001). We thus jointly estimated the population mutation rate ($\theta = 4N_e\mu$, where N_e is the effective population size and μ is the mutation rate per base pair) and levels of intragenic recombination, ρ ($= 4N_e r$, where r is the recombination rate per base pair) in *P. glaucus* and *P. canadensis* using an approximate Bayesian method with rejection sampling (Haddrill et al. 2005; Thornton, unpubl. data). Posterior distributions are based on 5000 acceptances of ρ and θ and were estimated for each individual locus and jointly over all loci. Wide uniform priors were chosen for ρ (0, 0.9) and θ (0.003, 0.02). A fixed tolerance (ε) was set to 0.1 for both parameters to reduce computational demand. The analysis took three weeks on two G5 processors.

Levels of silent nucleotide variability within species were summarized using Watterson's estimator, θ_W (Watterson 1975), and the average pairwise diversity per nucleotide, π (Tajima 1983). Divergence (D_{XY}) between *P. glaucus* and *P. canadensis* was estimated as the average pairwise number of nucleotide substitutions per site between species (Nei 1987). The linkage relationships of our Z-linked markers relative to each other are not known, but throughout this paper we assume that they are sufficiently loosely linked such that they can be treated independently. We performed an exact test for linkage disequilibrium between Z-linked loci as implemented in Arlequin version 2.0 (Schneider et al. 2000). No significant levels of linkage disequilibrium among loci were detected in either *P. glaucus* or *P. canadensis* ($P > 0.05$).

NEUTRALITY TESTS

We employ a series of common tests of the neutral equilibrium model in two contexts. First, as the methods employed to estimate divergence times assume neutrality, we used these tests to assess whether there are signs of direct or linked selection acting on the loci used. In a second context, we ask whether our candidates for linkage to hybrid incompatibility loci in the hybrid zone, *Ldh* and the mtDNA, show any signs of recent or ongoing selection, and look any different than other loci we surveyed. In particular, we used two summaries of the distribution of polymorphism frequencies: Tajima's D (Tajima 1989), a measure of the standardized difference between π and θ_W, and Fay and Wu's H (Fay and Wu 2000), which measures the difference between π and θ_H, a summary of θ that weights derived variants by the square of their frequencies. For Fay and Wu's H, we used sequences from *P. rutulus* and *P. multicaudatus* to infer the ancestral state of each nucleotide using standard parsimony criteria with a correction for multiple hits. Under the standard neutral model, both tests are expected to give values close to zero. We also performed three multilocus tests of neutrality as implemented by Haddrill et al. (2005). These tests compared the average Tajima's D and Fay and Wu's H across loci to simulated distributions, and the Hudson, Kreitman, Aguadé (HKA) test, which compares levels of polymorphism (θ_W) to levels of interspecific divergence (D_{XY}) across loci. Analysis of nucleotide variation and tests of neutrality were implemented using programs available at www.biology.ucsd.edu/labs/andolfatto/programs.html.

ESTIMATING DIVERGENCE TIMES

Divergence time (T) is estimated using a novel approximate Bayesian inference developed from previous methods that use the HKA test (Hudson et al. 1987) as a framework (Wakeley and Hey 1997; Bachtrog et al. 2006). Here we develop a Bayesian extension of the likelihood method of Bachtrog et al. (2006), that incorporates information on the number of shared and fixed polymorphisms. We assume a model of simple allopatric speciation where an ancestral population of size θ_A splits into two equal-sized populations of size θ with no gene flow (Fig. 2), and θ and θ_A remain constant over time. For each locus, j, θ is estimated as

$$\theta_j = S_j / \sum_{i=1}^{n-1} \frac{1}{i}, \qquad (1)$$

where n is the sample size and S is the number of segregating sites, Using the same θ for *P. glaucus* and *P. canadensis* is justified because their estimates of θ are not significantly different (results

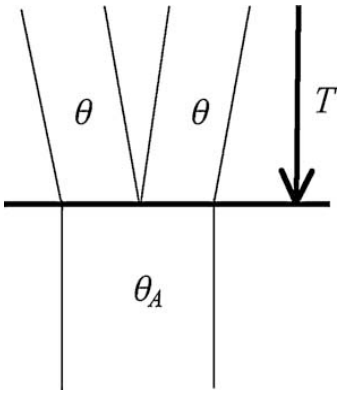


Figure 2. Model of allopatric species divergence. An ancestral population of size θ_A splits into two species of size θ at time *T* with no migration.

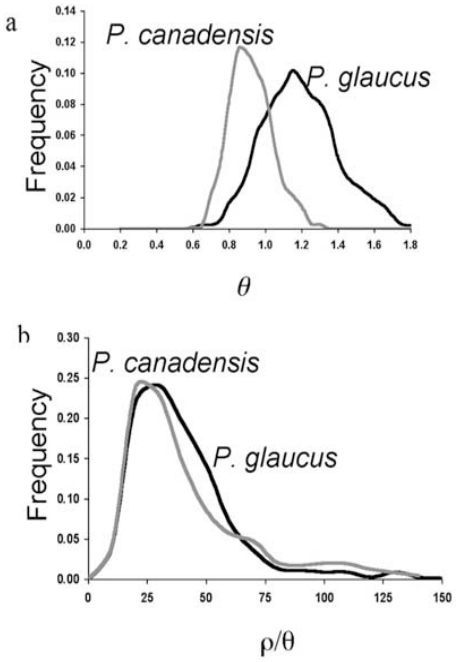


Figure 3. Approximate Bayesian joint posterior distribution of (a) θ and (b) ρ/θ for the five Z-linked markers in *Papillio glaucus* (black) and *P. canadensis* (grey).

not shown); however, our method can be modified to accommodate different populations sizes in the two species (see Bachtrog et al. 2006).

For each locus we perform the following set of steps:

We summarize the observed (obs) data by the sample sizes, locus length in base pairs (representing the total alignment length), levels of variability estimated from the number of segregating sites within a species (θ_j), the number of shared polymorphisms ($S_{j,obs}$), the number of fixed differences between species ($F_{j,obs}$), and the divergence time estimated as

$$T_{j,obs} = (D_{XY,j}/\theta_j) - 1, \quad (2)$$

where θ_j is the average of θ estimates for *P. glaucus* and *P. canadensis* (Hudson et al. 1987).

Using *ms* (Hudson 2002), simulate the neutral coalescence with recombination of samples drawn from two subdivided populations with no gene flow that diverged at time, T . This coalescence method is based on the standard Fisher-Wright neutral model, which assumes a large, panmictic population and mutations occurring according to the infinites sites model (Hudson 1983). An infinite sites model is a reasonable assumption in our case (because levels of variability and divergence are low), but may not be appropriate for highly diverged species or genomes with high mutation rates (e.g., some viruses). In these simulations, we use the point estimate of θ_j for θ and assume $\rho = 33.5\theta$, which represents the joint maximum a posteriori (MAP) estimate of ρ in *P. glaucus* (Fig. 3). The simulated divergence time, *T,* is the only free parameter. We use an uninformative (i.e., uniform) prior for *T* that is sampled from the interval 0-16 N_e generations.

We summarize the simulated data (sim) in the same way we summarized the observed data. We accept the simulated value of *T* if $S_{j,sim} = S_{j,obs}$, $F_{j,sim} = F_{j,obs}$, and $|T_{j,sim} - T_{j,obs}| < \delta$, where δ is a fixed tolerance. A drawback of rejection sampling is that the tolerance parameter affects the efficiency of inference and as a result, acceptance rates may be prohibitively low if a very stringent δ is used (Beaumont et al. 2002). Owing to computational constraints, δ was set to 0.05, and the average acceptance rate was $\sim 10^{-3}$ per locus. The simulations took approximately three weeks on four G5 processors. Using a lower tolerance (0.001) had little effect on the posterior distribution (results not shown). To calculate a joint, multilocus estimates of *T*, the same priors and tolerance must be used for each locus.

We repeated steps 2 and 3 until 2000 draws of the posterior distribution were collected.

The method produces posterior distributions of *T* for each locus. We summarized the posterior distributions and obtained the MAP estimate and 95% confidence interval as implemented in the "locfit" statistical package (Loader 2006) in the library for *R*. To obtain a joint Z-linked MAP estimate of *T*, posterior distributions for each locus were binned in increments of 0.2 N_e

and probabilities for each bin were multiplied across loci. Likelihood ratio tests were used to test whether assuming unique divergence times for each of the Z-linked markers fit the data significantly better than one (i.e., strictly allopatric) divergence time. We assume the likelihood ratio statistic is chi-squared distributed with degrees of freedom equal to the difference in the number of free parameters. To examine the effect of the assumed ancestral population size (θ_A) on divergence time estimates, we carried out simulations with an ancestral population size that was 1, 2, 5, 8, and 10 times the current population size (θ). In this implementation, we assume that the change in population size is instantaneous, though this assumption can be relaxed (see the *ms* program documentation).

We tested the performance of our method on 100 datasets simulated under the parameters we estimated from the data (i.e., sample size, and our estimates of θ, ρ and T), and assuming that $\theta_A = \theta$. The method works well under parameters that closely match our data, with a bias of 0.2 N_e and root mean square error of 4.8 N_e (Supplementary Material, Fig. S1). A library of scripts and programs to implement this procedure, called STE (Speciation Time Estimator), is available from the website www.biology.ucsd.edu/labs/andolfatto/programs.html.

We compared our divergence time estimates among Z-linked markers to results from an alternative method (*WH*) developed by Wakeley and Hey (1997). This program uses the number of exclusive and shared polymorphisms, and the number of fixed differences in the observed data to estimate the population sizes of *P. glaucus* and *P. canadensis* (θ_1 and θ_2, respectively), the size of the ancestral population (θ_A), and the divergence time estimate (T) for each locus and jointly across all Z-linked loci. Point estimates of these parameters were then used in neutral coalescent simulations with recombination to test the fit of the data to a strict allopatric speciation model (Wang et al. 1997; Kliman et al. 2000). Whereas both approaches are multilocus methods that use coalescence with recombination to test the fit of observed data to an allopatric speciation model, *WH* differs from our method in that it has an additional free parameter (θ_A), estimates θ differently, and is a moment-based method.

EVALUATING MODELS USING THE POLYMORPHISM FREQUENCY SPECTRUM

We used two summaries of the frequency spectrum (Tajima's D and Fay and Wu's H) to evaluate the fit of parameters estimated both under our method and the *WH* method to the observed data. We summarized the observed data as the average D and H across the five Z-linked loci. For each locus, we used *ms* to simulate 10,000 neutral genealogies with the following parameters: θ_j, the joint Z-linked mode for recombination ($\rho = 33\theta$), and T, drawn from the posterior distribution of T obtained from our approximate Bayesian analysis and varied θ_A to be 1×, 5×, 8×, and 10× the current population size θ. For each simulated replicate, we recorded the average Tajima's D and average Fay and Wu's H across loci and compared these distributions to the observed averages. Similarly, we evaluated estimates from the *WH* approach by simulating data using point estimates of θ_1, θ_2, θ_A, T and assumed that $\rho = 33\theta$.

Results

LEVELS OF DIVERSITY AND RECOMBINATION

Levels of silent variability in *P. glaucus* and *P. canadensis* (∼1%, see Table 1 and Supplementary Material, Fig. S2) are comparable to *Drosophila* (Moriyama and Powell 1997) and other Lepidoptera surveyed so far (Beltran et al. 2002; Dopman et al. 2005). *Papilio canadensis* has slightly lower levels of variability (0.9%, 95% CI 0.7–1.2) on average than *P. glaucus* (1.1%, 95% CI 0.8–1.7); however, three of the five Z-linked markers are actually more variable in *P. canadensis*. Although suggestive of a smaller effective population size in *P. canadensis*, the lack of a systematic trend implies that we cannot reject the hypothesis that the two species have equal effective population sizes. Levels of diversity on the Z chromosome and mtDNA are similar, but levels of divergence are three-fold higher on the mtDNA (average D_{XY} is 2.3% for Z-linked and 6.3% for mtDNA). The smaller ratio of polymorphism to divergence for the mtDNA relative to nuclear genes is expected in a neutral equilibrium population with equal numbers of males and females (Kimura 1983; Birky et al. 1989). The higher levels of silent divergence on the mtDNA relative to nuclear genes is typical of arthropods (Moriyama and Powell 1997).

The population recombination rate, ρ, is inversely proportional to levels of intragenic linkage disequilibrium. We estimate ρ per site to be lower (but not significantly so) in *P. canadensis* (mode 0.15, 95% CI 0.06–1.5) than in *P. glaucus* (mode 0.35, 95% CI 0.11–1.0). This difference may in part be attributed to a smaller population size, as reflected by lower levels of diversity on average in *P. canadensis* (Fig. 2A). The ratio of the recombination rate, ρ, relative to the mutation rate, θ, however, is expected to be similar in two equilibrium populations of different size (Hudson et al. 1987; Andolfatto and Przeworski 2000). Interestingly, the mode of ρ/θ in *P. glaucus* (33.5, 95% CI 9.0–90.7) that is very close to that for *P. canadensis* (33.0, 95% CI 9.1–110.2; see Fig. 2B). This estimate is considerably higher than recent estimates from *D. melanogaster* ($\rho/\theta = 10$; Thornton and Andolfatto 2006), which is consistent with *Papilio* having at least a three-fold longer genetic map (Ashburner 1989; Winter and Porter, pers. comm.) but roughly only two times more genomic DNA (Celniker and Rubin 2003; Gregory and Herbert 2003).

Per locus posterior distributions of ρ/θ are broad (Supplementary Material, Fig. S3) and MAP estimates of ρ/θ appear to

Table 1. Polymorphism and divergence statistics for *Papilio glaucus* and *P. canadensis*

Gene	Species	Sample size	Total length	Silent sites	S[1]	π[2] (%)	θ[3] (%)	D_{XY}[4] (%)	Sh[5]	F[6]	E(ss)[7]
Kettin	*P. glaucus*	12	1206	251	4	.4	.5	3.8	0	7	.08
	P. canadensis	10			5	.6	.7				
Ldh	*P. glaucus*	11	439	317	16	1.9	1.7	2.9	0	3	.25
	P. Canadensis	9			5	.7	.6				
Period	*P. glaucus*	12	212	162	3	.5	.6	2.5	0	3	.56
	P. canadensis	7			3	.7	.8				
Titin	*P. glaucus*	12	611	410	13	1.2	1.1	1.7	10	1	.41
	P. canadensis	8			13	1.4	1.2				
Tpi	*P. glaucus*	12	296	211	15	2.4	2.4	4.4	3	0	.64
	P. canadensis	10			9	1.2	1.3				
Z-linked	*P. glaucus*	12[8]	2764	1351	51	1.3	1.3	2.8[8]	13	14	1.4
	P. canadensis	9[8]			36	.9	.9				
COI/COII	*P. glaucus*	29	2289	494	24	.7	1.3	6.3	1	19	.49
	P. canadensis	13			10	.3	.6				

[1] Total number of polymorphisms observed.
[2] Average pairwise diversity per site.
[3] Estimate of $\theta = 3N_e\mu$ ($N_e\mu$ for mtDNA) per site using the number of polymorphic sites.
[4] Average pairwise divergence per silent site.
[5] The number of shared polymorphisms at silent sites.
[6] The number of fixed differences at silent sites.
[7] The expected number of shared mutations from recurrent mutation (Clark 1997; Kliman et al 2000).
[8] Averages.

vary somewhat among loci. We implemented a likelihood ratio test to evaluate whether this reflected significant heterogeneity in recombination rates among loci. In particular, we tested whether assuming one ρ/θ for all loci is a better fit to the data than assuming five different ρ/θ (i.e., one for each locus). Using this approach, we found no significant evidence for recombination rate heterogeneity in either species (*P. glaucus*: $P = 0.66$, *P. canadensis*: $P = 0.21$). Given lack of evidence for heterogeneity in recombination rates and the worry that per locus MAP estimates may be biased due to the sample size and number of segregating sites (Andolfatto and Wall 2003), we prefer to use the joint multilocus estimate of ρ/θ in lieu of locus-specific estimates.

PATTERNS OF VARIATION AT CANDIDATE LOCI *Ldh* AND mtDNA

Transects through the hybrid zone between *P. glaucus* and *P. canadensis* have revealed that *Ldh* harbors a single allozyme variant that shows strong clinal differentiation between species (Hagen 1990). Because most polymorphic allozymes show little differentiation, our working hypothesis is that *Ldh* is tightly linked to a factor causing reproductive isolation between the species. Our sequenced *Ldh* clone is 6.4 kilobases long and contains five exons, which comprises most of the *Ldh* protein (we are missing 40 amino acids on the 5′ end). We sequenced all exons in both species and found three nonsynonymous differences in the fourth exon of our clone (Fig. 4). Two of the mutations are conservative amino acid changes. One mutation is from a serine to threonine substitution that is only polymorphic in one *P. canadensis*, and the other is a glycine to valine substitution that is polymorphic in *P. glaucus*. The third nonsynonymous mutation is a nonconservative change from an uncharged glutamine in *P. glaucus* to a negatively charged lysine in *P. canadensis* that appears to be fixed between species. This likely represents the electrophoretic difference underlying the *Ldh* allozyme variants that distinguish these species. Using *P. rutulus* and *P. multicaudatus* as outgroup species, the fixed Gln to Lys change appears to be derived in the *P. canadensis* lineage. We surveyed a 495 base pair region that included this exon and 306 base pairs of the upstream intron in population samples from both species. Surprisingly, despite being a candidate for selection in the hybrid zone, *Ldh* does not stand out as unusual when compared to other loci. Levels of variability, Tajima's *D*, and Fay and Wu's *H* for *Ldh* in both species were close to the average across Z-linked markers, and did not significantly differ from neutral expectations (Supplementary Material Table S3). A possible explanation for this pattern is that if selection is weak compared to levels of recombination in *Ldh* a signature of historical divergent selection at linked sites might be obscured.

The mitochondrion is another candidate target of selection opposing introgression owing to its expected linkage to the W chromosome. Interestingly, the mtDNA marker, *COI/COII*, is the only marker surveyed in this study that significantly departs from

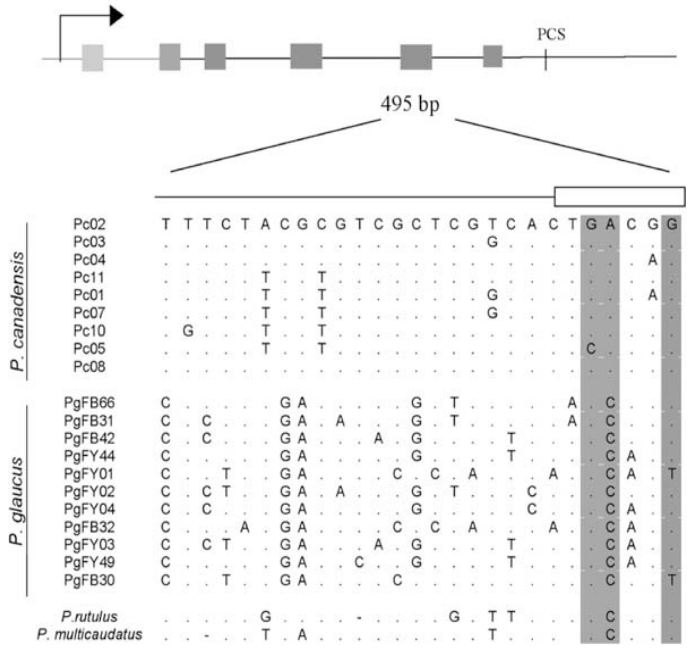


Figure 4. Diagram of the *Ldh* gene and nucleotide variants in *Papillio glaucus* and *P. canadensis*. In the gene diagram the grey box represents an exon that was not sequenced in this study. Shaded nucleotides indicate nonsynonymous substitutions. Outgroup sequences of *P. rutulus* and *P. multicaudatus* are included. Dots indicate identity to the reference sequence; hyphens indicate a gap. The PolyA cleavage site is indicated as PCS.

neutral expectations (Supplementary Material Table S3). A significantly negative Tajima's D ($P = 0.02$, without correcting for multiple tests) and somewhat reduced polymorphism in *P. canadensis* is suggestive of a possible selective sweep involving the mitochondrion or the W chromosome. Additional markers will be necessary to distinguish between demographic and selective causes for this departure from the neutral equilibrium model.

MULTILOCUS PATTERNS OF DIFFERENTIATION

Shared polymorphisms can result from the persistence of ancestral polymorphism, introgression, or recurrent mutation. The distribution of shared and fixed, and exclusive polymorphisms between the species varies considerably among loci (Table 1). Three of the five Z-linked loci exhibited fixed differences between species and no shared polymorphisms. *Titin* has 10 shared polymorphisms (of 16 total) with one fixed difference and *Tpi*, has three shared polymorphisms (of 21 total) and no fixed differences. The mtDNA marker, *COI/COII*, contained 19 fixed differences and O shared polymorphism (of 33 total).

The source of shared polymorphisms of nuclear genes and the mtDNA marker may have different causes. The probability of shared polymorphisms by parallel mutation was determined following the expectation of a hypergeometric distribution in the two species (Clark 1997). In *P. glaucus* and *P. canadensis* 13 of the 74 Z-linked polymorphisms are shared but only 1.4 are expected to be due to recurrent mutation ($P < 0.001$, Table 1). This rules out recurrent mutation as the cause of shared polymorphism and leaves ancestral polymorphism or introgression as possible causes. However, recurrent mutation cannot be ruled out as the cause of the one shared of 33 polymorphisms observed at *COI/COII* ($P = 0.31$). This shared difference was excluded from our estimation of divergence time for *COI/COII* (see below).

DIVERGENCE TIME ESTIMATES

To estimate the time that *P. glaucus* and *P. canadensis* began to diverge, T, we implemented a novel approximate Bayesian method that is an extension of previous approaches (Hudson et al. 1987; Wakeley and Hey 1997; Bachtrog et al. 2006). These estimates are based on average pairwise divergence between species, D_{XY}, and within species polymorphism, θ, and the number of shared and fixed differences observed in samples from both species. Posterior distributions of T for each locus, and a joint posterior distri-

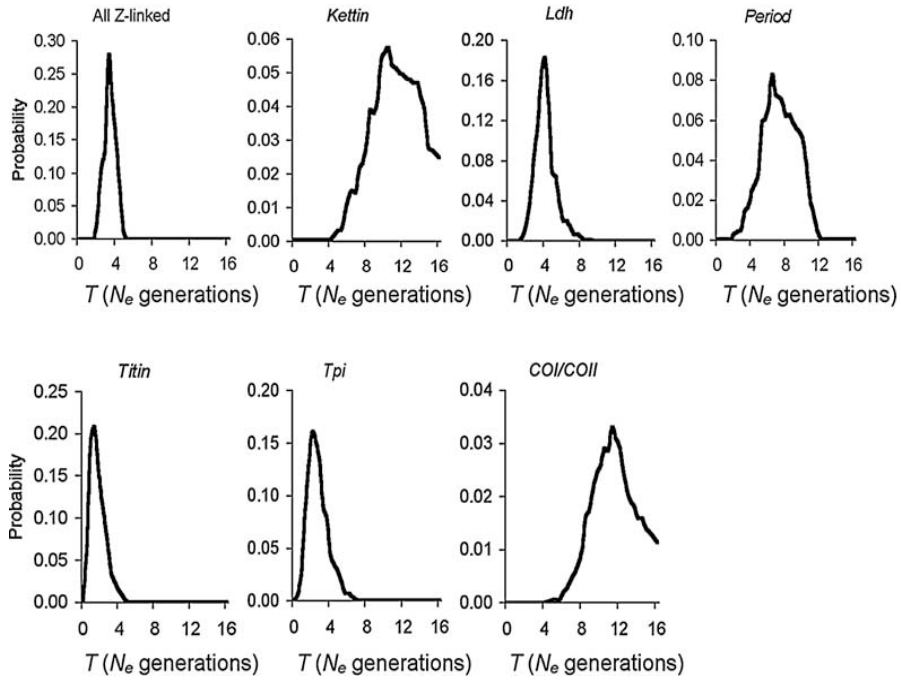


Figure 5. Posterior distributions of T (in units of N_e generations) for *Papilio glaucus* and *P. canadensis*, assuming $\theta_A = \theta$.

bution for all Z-linked loci are shown in Figure 5. The joint maximum a posteriori (MAP) estimate for the Z-linked loci is 3.2 N_e generations ago. The mutation rate in *Papilio* is not known, and the generation time of *P. glaucus* and *P. canadensis* are likely to differ (Hagen et al. 1991). However, if we crudely assume that the mutation rate per generation is similar to that of *Drosophila* (i.e., 1.5×10^{-8} per year; Li 1997) and assume one generation per year, this implies that these species began to diverge around 0.6 million years ago.

In this study, we were interested in comparing divergence time estimates for two candidate loci—*Ldh* and the mitochondrion—to other Z-linked loci, with the expectation that these loci should yield deeper divergence times due to their putative linkage to reproductive isolation factors. Surprisingly, the divergence time estimate for *Ldh* is close to the joint estimate based on all Z-linked loci. The divergence time estimate for the mtDNA (11.7 N_e generations ago) is difficult to compare to nuclear genes because the appropriate coalescent scaling factor relative to the Z chromosome is not known. However, if we assume a neutral equilibrium population with equal numbers of males and females, the appropriate scaling factor would be 3, suggesting that the scaled divergence time for the mtDNA (3.9 N_e generations ago) also agrees well with the joint estimate of T for the Z chromosome. Thus there is no evidence that these candidate loci have unusual patterns of divergence compared to other Z-linked loci.

TESTING THE ALLOPATRIC SPECIATION MODEL

We address the validity of the purely allopatric speciation model using only Z-linked genes. MAP estimates of T varied considerably among Z-linked loci ranging from 1.2 (*Titin*) to 10.5 (*Kettin*) N_e generations ago. A likelihood ratio test was implemented to ask whether a model positing unique divergence times for each locus fit the data significantly better than a single divergence time equal to the joint MAP estimate. We exclude the mtDNA because of uncertainty about appropriate scaling factor and possible evidence for selection (Supplementary Material Table S3). Using this test, we can reject the simple allopatric speciation model depicted in Figure 2 ($P < 0.001$ assuming $\theta_A = \theta$; Table 2). Because θ_A is not a free parameter in our approach, we performed our test of allopatric model assuming θ_A is $2\times$, $5\times$, $8\times$, and $10\times$ θ. We find that allopatry can be rejected when the assumed size of the ancestral population was 2, 5, and 8 times the current popula-

Table 2. Divergence time, T, in units of N_e, estimated under a simple speciation model with no migration. T is estimated from the raw data (using $D_{XY}/\theta - 1$) when $N_A = 1N_e$, and as the MAP estimate when the ancestral population size is increased 2, 5, 8, and 10 times.

Gene	Estimated divergence time (T)				
	$N_A = 1N_e$	$N_A = 2N_e$	$N_A = 5N_e$	$N_A = 8N_e$	$N_A = 10N_e$
Kettin	10.5 (5.9–16.0)[1]	10.4 (5.6–15.2)	5.7 (3.0–8.0)	4.8 (2.0–6.9)	4.0 (1.6–6.6)
Ldh	4.0 (1.8–6.8)	3.6 (1.6–6.4)	2.1 (1.4–4.2)	2.0 (1.4–4.1)	1.6 (1.0–3.4)
Period	6.3 (3.4–11.0)	4.5 (2.8–8.8)	2.8 (1.2–3.6)	2.8 (1.2–3.6)	2.8 (1.2–3.4)
Titin	1.2 (.4–3.6)	.9 (.4–2.4)	.8 (.4–2.0)	.6 (.2–1.5)	.4 (.1–1.2)
Tpi	2.8 (.7–5.1)	2.4 (.6–4.9)	1.6 (.6–4.2)	1.6 (.5–4.0)	1.6 (.4–4.0)
All Z-linked	3.2 (1.8–4.2)	2.8 (1.3–3.6)	2.4 (1.1–2.2)	2.4 (1.0–2.2)	2.4 (1.0–2.0)
mt_COI/COII[2]	3.9 (1.3–5.3)	3.6 (1.2–4.6)	1.9 (1.1–3.6)	1.6 (1.0–3.5)	1.4 (.6–3.4)

[1] Confidence interval of T estimated as two log-likelihood units around the maximum.
[2] Divergence time is scaled to 1/3 N_e of Z-linked markers.

tions size ($P = 0.001$, $P = 0.007$, $P = 0.02$, respectively) but not 10 times ($P = 0.09$). These results suggest that shared ancestral polymorphism cannot account for the variation in divergence time estimates across loci unless the ancestral population was 10 times the size of the current population size.

We compared the results of our approximate Bayesian method to the *WH* method of Wakeley and Hey (1997). Estimated parameter values from the *WH* analysis were $\theta_1 = 6.77$, $\theta_2 = 5.48$, $\theta_A = 49.05$, and $T = 0.7\ N_e$, and this approach failed to reject the allopatric speciation model using the *WH* test statistic ($P = 0.5$). In the *WH* method, θ_1,θ_2, and θ_A are free parameters, and θ_A is estimated to be about eight times larger than θ_1 and θ_2. Based on coalescent simulations with recombination, we found the estimates of θ_1 and θ_2 are not significantly different (results not shown). However, using our approximate Bayesian approach we reject allopatry with a likelihood ratio test when $\theta_A = 8x\theta$ (see above; $P = 0.02$). It is unclear whether it is the *WH* statistic or the more recent divergence time inferred by the *WH* method compared to our approximate Bayesian approach that explains the discrepancy between the two methods.

A large variance among loci in the number of shared and fixed polymorphisms (and thus estimates of divergence time in models with no migration) is expected when the ancestral population size is large compared to the current population size (Wakeley and Hey 1997). Both tests of the allopatric model suggest that the data can be reconciled with an allopatric model only if that size of the ancestral population was large relative to the current size (i.e., 10× by our approach and 8× by the *WH* approach). However, drastic changes in population size are expected to leave characteristic signatures in the frequency spectrum of polymorphisms (Tajima 1989; Fay and Wu 1999; Haddrill et al. 2005). We thus used information from the frequency spectrum of polymorphisms (summarized as the averages of Tajima's D and Fay and Wu's H across loci) to evaluate the fit of the data to different assumptions about θ_A. We compared the average Tajima's D and Fay and Wu's H for our observed data to simulations using point estimates of parameters from the *WH* method and the posterior distributions of parameters from our approximate Bayesian approach (Fig. 6). The observed average values for Tajima's D and Fay and Wu's H in *P. glaucus* are significantly different from the distributions of average D and average H obtained under the *WH* parameters ($P = 0.03$). Similarly, the average Fay and Wu's H is significantly different from the observed average for *P. glaucus* using parameters estimated by our approximate Bayesian method assuming that θ_A is 5× and 10× larger than θ_1 ($P = 0.02$). In contrast, both the observed D and H are compatible with the distributions obtained assuming $\theta_A = \theta$ ($P = 0.34$ and $P = 0.10$). These results suggest that the ancestral population size is unlikely to be larger than 5× the current size as suggested by the *WH* method, and allows us to reject the allopatric model for this data. Incorporating information from the frequency spectrum may be a valuable addition to methods for estimating current and ancestral population sizes under speciation models.

Discussion

TESTING ALTERNATIVE MODES OF SPECIATION

Multilocus coalescent methods provide a potentially powerful means to piece together the history of speciation for recently diverged species. The role of continuing gene flow in the speciation process is much debated (reviewed in Coyne and Orr 2004). Two types of models have been developed to address different speciation scenarios. The first are simple isolation models with no gene flow that mimic strictly allopatric speciation (Takahata and Nei 1985; Hudson et al. 1987; Hey 1994; Wakeley and Hey 1997; Wang et al. 1997; Bachtrog et al. 2006). A second class of models allow for continuing migration between populations (Nath and Griffiths 1996; Beerli and Felsenstein 1999; Nielsen and Wakeley

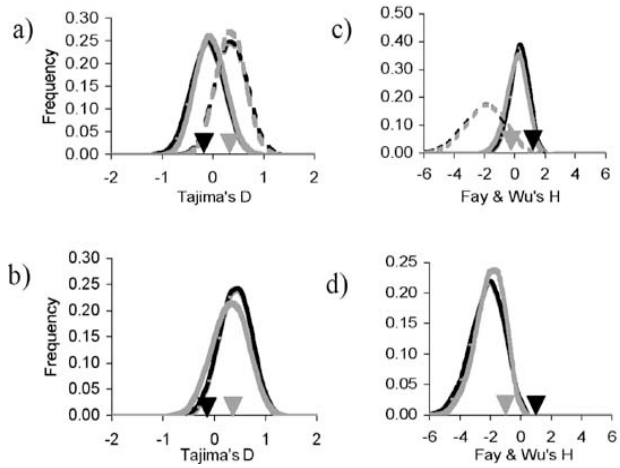


Figure 6. The polymorphism frequency spectrum under alternative allopatric speciation models. For all figures, *Papilio glaucus* is represented in black and *P. canadensis* is represented in gray. Observed averages are indicated by triangles. Each graph plots 10,000 simulated replicates of the average *D* and *H* for five loci. The distributions of Tajima's *D* and Fay and Wu' s *H* in (a) and (c) are based on neutral simulations using parameters estimated under the approximate Bayesian method. Solid lines represent distributions where $\theta_A = \theta$, and dashed lines where $\theta_A = 5\theta$. Distributions shown in (b) and (d) are based on neutral simulations using parameters estimated by the *WH* method (Wakeley and Hey 1997). See Methods for simulation parameters.

2001) as expected under parapatric or sympatric speciation. By assessing the fit of data from multiple unlinked regions of the genome to these models, we can begin to assess the prevalence of one mode of speciation relative to another, and the general importance of continuing gene flow in particular.

Table 3. Goodness-of-fit tests for one versus five divergence times among Z-linked markers.

N_A[1]	Divergence times[2]	ln L	Likelihood-ratio statistic (df = 4)	*P*
1×	1	−19.78	18.36	.001
	5	−10.60		
2×	1	−19.27	18.23	.001
	5	−10.16		
5×	1	−15.01	14.14	.007
	5	−7.93		
8×	1	−13.95	12.15	.016
	5	−7.88		
10×	1	−11.74	8.13	.087
	5	−7.67		

[1]The size of the ancestral population relative to the current population size of *P. glaucus* and *P. canadensis*.

[2]The two models tested posit one divergence time for all loci versus a unique divergence time for each of the five Z-linked loci.

In the absence of recombination and recurrent mutations, genealogies will show either fixed differences or shared polymorphisms, but not both (Wakeley and Hey 1997). The presence of both shared and fixed polymorphisms among loci on the Z chromosome of *P. glaucus* and *P. canadensis* can thus be best explained by different evolutionary histories for different parts of the Z chromosome. Here we have combined a likelihood ratio test of the allopatric model with information from the frequency spectrum to show that a strictly allopatric model can be rejected for these two *Papilio* species (Table 3, Fig. 6). The wide range of divergence time estimates among loci strongly suggests that historical introgression has occurred for some parts of the Z chromosome but not in others. This pattern is consistent with the observation of differential introgression of molecular markers through the hybrid zone between these two species (Hagen 1990). Given our inference of continuing gene flow between species, we conclude that our joint speciation time estimate across loci (3.2 N_e generations ago) is most likely an underestimate of the true time the species began to diverge (Wakeley and Hey 1997; Osada and Wu 2005).

LOCUS-SPECIFIC DIVERGENCE TIME ESTIMATES AND IMPLICATIONS

An additional use of the coalescent-based approach is to test hypotheses about the evolutionary history of specific genes or regions

of the genome. In particular, by treating loci separately, we can assign each with an estimated divergence time. In the case of two hybridizing incipient species, we expect that regions of the genome linked to hybrid incompatibility alleles will begin to diverge as soon as these incompatibilities arise. It is also possible that some regions of the genome began to differentiate between species prior to the evolution of reproductive isolation between these species. Identifying both types of regions can potentially yield information about the genetic basis of speciation.

The accumulation of fixed differences and lack of shared polymorphism at three of five of the Z-linked loci surveyed (*Kettin*, *Ldh*, and *Period*) suggests there has been little introgression at these markers. Of particular interest is *Ldh*, for which we had prior information that alternative allozyme alleles show strong clinal differentiation between species in transects through the hybrid zone, suggesting either selection on *Ldh* itself or a linked character such as diapause, Batesian mimicry, or hybrid inviability factors. We have identified a single nonsynonymous fixed difference between *P. glaucus* and *P. canadensis* that results in a change in amino acid charge (Fig. 4) and may account for the diagnostic alleles observed in a previous allozyme study of transects through the hybrid zone (Hagen 1990). Surprisingly, however, our divergence time estimate for *Ldh* (4.0 N_e generations ago) is intermediate compared to other loci, and very close to the average for Z-linked genes (3.2 N_e generations ago). *Kettin* and *Period*, for which we have no prior hypotheses, have similar patterns of polymorphism to *Ldh* but deeper divergence estimates and thus are more likely to be linked to loci causing reproductive isolation. Such hypotheses could be confirmed by reciprocally introgressing these regions from one species into the other and testing for effects on hybrid inviability, or surveying patterns of introgression through the hybrid zone at these loci relative to the rest of the genome. Such regions are predicted to have a larger effect on hybrid viability than *Titin* and *Tpi*, for example, where many shared polymorphisms and few fixed differences are observed. Thus, coalescent methods may prove to be a useful tool in partitioning the Z chromosome into candidate regions of species-specific functional and/or ecological importance.

The mtDNA is a second candidate locus for selection in hybrids due to its expected linkage to the W chromosome in Lepidoptera (both are maternally transmitted), which in *P. glaucus* and *P. canadensis* carries alternative alleles for a diagnostic mimicry difference between species (Clarke and Sheppard 1962; Scriber et al. 1996). A relatively deep divergence time estimate might be expected for the mitochondrion if mimicry contributes to reproductive isolation between these species. At first glance, the divergence time for the mtDNA does appear to be deeper than for most Z-linked markers. However, given the lower expected effective population size for the mtDNA relative to the Z chromosome, it is probably more appropriate to scale the divergence time by a factor of three. This correction results in a divergence time estimate for the mtDNA that is similar to joint estimate for the Z chromosome (3.9 and 3.2 N_e generations, respectively). It should be noted that possible recombination between the mtDNA and the W chromosome in *Papilio* (Andolfatto et al. 2003) may uncouple their evolutionary histories to some extent, weakening the association between the mtDNA and the W chromosome.

HITCHHIKING EFFECTS AND ESTIMATING *T*

Inference of speciation times from population genetic data assume that the markers used are neutral and are not closely linked to other loci experiencing purifying or positive selection. In general, because speciation times are estimated in units of the effective population size, any form of selection that influences the effective population size (and thus neutral variation) in a genomic region—such as recurrent selective sweeps (Kaplan et al. 1989), background selection (Charlesworth et al. 1993), or balancing selection (Kaplan et al. 1988)—can lead to incorrect time estimates. In particular, selection that reduces variation will lead to overestimates of the divergence time whereas selection that enhances variation will lead to underestimates. In principle, hitchhiking effects could account for the greater than expected variance in estimated divergence times among loci under a strictly allopatric model that has now been described in several species (Wang et al. 1997; Machado et al. 2002; Llopart et al. 2005).

In our study, we can rule out the possibility that recent species-specific selective sweeps in either species contributes to deep divergence times estimated at *Kettin* and *Period* by considering intraspecific estimates of θ (Supplementary Material Fig. S2). At these two markers, levels of variability are very similar in the two species, whereas a recent selective sweep in either species would be apparent as a species-specific reduction in levels of variability. More difficult to rule out is the possibility of reduced variability in both species. In *Drosophila*, regions of reduced crossing-over harbor reduced levels of variability consistent with the effects of recurrent hitchhiking and/or background selection (Begun and Aquadro 1992; Andolfatto 2001). Currently, we cannot rule out the possibility that *Kettin* and *Period* are simply located in regions of reduced recombination on the Z chromosome, and as a result are more prone to linked variation-reducing selection. Within *P. glaucus* and *P. canadensis*, we did not detect significant variation in the recombination rate among loci; however, our test is weak and clearly more loci will be necessary to determine if there is a negative correlation between recombination rates and divergence time estimates. Uncertainty about recombination rates, and the effects of linked selection, are a general problem for using coalescent-based approaches to estimate speciation times in organisms that lack genetic and physical maps. Hey and Nielsen (2004) address

this issue by adding an inheritance scalar to each locus as an additional free parameter in their model (noting that they also assume no recombination within loci). Unfortunately, this approach would be computationally prohibitive using our method.

FUTURE PROSPECTS

Our coalescent-based approach to divergence time estimation provides the means to test models of speciation as well as identify regions of a genome potentially contributing to phenotypic species differences and factors underlying reproductive isolation between them. Here we have employed a simple allopatric model with no migration to show that such a model can be rejected based on the greater than expected variance in divergence time estimates among loci under this model. Although computationally intensive, the framework can be extended to add ancestral and current population sizes and continuing migration as free parameters within the rejection-sampling method. Here we have also shown the frequency spectrum distribution can yield useful information about the relative sizes of current and ancestral populations, and thus may be useful in estimating speciation parameters. A disadvantage of our rejection-sampling method is that it becomes less computationally feasible as the number of free parameters increases. In contrast, Markov Chain Monte Carlo-based approaches (i.e., Nielsen and Wakeley 2001; Hey and Nielsen 2004) do not suffer from this limitation. A clear advantage of our approach over these methods is the inclusion of intragenic recombination, which appears to be very strong relative to mutation in *Papilio*.

Our method can also be used to assign a temporal sequence to the appearance of reproductive isolation factors. Because reproductive isolation is not complete between *P. glaucus* and *P. canadensis*, each reproductive isolation factor that is found can be thought of as actively contributing to the on-going speciation process. Such is not the case for completely reproductively isolated species that may harbor hundreds of reproductive isolation factors, most of which were likely not involved in the speciation process (Coyne and Orr 2004). In principle, our approach could be used to distinguish between true speciation genes (i.e., those genes that were actively involved in restricting gene flow between species during speciation) and secondary reproductive isolation factors that have accumulated long after the speciation process has been completed. It will be of considerable interest to determine if genomic regions underlying locally adapted traits in *P. glaucus* and *P. canadensis*, such as diapause, host-plant preferences, and mimicry differences, show evidence for reduced gene flow between species and higher levels of differentiation than the genomic background.

Our approach, like all methods so far, makes a number of simplifying assumptions that are necessary given computational constraints. Although clearly we can look forward to more advances in coalescent-based methods, these approaches should probably at best be used as guides to identify potentially interesting regions of the genome. Coalescent approaches should be complementary to examining patterns of introgression of candidate regions through the hybrid zone or introgressing these regions from one species to another so we can potentially verify their effect on phenotypic differences and/or reproductive isolation between species.

ACKNOWLEDGMENTS

We thank K. Thornton and J. Huelsenbeck for helpful discussions, and D. Bachtrog for comments on the manuscript. JMS was supported by MAES Project no. 01644 at Michigan State University. PA was supported by an Alfred P. Sloan Research Fellowship in Molecular and Computational Biology.

LITERATURE CITED

Andolfatto, P. 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans.* Mol. Biol. Evol. 18:279–290.

Andolfatto, P., and M. Przeworski. 2000. A genome wide departure from the standard neutral model in natural populations of *Drosophila.* Genetics 156:257–268.

Andolfatto, P., and J. D. Wall. 2003. Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster.* Genetics 165:1289–1305.

Andolfatto, P. J., Scriber, and B. Charlesworth. 2003. No association between mitochondrial DNA haplotypes and a female-limited mimicry phenotype in *Papilio glaucus*. Evolution 57:305–316.

Ashburner, M. 1989. *Drosophila*: a laboratory handbook. Cold Spring Harbor Laboratory Press, New York.

Bachtrog, D., K. Thornton, A. Clark, and P. Andolfatto. 2006. Extensive introgression of mitochondrial DNA in the absence of nuclear gene flow in *Drosophila yakuba* species group. Evolution 60:292–302.

Barbash, D., D. Sinno, A. Tarone, and J. Roote. 2003. A rapidly evolving MYB-related protein causes species isolation in *Drosophila*. Proc. Natl. Acad. Sci. USA 100:5302–5307.

Barton, N., and K. Gale. 1993. Genetic analysis of hybrid zones. Pp. 13–45 *in* R. Harrison, ed. Hybrid zones and the evolutionary process. Oxford Univ. Press, Oxford, U.K.

———. 2001. The role of hybridization in evolution. Mol. Ecol. 10:551–568.

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

Begun, D. J., and C. F. Aquadro. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rate in *D. melanogaster*. Nature 365:519–520.

Beerli, P., and J. Felsenstein. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152:763–773.

Beltran, M., C. Jiggins, V. Bull, M. Linares, J. Mallet, W. McMillan, and E. Bermingham. 2002. Phylogenetic discordance at the species boundary: comparative gene genealogies among rapidly radiating *Heliconius* butterflies. Mol. Biol. Evol. 19:2176–2190.

Birky, C., P. Fuerst, and T. Maruyama. 1989. Organelle gene diversity under migration, mutation, and drift: equilibrium expectation, approach to equilibrium, effects of heteroplasmic cells, and comparison to nuclear genes. Genetics 121:613–627.

Bossart, J. L., and J. M. Scriber. 1995. Maintenance of ecologically significant

genetic variation in the tiger swallowtail butterfly through differential selection and gene flow. Evolution 49:1163–1171.

Bull, V., M. Beltran, C. D. Jiggins, W. O. McMillan, E. Bermingham, and J. Mallet. 2006. Polyphyly and gene flow between non-sibling *Heliconius* species. BMC Biol. 4:11.

Celniker, S. E., and G. M. Rubin. 2003. The *Drosophila melanogaster* genome. Ann. Rev. Genom. Hum. Genet. 4:89–117.

Charlesworth, B., J. Coyne, and N. Barton. 1987. The relative rates of evolution of sex chromosomes and autosomes. Am. Nat. 130:113–146.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303.

Clark, A. 1997. Neutral behavior of shared polymorphism. Proc. Natl. Acad. Sci. USA 94:7730–7734.

Clarke, C. A., and P. M. Sheppard. 1962. The genetics of the mimetic butterfly, *Papilio glaucus*. Ecology 43:159–161>.

Coyne, J., and H. Orr. 2004. Speciation. Sinauer Associates, Sunderland, MA.

Dobzhansky, T. 1937. Genetics and the origin of species. Columbia Univ. Press, New York.

Dopman, E., L. Perez, S. Bogdanowicz, and R. Harrison. 2005. Consequences of reproductive barriers for genealogical discordance in the European corn borer. Proc. Natl. Acad. Sci. USA 102:14706–14711.

Endler, J. 1977. Geographic variation, speciation and gene flow. Princeton Univ. Press, Princeton, NJ.

Emelianov, I., F. Marec, and J. Mallet. 2004. Genomic evidence for divergence with gene flow in host races of the larch budmoth. Proc. R. Soc. Lond. B, Bio. Sci. 271:97–105.

Fay, J. C., and C.-I. Wu. 1999. A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol. Biol. Evol. 16:1003–1005.

Fay, J., and C. Wu. 2000. Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413.

Grahame, J. W., C. S. Wilding, and R. K. Butlin. 2006. Adaptation to a steep environmental gradient and an associated barrier to gene exchange in *Littorina saxatilis*. Evolution 60:268–278.

Gregory, T. R., and P. D. N. Hebert. 2003. Genome size variation in lepidopteran insects. Can. J. Zool. 81:1399–1405.

Haddrill, P., K. Thornton, B. Charlesworth, and P. Andolfatto. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Res. *15*:790–799.

Hagen, R., and M. Scriber. 1989. Sex-linked diapause, color and allozyme loci in *Papilio glaucus*: linkage analysis and significance in a hybrid zone. J. Hered. 80:179–185.

———. 1990. Population structure and host use in hybridizing subspecies of *Papilio glaucus*. Evolution 44:1914–1930.

Hagen, R., R. Lederhouse, J. Bossart, and M. Scriber. 1991. *Papilio canadensis* and *P. glaucus* are distinct species. J. Lepidopterist Soc. 45:245–258.

Harrison, R. 1990. Hybrid zones: windows on evolutionary processes. Oxf. Surv. Evol. Biol. 7:69–128.

Hey, J. 1994. Bridging phylogenetics and population genetics with gene tree models. Pp. 435–447 *in* B. Schierwater, B. Streit, G. Wagner and R. DeSalle, eds. Molecular ecology and evolution: approaches and applications. Birkhauser Verlag, Basel, Switzerland.

Hey, J., and R. Nielsen. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167:747–760.

———. 2005. On the number of New World founders: a population genetic portrait of the peopling of the Americas. PLoS Biol. 3:e193.

Hudson, R. 1983. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23:183–201.

Hudson, R. R. 1987. Estimating the recombination parameter of a finite population model without selection. Genet. Res. 50:245–250.

Hudson, R., M. Kreitman, and M. Aguade. 1987. A test of neutral molecular evolution based on nucelotide data. Genetics 116:153–159.

———. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Kaplan, N. L., T. Darden, and R. R. Hudson. 1989. The coalescent process in models with selection. Genetics 120:819–829.

Kaplan, N. L., R. R. Hudson, and C. H. Langley. 1988b. The "hitchhiking effect" revisited. Genetics 123:887–899.

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

Kliman, R. M., P. Andolfatto, J. A. Coyne, F. Depaulis, M. Kreitman, A. J. Berry, J. McCarter, and J. Hey. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. Genetics. 156:1913–1931.

Kronforst, M. R., L. G. Young, L. M. Blume, and L. E. Gilbert. 2006. Multilocus analyses of admixture and introgression among hybridizing Heliconius butterflies. Evolution. 60:1254–1268.

Lushai, G., D. A. S. Smith, I. J. Gordon, D. Goulson, J. A. Allen, N. Maclean. 2003. Incomplete sexual isolation in sympatry between subspecies of the butterfly *Danaus chrysippus* (L.) and the creation of a hybrid zone. Heredity 90:236–246.

Li, W. 1997. Molecular evolution. Sinauer Associates, Sunderland, MA.

Llopart, A., D. Lachaise, J. A. Coyne. 2005. Multilocus analysis of introgression between two sympatric sister species of *Drosophila*: *Drosophila yakuba* and *D. santomea*. Genetics 171:197–210.

Loader, S. 2006. locfit: local regression, likelihood and density estimation. R package ver. 5–3. Available from http:/www.locfit.info/.

Logsdon, J., M. Tyshenko, C. Dixon, J. -Jafari, V. Walker, and J. Palmer. 1995. Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. Proc. Natl. Acad. Sci. USA 92:8507–8511.

Machado, C., R. Kliman, J. Markert, and J. Hey. 2002. Inferring the history of speciation from multilocus DNA sequence data. Mol. Biol. Evol. 19:472–488.

Mallet, J., and N. Barton. 1989. Strong natural selection in a warning-color hybrid zone. Evolution 43:421–431.

Mayr, E. 1942. Systematics and the origins of species. Columbia Univ. Press, New York.

Moehring, A., A. L. Lopart, S. Elwyn, J. Coyne, and T. Mackay. 2006. The genetic basis of postzygotic reproductive isolation between *Drosophila santomea* and *D. yakuba* due to hybrid male sterility. Genetics 173:225–233.

Moriyama, E., and J. Powell. 1997. Synonymous substitution rates in *Drosophila*: mitochondrial versus nuclear genes. J. Mol. Evol. 45:378–391.

Nath, H., and R. Griffiths. 1996. Estimation in an island model using simulation. Theor. Popul. Biol. 50:227–253.

Nei, M. 1987. Molecular evolutionary genetics. Columbia Univ. Press, New York.

Nielsen, R., and J. Wakeley. 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158:885–896.

Orr, H. 1995. The population genetics of speciation: the evolution of hybrid incompatibilities. Genetics 139:1805–1813.

Orr, H., J. Maslly, and D. Presgraves. 2004. Speciation genes. Curr. Opin. Genet. Dev. 14:675–679.

Osada, N., and C.-I. Wu. 2005. Testing the mode of speciation with genomic data—examples from the great apes. Genetics 169:259–264.

Palopoli, M., A. Davis, and C.-I. Wu. 1996. Discord between the phylogenies inferred from molecular versus functional data: uneven rates of functional evolution or low levels of gene flow? Genetics 144:1321–1328.

Payseur, B., and M. Nachman. 2005. The genomics of speciation: investigation the molecular correlates of X chromosome introgression across the hybrid zone between *Mus domesticus* and *Mus musculus*. Biol. J. Linn. Soc. 84:523–534.

Presgraves, D. 2003. A fine-scale genetic analysis of hybrid incompatibilities in Drosphila. Genetics 163:955–972.

Presgraves, D., L. Balagopalan, S. Abmayr, and H. Orr. 2003. Adaptive evolution drives divergence of a hybrid inviability gene between two species of *Drosophila*. Nature 243:715–719.

Prowell, D. P., M. McMichael, J-F. Silvain. 2004. Multilocus genetic analysis of host use, introgression, and speciation in host strains of fall armyworm (Lepidoptera: Noctuidae). Ann. Entomol. Soc. Am. 97:1034–1044.

Przeworski, M., J. Wall, and P. Andolfatto. 2001. Recombination and the frequency spectrum in *Drosophila melanogaster* and *D. simulans*. Mol. Biol. Evol. 18:291–298.

Regier, J., Q. Fang, C. Mitter, R. Peigler, T. Friedlander, and M. Solis. 1998. Evolution and phylogenetic utility of the *period* gene in Lepidoptera. Mol. Biol. Evol. 15:1172–1182.

Rieseberg, L., J. Whitton, and K. Gardner. 1999. Hybrid zones and the genetic architecture of a barrier to gene flow between two wild sunflower species. Genetics 152:713–727.

Rozas, J., and R. Rozas. 1999. DnaSP ver. 3: an integrated program for molecular population genetics and molecular evolution. Bioinformatics. 15:174–175.

Sambrook, J., and D. Russell. 2001. Molecular cloning, a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sawamura, K., J. Roote, C.-I. Wu, and M. Yamamoto. 2004. Genetic complexity underlying hybrid male sterility. Genetics 166:789–796.

Schneider, S., D. Roessli, and L. Excoffier. 2000. Arlequin ver. 2000 a software for genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Scriber, J., B. Giebink, and D. Snider. 1991. Reciprocal latitudinal clines in oviposition behaviour of *Papilio glaucus* and *P. canadensis* across the Great Lakes hybrid zone. Oecologia 87:360–368.

Scriber, J. M., R. H. Hagen, and R. C. Lederhouse. 1996. Genetics of mimicry in the tiger swallowtail butterflies, *Papilio glaucus* and *P. canadensis*. (Lepidoptera: Papilionidae). Evolution 50:222–236.

Slatkin, M. 1973. Gene flow and selection in a cline. Genetics 75:733–756.

Sperling, F. A. H. 1993. Mitochondrial DNA variation's rule in *Papilio glaucus* and *Papilio troilus* groups. Heredity 71:227–233.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Takahata, N., and M. Nei. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325–344.

Tao, Y., S. Chen, D. Hartl, and C. Laurie. 2003. Genetic dissection of hybrid incompatibilities between *Drosophila simulans* and D. mauritana. I. Differential accumulation of hybrid male sterlility factors on the X and autosomes. Genetics 164:1383–1397.

Thompson, J., T. Gibson, F. Plewniak, F. Jeanmougin, and D. Higgins. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 24:4876–4882.

Thornton, K. R., and P. Andolfatto. 2006. Approximate bayesian inference reveals evidence for a recent, severe, bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics 172:1607–1619.

Ting, C., S.-C. Tsaur, and C.-I. Wu. 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. Science 282:1501–1504.

True, J., B. Weir, and C. Laurie. 1996. A genome-wide survey of hybrid incompatibility factors by the introgression of marked segments of *Drosophila mauritiana* chromosomes into *Drosophila simulans*. Genetics 142:819–837.

Turner, T. L., M. W. Hahn, S. V. Nuzhdin. 2005. Genomic islands of speciation in *Anopheles gambiae*. PLoS Biol. 3:1572–1578.

Wakeley, J., and J. Hey. 1997. Estimating ancestral population parameters. Genetics 145:847–855.

Wang, R., J. Wakeley, and J. Hey. 1997. Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. Genetics 147:1091–1106.

Watterson, G. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7:256–276.

Wittbrodt, J., D. Adam, B. Malitschek, W. Maueler, and F. Raulf. 1989. Novel putative receptor tyrosine kinase encoded by the melanoma-inducing *Tu* locus in *Xiphorphorus*. Nature 341:415–421.

Wu, C.-I. 2001. The genic view of the process of speciation. J. Evol. Biol. 14:851–865.

Associate Editor: M. Noor

Supplementary Material

The following supplementary material is available for this article:

Table S1. Sampling locations of *P. glaucus* and *P. canadensis*.
Table S2. Primer sequences for Z-linked loci used in this study. Genbank accession numbers are listed.
Table S3. Summary of the frequency distribution of polymorphisms.
Figure S1. Performance of divergence time estimation by our approximate Bayesian approach. Posterior distributions of 100 datasets simulated using parameters estimated from the data (see Materials and Methods) are shown. The vertical line shows the true divergence time of 3.2 *Ne* generations.
Figure S2. Approximate Bayesian posterior distributions of θ for each Z-linked locus for *P. glaucus* (black) and *P. canadensis* (gray).
Figure S3. Approximate Bayesian posterior distributions of ρ/θ for each Z-linked locus for *P. glaucus* (black) and *P. canadensis* (gray).

This material is available as part of the online article from:
http://www.blackwell-synergy.com/doi/abs/10.1111/j.1558-5646.2007.00076.x
(This link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Online Appendix Table S1: Sampling locations of *P. glaucus* and *P. canadensis*.

Individual	Location	Markers sequenced from individuals
Pc01	Oscoda Co., MI	*mt, Kettin, Ldh*
Pc02	Oscoda Co., MI	*mt, Kettin, Ldh, Per, Tpi, Titin*
Pc03	VT	*mt, Kettin, Ldh, Tpi*
Pc04	VT	*mt, Kettin, Ldh, Tpi, Titin*
Pc05	VT	*mt, Kettin, Ldh, Tpi, Titin*
Pc06	VT	*mt, Kettin, Per, Tpi*
Pc07	VT	*mt, Kettin, Ldh, Per, Tpi, Titin*
Pc08	VT	*mt, Kettin, Ldh, Per, Tpi, Titin*
Pc09	Oscoda Co., MI	*Ldh, Tpi, Titin*
Pc10	Emmet Co., MI	*mt, Per, Tpi, Titin*
Pc11	Cheboygan Co., MI	*mt, Kettin, Ldh, Per, Tpi, Titin*
Pc12	Cheboygan Co., MI	*mt, Kettin, Per, Tpi, Titin*
Pcan_sp	Edmonton,Canada	*mt*
PgFB30	Levy Co., FL	*mt, Kettin, Ldh, Per, Tpi, Titin*
PgFB31	Levy Co., FL	*mt, Kettin, Ldh, Tpi, Titin*
PgFB32	Levy Co., FL	*mt, Kettin, Ldh, Per, Tpi, Titin*
PgFB42	Levy Co., FL	*mt, Ldh*
PgFB43	Levy Co., FL	*mt, Kettin, Per, Tpi, Titin*
PgFB66	Levy Co., FL	*mt, Kettin, Ldh, Per, Tpi, Titin*
PgFB69	Levy Co., FL	*mt*
PgFB72	Levy Co., FL	*mt, Kettin, Tpi*
PgFY01	Levy Co., FL	*mt, Kettin, Ldh, Per, Tpi*
PgFY02	Levy Co., FL	*mt, Kettin, Ldh, Titin*
PgFY03	Levy Co., FL	*mt, Kettin, Ldh, Per, Titin*

PgFY04	Levy Co., FL	*mt, Kettin, Ldh, Per, Titin*
PgFY14	Levy Co., FL	*mt*
PgFY44	Levy Co., FL	*mt, Kettin, Ldh, Per, Tpi, Titin*
PgFY66	Levy Co., FL	*mt, Kettin,Ldh,Per, Titin*
PgFY73	Levy Co., FL	*mt*
PgFY75	Levy Co., FL	*mt*
PgGB13197	Clarke Co., GA	*mt*
PgGB13294	Clarke Co., GA	*mt*
PgGB13297	Clarke Co., GA	*mt*
PgGY04	Clarke Co., GA	*mt*
PgMB11	Warren Co., MO	*mt*
PgMB12	Warren Co., MO	*mt*
PgPY13	PA	*mt*
PgPY15	PA	*mt*
Pg_sp	PA	*mt*
PgOB03	Ohio	*mt*
PgGB59	Clarke Co., GA	*Tpi*
PgGB01	Clarke Co., GA	*mt, Tpi, Per*
PgGB02	Clarke Co., GA	*mt, Tpi, Per*
PgGB03	Clarke Co., GA	*Tpi, Per*

Online Appendix Table S2: Primer sequences for Z-linked loci used in this study. Genbank accession numbers are listed.

Locus	Primer	Total length	*B. mori* position/Accession	Sequence (5'-3')
Kettin	Ket1F	1206	112321/AB090307	CAGCACCCCGAAGGTGAAA
	Ket2R		113625/AB090307	CAACATCCCCAAGGCAAGGCA
Ldh	Ldh817F	439	36821/AADK01000293	GCGAGCAACCCCGTGGACATC
	Ldh2201R		37812/AADK01000293	CCAGTCTCCACCTACATCAAG
Period	Per1F	212	984/AY526605	GCGACTCCATTCTTCTCAGC
	Per2R		1150/AY526605	CTATTCATCATTCGGCATGC
Titin	Titin1F	611	15126/AB079867	AGCTCAATGCCAACCTAACG
	Titin2F		15341/AB079867	CATTCCAGTTATCAAAGAGAGACCAA
	Titin3R		15632/AB079867	GGCTACGTCTCAGCAAGTTGA
Tpi	Tpi1F	296	785 /AY734490	CAGGACCATCTTTGGTGAAA
	Tpi1R		2063/AY734490	GAGACCTTGGATGAAAGGGA

Online Appendix Table S3: Summary of the frequency distribution of polymorphisms. *P* values of Tajima's *D* (Tajima 1989) and Fay & Wu's *H* (Fay & Wu 2000) are listed in parentheses. The asterisk indicates $P < 0.05$. For the *H* test, *P. rutulus* was used as an outgroup and a correction for multiple hits was implemented (see Fay & Wu 2000).

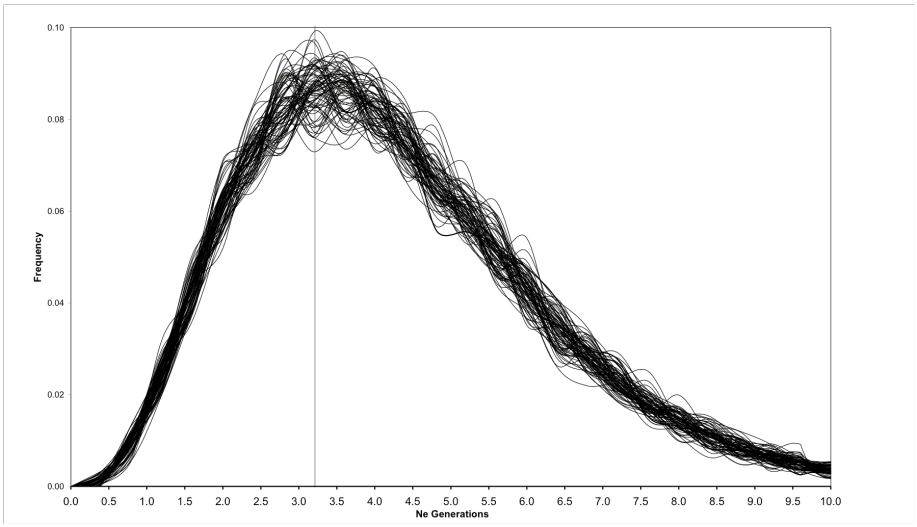
Gene	Species	Tajima's *D*	Fay & Wu's *H*
Kettin	*P. glaucus*	-1.02 (0.11)	0.79 (0.19)
	P. canadensis	-0.63 (0.23)	-0.80 (0.19)
Ldh	*P. glaucus*	0.41 (0.29)	2.70 (0.01)*
	P. canadensis	0.84 (0.10)	0.00 (0.43)
Period	*P. glaucus*	-0.58 (0.25)	0.55 (0.22)
	P. canadensis	0.76 (0.22)	-0.65 (0.16)
Titin	*P. glaucus*	0.56 (0.19)	1.75 (0.14)
	P. canadensis	0.67 (0.12)	-0.93 (0.26)
Period	*P. glaucus*	0.05 (0.48)	-1.94 (0.13)
	P. canadensis	-0.45 (0.26)	0.18 (0.50)
Mean Z-linked	*P. glaucus*	-0.12 (0.44)	0.77 (0.31)
	P. canadensis	0.24 (0.36)	-0.44 (0.33)
COI/COII	*P. glaucus*	-1.17 (0.13)	-6.68 (0.06)
	P. canadensis	-1.88 (0.02*)	-4.14 (0.07)

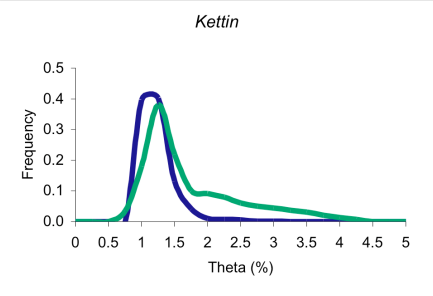
Appendix Figure Legend

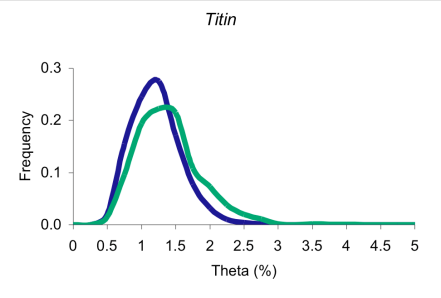
Appendix Figure 1: Performance of divergence time estimation. Posterior distributions of 100 simulations using parameters obtained from the average of the 5 Z-linked loci used in the study (θ = 1.1%, $\rho = 33.0\theta$, shared polymorphism =0, fixed polymorphism = 3) are shown. The real divergence time of 3.2 N_e generations is shown by the vertical line. The root mean square error between the maximum likelihood estimate and the real divergence time is 4.8 N_e generations.

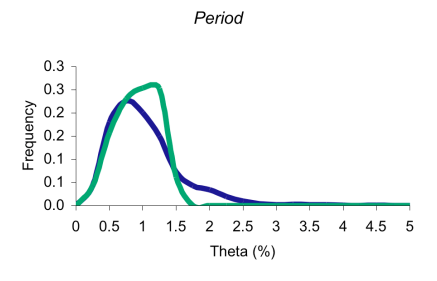
Appendix Figure 2: Approximate Bayesian posterior distribution comparing θ for each Z-linked locus between *P. glaucus* (blue) and *P. canadensis* (green).

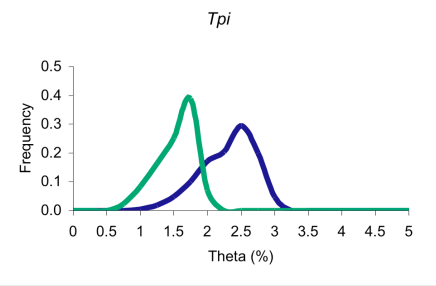
Appendix Figure 3: Approximate Bayesian posterior distribution comparing of ρ/θ for each Z-linked locus between *P. glaucus* (blue) and *P. canadensis* (green).

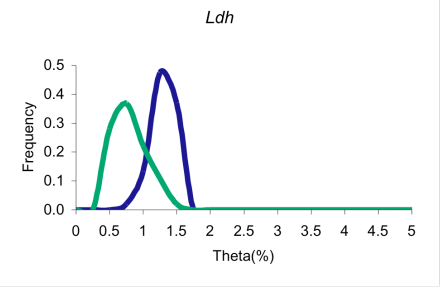

0.10
0.08
0.06
0.04
0.02
0.00
Frequency
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0 9.5 10.0
Ne Generations

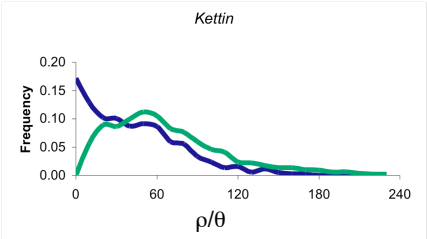

Kettin
Frequency
0.5
0.4
0.3
0.2
0.1
0.0
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5
Theta (%)

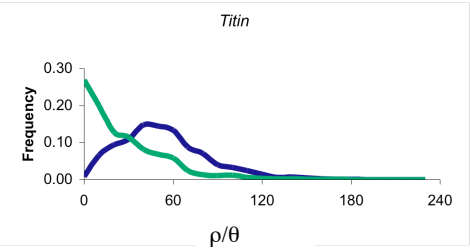

Titin
Frequency
0.3
0.2
0.1
0.0
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5
Theta (%)

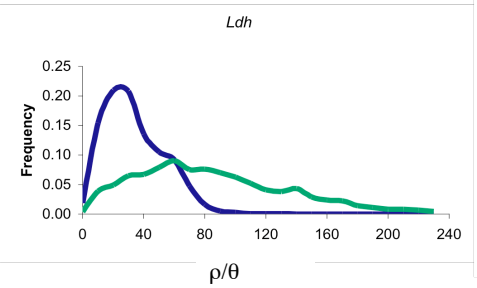

Period
Frequency
0.3
0.3
0.2
0.2
0.1
0.1
0.0
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5
Theta (%)

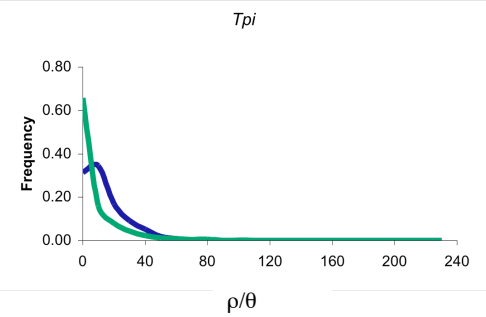

Tpi
Frequency
0.5
0.4
0.3
0.2
0.1
0.0
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5
Theta (%)

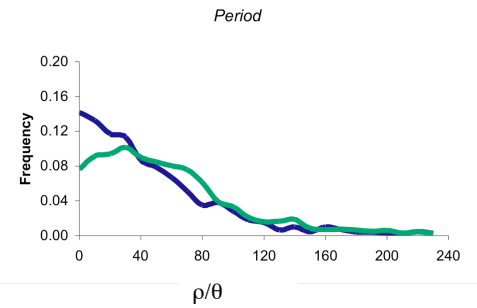

Ldh
Frequency
0.5
0.4
0.3
0.2
0.1
0.0
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5
Theta(%)


Kettin
Frequency
0.20
0.15
0.10
0.05
0.00
0
60
120
180
240
ρ/θ


Titin
Frequency
0.30
0.20
0.10
0.00
0
60
120
180
240
ρ/θ


Ldh
Frequency
0.25
0.20
0.15
0.10
0.05
0.00
0
40
80
120
160
200
240
ρ/θ


Tpi
Frequency
0.80
0.60
0.40
0.20
0.00
0
40
80
120
160
200
240
ρ/θ


Period
Frequency
0.20
0.16
0.12
0.08
0.04
0.00
0
40
80
120
160
200
240
ρ/θ

Chapter 2

A stepwise approximate Bayesian approach to estimating speciation times and ancestral population size in closely related species

A STEPWISE APPROXIMATE BAYESIAN APPROACH TO ESTIMATING SPECIATION TIMES AND ANCESTRAL POPULATION SIZE IN CLOSELY RELATED SPECIES

Andrea S. Putnam[1] and Peter Andolfatto[1]

[1] Section of Ecology, Behavior, and Evolution, Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093. E-mail: *asputnam@biomail.ucsd.edu*

Correspondence:
Andrea Putnam
Section on Ecology, Behavior, and Evolution
Division of Biological Sciences
University of California, San Diego
9500 Gilman Drive, MC0116
La Jolla, CA 92093
Tel: 858-822-1832
Email: asputnam@biomail.ucsd.edu

Keywords: Approximate Bayesian Computation, introgression, speciation, divergence, hybrid zones, Lepidoptera, *Papilio*, *Homo sapiens*, *Pan,* ancestral population size

Abstract

Estimating divergence times across the genomes of sister species can provide insight to the commonly accepted predominance of allopatric speciation. We improve upon an existing approach that uses approximate Bayesian computation (ABC) methods to test a null model of allopatric speciation (Putnam et al 2007). A joint ancestral population size is inferred using ABC with regression to speed computation time. In addition, the joint posterior distribution of ancestral population size is used as a prior parameter to infer and per locus and joint divergence times using ABC with rejection. The method is tested using two very different datasets. In two *Papilio* butterfly species the joint ancestral population size is very close to the current population size. A candidate locus for reproductive isolation *TH,* does not appear to have an unusual divergence time estimate compared to other Z chromosome–linked loci. However, allopatric speciation across the Z chromosome is rejected in favor of a more complex speciation process. Alternatively, using 14 autosomal markers in populations of humans and chimpanzees we are not able to reject a model of simple allopatric speciation. We find a recent joint divergence time estimate (4.5 million years) and an ancestral population size to be 2.5 times larger than the current population size. In summary, our method requires very little computation time and can be used to test modes of speciation and identify candidates for reproductive isolating factors.

Introduction

Allopatric speciation, an instantaneous speciation event, has long been considered the most common mode of speciation (Dobzhansky 1940, Mayr 1963). However, recent empirical studies have provided evidence for a more complex picture of the speciation process in many organisms (Osada and Wu 2005, Putnam et al 2007, Patterson et al 2006, Becquet and Przeworski 2007). These studies reveal that parapatric speciation is a better fit to multilocus data from many closely related species. In parapatric speciation, gene flow continues across some loci longer than others until genomic incompatibilities accumulate and eventually there is complete reproductive isolation (reviewed in Coyne and Orr 2004). Thus developing models that lend empirical support in testing speciation hypotheses in closely related species is of primary importance.

Here we use the method of approximate Bayesian computation (ABC) as the framework for estimating divergence times and ancestral population sizes. ABC has recently proved to be a successful alternative to likelihood methods (Beaumont et al 2002, Excoffier et al 2005, Hickerson et al 2006, reviewed in Beaumont and Rannala 2004). Additionally, analyses show that the results of ABC and likelihood methods are comparable under many evolutionary models (Beaumont *et al.* 2002; Marjoram *et al.* 2003, Tallmon et al 2004). Because it is

not necessary to calculate full likelihoods, ABC has the benefit of adding parameters with little computational trade-off (Beaumont and Rannala 2004). A disadvantage of ABC, however, is its reliance on simulations that use summary statistics of the observed data instead of using all available haplotype information as the full likelihood framework does. The key to effectively utilizing ABC thus becomes using summary statistics that collect the most important information from the data.

ABC is performed by obtaining summary statistics from the observed polymorphism data, which are then compared to the same statistics simulated under a specified model. It is most frequently implemented by either rejection or regression methods. For rejection sampling, only summary statistics from simulations that occur within in a tolerance are accepted and used to estimate the parameter of interest. The regression method differs by taking those remaining summary statistics that have been accepted and performing a weighted linear regression on them. Regression is considered to be faster and just as accurate as the rejection method (Beaumont et al. 2002). However, marginal likelihoods can be obtained using the rejection method but not regression. With the rejection method posteriors are binned per locus then multiplied across loci for a joint estimate of parameters. This enables one to test alternative speciation hypotheses using likelihood ratio tests (Bachtrog et al 2006, Putnam et al 2007).

Our aim is to test allopatry between two closely related species and identify reproductive isolating factors through estimating divergence times at individual loci and jointly across loci. Previously we developed a speciation time estimator (*STE*) to test a null hypothesis of allopatric speciation (Putnam et al 2007). The *STE* simulations used point estimates of observed data summary statistics to estimate the parameter of interest, divergence time, per locus and jointly using a rejection

method. A likelihood ratio test was then used to test whether one divergence time (allopatry; depicted in Figure 1) fits the data better than independent divergence times for each locus (parapatry). There are, however, limitations of this method. *STE* did not include estimation of the ancestral population size as a free parameter. Estimating the ancestral population size is important because it affects the extent of linkage disequilibrium and patterns of polymorphism and thus is an important consideration for association studies. In addition, if the ancestral population size is estimated to be large, gene and species genealogies will not always match because of greater variability in the coalescence. This can be seen in human data where gene trees are frequently more similar to gorillas than chimpanzees (Wall 2003, Chen and Li 2001). In addition, adding free parameters to STE, although still faster than MCMC methods, can become computationally prohibitive. Here we improve upon the *STE* method by 1) including a weighted linear regression to quicken the computation time and, 2) the estimation of a joint ancestral population size (θ_A).

Using the improved *STE*, called *STE2*, the method is tested in two datasets. The first dataset are Z-linked markers in the *Papilio glaucus* and *Papilio canadensis* species pair used in Putnam et al. 2007 with the addition of a new Z-linked marker, *tyrosine hydroxylase* (*TH*). *TH* is an enzyme in the *Papilio* pigmentation pathway (Koch et al 2000) and is a candidate as a reproductive isolating factor. Melanic females that mimic the unpalatable *Battus philenor* exist in *Papilio glaucus* but not *Papilio canadensis* females, although gene flow between the two species occurs through a hybrid zone. If *TH* is important in maintaining reproductive isolation between the two species we expect the estimated divergence time of the locus to be deeper than at the other loci.

The second dataset is a multilocus *Homo sapiens* and *Pan troglodytes* dataset. This is the first time a coalescent-based approach has been used to estimate human-chimp divergence times with population-level data. Recent studies using single species alignments have rejected allopatric speciation in humans and chimps (Osada and Wu 2005, Patterson et al 2006), suggesting gene flow may have occurred as recently as 4 MYA (Patterson et al. 2006). Population level analyses can shed light on this complex speciation process, testing whether a single or multiple divergence times is more likely among loci.

Methods

Papilio samples and loci

Papilio glaucus and *P. canadensis* were collected from a broad geographic sample of their ranges and identified by hind wing size and band width.

Sequences for all *Papilio* loci except for *tyrosine hydroxylase* (*TH*) have been previously published (Putnam et al, 2007). Primers for *TH* were designed from degenerate primers based on *Papilio xuthus* mRNA (Futahashi and Fujiwara, 2005). *TH* primers used in this study are TH_313F: 5' CCAAACAAAGTGTGCTCGAA3' and TH_1093R: 5' CTGGACCTGATGCCCAAGG3'.

PCR conditions in a thermocycler (Bio-Rad, Hercules, CA) to amplify *TH* included an initial denaturing step at 95°C for 2 min. followed by 40 cycles of 95°C for 30 sec, 53 °C for 45 sec, 72°C for 2 min with a final 5 min at 75°C. PCR product clean-up was performed using Exo/SAP reagents (Fermentas, Hanover, MD). Templates were directly sequenced on both strands and the BigDye sequencing kit (ver 3.1, Roche). Sequence reactions were run on an ABI 3730 sequencer (Applied Biosystems, Foster City, CA). *TH* is known to be Z-linked in *Heliconius* (Simon Baxter, personal communication) and was sequenced in a panel of male and female *P. glaucus* to confirm Z-linkage. Heterozygous sites were identified in males and never in females. Nucleotide sequences were edited and aligned using Sequencher 4.5 software (Gene Codes, Ann Arbor, MI) and manually adjusted (GenBank accession ___).

Human and Chimpanzee samples and loci

Sequences from a sample size of 10 eastern (Kenyan) chimpanzees were previously published in Fischer et al (2006). Human sequences from a sample

size of 15 African (Cameroonian) individuals were previously published in Frisse et al (2001) and Voight et al (2005). From 26 regions sequenced in both chimpanzee and human populations (Fischer et al 2006, Voight et al 2005), 15 regions had polymorphisms in both species, could be aligned and phase determined with confidence and are used in this study.

Polymorphism, divergence and recombination

All human and chimp loci are from unlinked, noncoding, intergenic, autosomal regions that have recombination rates close to the genome average (Frisse et al 2001, Voight et al 2005). Haplotype phase was determined using the software Phase 2.1 (Stephens et al 2001, Stephens and Scheet 2005). This software program uses a Bayesian method to determine the probability of reconstructed haplotypes. For each locus, homologous gorilla and orangutan sequences were obtained from Fischer et al (2006). These sequences were used to infer the ancestral state. Polymorphisms at CpG sites were filtered out and a correction for recurrent mutation was performed (Jukes and Cantor 1969). *Papilio* loci are Z-linked, exon and intron regions sequenced only in females, thus haplotypes reconstruction was not necessary.

All silent sites (synonymous and non-coding) were considered for both datasets, while sites with insertions and/or deletions, and more than two variants were excluded. Watterson's estimator, θ_w (Watterson 1975), and the average pairwise diversity per nucleotide, π (Tajima1983) were used to estimate levels of

nucleotide diversity. Divergence was estimated using D_{XY}, the average pairwise number of nucleotide substitutions per site between species (Nei 1987).

The population mutation rate ($\theta = 4N_e\mu$, where N_e is the effective population size and μ is the mutation rate per base pair) and the population recombination rate ($\rho = 4N_e r$ where r is the recombination rate per base pair) were jointly estimated using ABC with rejection sampling (Thornton and Andolfatto 2006). This method jointly estimates ρ and θ based on summary statistics of the data.

Papilio glaucus and *P. canadensis* are known to have high rates of recombination compared to the most commonly studied invertebrate, *Drosophila melanogaster* (Putnam et al. 2007, Thornton and Andolfatto 2006). For *Papilio,* recombination parameters were the same as in Putnam et al. (2007) with the *TH* locus added to the previously published dataset. Indeed, in *P. glaucus* the average joint multilocus mode of ρ is 33θ and is 30.5θ in *P. canadensis*. These estimates are similar to one another and close to the previous *Papilio* average recombination estimate ($\rho \approx 33.5\theta$*;* Putnam et al 2007).

Using a high-resolution recombination map available on the UCSC database, the average recombination rate in each human region examined in this study varies from 0.8 – 2.9 cM/Mb (Kong et al, 2002). Variation in recombination rates within and between species can affect divergence time estimates. Loci in areas of low recombination may be subject to increased hitchhiking effects that reduce

nucleotide variability. Thus, as fine-scale recombination rates appear to vary between humans and chimps (Ptak et al 2005), per locus and joint intraspecific estimates were obtained for both species using the ABC approach. Uniform priors for θ (0.0001, 0.02) and ρ (0, 1.0) were chosen. For both parameters, the tolerance for acceptance was set to 0.01. The posterior distribution was based on 1000 acceptances.

The mode of ρ was similar between humans and chimps (Supplementary Figure 1). This observation of the similarity in recombination rate is in line with previous studies using larger datasets (Ptak et al 2004b). The joint multilocus mode of $\rho = 0.9\theta$ in the Hausa human population, and slightly higher, $\rho = 1.2\theta$, in the chimpanzee population. Again, these estimates are quite close to previous recombination rate estimates (Frisse et al 2001, Ptak et al 2004a, Ptak et al 2004b). For simulations, $\rho = 31.8\theta$ is used for *Papilio* and 1.05θ in humans and chimpanzees, which represents the average maximum a posteriori (MAP) estimate for each species pair.

Tests of neutrality

Patterns of polymorphism that do not deviate from neutrality are an assumption of the coalescent-based STE approach. The human, chimp loci and *Papilio TH* locus were tested for departures from neutrality using two summaries of the distribution of polymorphism frequencies. Tajima's *D* (Tajima 1989) is a measure of the standardized difference between π and θ_w. Fay and Wu's *H* measures the

difference between π and θ_H, an estimator of θ that weights derived variants by the square of their frequencies (Fu 1996; Fay and Wu 2000). Deviations from neutrality in *D* and *H* were ascertained by comparing the observed value to a distribution obtained from simulating 1,000 neutral genealogies with recombination in *ms* (Hudson 2002). A correction for multiple tests was calculated using the Q-value R package (Storey 2002). Simulations were performed with a point estimate of Watterson's θ and T_{obs}. $T_{obs} = (D_{XY}/ \theta_j) - 1$ where θ_j is the average θ between the two species (Hudson et al 1987).

Summary statistics of *TH* in *Papilio* are presented in Supplementary Table 1. The level of nucleotide diversity of *TH* in *Papilio canadensis* is close to the average for five previously reported neutral Z-linked markers (*P. canadensis* $\hat{\theta}_W$= 0.9%; Putnam et al. 2007). However, in *P. glaucus* diversity of *TH* is more than twice as high as the previously reported average (*P. glaucus* $\hat{\theta}_W$ = 1.3) Values of Tajima's D and Fay and Wu's H are not significant. Taken together *TH* does not appear to deviate from neutral expectations.

Human and chimpanzee polymorphism and divergence per locus are given in Supplementary Table 2. Human and chimpanzee polymorphism summary statistics are in line with those previously reported using these datasets (Frisse et al. 2001, Fischer et al 2006). The average nucleotide diversity is θ_W = 0.18%. Human nucleotide diversity is θ_W = 0.14%, slightly less than in chimpanzees. Average Tajima's *D* and Fay and Wu's *H* for chimps ($\hat{D}$ = -0.34 $\hat{H}$ = 0.27) and

humans ($\tilde{D}$ = -0.42, $\tilde{H}$ = -0.74) do not deviate from neutral expectations. Two human loci (regions 5 and 15) and two chimp loci (regions 8 and 16) have values of *H* that significantly deviate from neutral expectations. After correcting for multiple tests (false discovery rate = 5%), however, only region 15 remains significant (*P* =0.05) in humans. Region 15 also shows reduced levels of polymorphism compared to other human loci surveyed here, a signature of positive selection or demographic effects. This locus was removed from further analyses in both the human and chimp dataset.

Joint ancestral population size estimation

Ancestral population size (θ_A) is estimated using a modification of the *STE* program (Putnam et al 2007), which did not estimate θ_A in its original implementation. Here, θ_A is estimated jointly because per locus posterior distributions are broad (data not shown). We chose to use a joint estimate, which effectively pools information across loci resulting in a narrower posterior distribution. Additionally, in the original framework of *STE* acceptances are conditioned on the observed *T* and the number of shared and fixed differences between the two species using a rejection method. This method becomes computationally intensive when the number of loci is large or if more parameters (i.e. migration) are added because the number of simulations necessary to get a posterior is prohibitive. Instead, to estimate a joint θ_A, simulations performed where a proportion of the priors are accepted that are closest to the observed values (Beaumont et al 2002). This modification uses ABC with a weighted linear

regression instead of rejection sampling to obtain the joint posterior probability density for the parameter θ_A across all loci. A detailed description of the ABC method is outlined in Beaumont et al 2002.

To overview a generalization of our implementation:

1) For each locus the following summary statistics of the observed data (obs) are obtained: number of individuals (n), sequence length, (L), segregating sites (S), nucleotide diversity (θ), recombination (ρ), shared polymorphisms (Sh), fixed differences (Fi) and divergence time (T_{obs}). T_{obs} is estimated as $(D_{XY}/\ \theta_j) - 1$ where θ_j is the average θ between the two species (Hudson et al 1987).

2) One draw of the hyperparameter θ_A/θ_1, the ancestral population size relative to the current population size, is chosen from a uniform prior (0.1 – 10).

3) Simulations (sim) of each locus are performed under a model of strict allopatric divergence (Figure 1) using θ_A calculated from θ_A/θ_1 chosen in the previous step. Observed per locus point estimates for n, L, θ_1, θ_2, ρ, and T_{obs} are used as simulation parameters (Table 1).

4) For each draw of θ_A/θ_1 summaries Sh, Fi, and T of the simulated data (sim) are recorded for each locus in a matrix. θ_A/θ_1 is accepted if $|obs - sim| < \varepsilon$ for Sh, Fi, and T across all loci, where the tolerance (ε) is a proportion of draws accepted.

5) A weighted linear regression is performed on the accepted posterior distribution of θ_A/θ_1. This posterior distribution is then used as a prior distribution for estimation of per locus and joint *T*.

Table 1 contains the summary statistics of the observed data for both datasets. For this analysis a total of 1.0 x 10^6 draws of θ_A/θ_1 are performed and the tolerance used here is 0.001. Thus, the posterior distribution is made up of 1000 acceptances.

Per locus and joint estimates of divergence time

Once a joint multilocus estimate of θ_A/θ_1 is obtained, per locus and a joint *T* are inferred using ABC with rejection to obtain marginal likelihoods. This estimation differs from the previous step because there is no hyperparameter. Here, *T* for each locus is estimated independently. For each simulation *T* is chosen from a uniform prior (0 – 16 N_e generations). However, for each locus we perform coalescent simulations that draw from the joint posterior θ_A/θ_1 estimate. Point estimates of per locus observed summary statistics are used for n, L, θ_1, θ_2, and ρ. *T* is accepted conditioning on |sim –obs| < ε for the parameters Sh, Fi, and T_{sim}. Simulations are performed until there is a posterior distribution of 1000 acceptances. The fixed tolerance (ε) was set to 0.05, and the average acceptance rate was ~10^3 per locus. Using the posterior distribution for each locus we obtained a MAP estimate and 95% confidence interval for *T* using the

'locfit' statistical package in *R* (Loader 2006). Next we obtained a joint estimate for *T*. This is done by binning the posterior distribution of per locus estimates in increments of 0.2 N_e (Putnam et al. 2007). Probabilities for each bin are multiplied across loci. The bin with the largest likelihood is the MAP estimate for the joint *T*.

A library of scripts and programs to implement this procedure is available on request. Simulations were performed using *ms* (Hudson 2000). P. Andolfatto, D. Bachtrog, and A. Putnam wrote scripts to calculate summary statistics of the simulations. K. Thronton wrote the scripts to perform the weighted linear regression.

Test of allopatry

Once a per locus and joint marginal likelihoods for *T* is inferred, a likelihood ratio test (LRT) can be employed to test various hypotheses about species' divergence. Here we are interested in testing the fit of the data using one divergence time for all loci (strict allopatry) compared to individual divergence times for each locus (parapatry). We perform a LRT using the joint MAP estimate of *T* and compare it to the MAP estimates from the per locus *T* estimates. If the LRT is significant allopatry can be rejected.

Mutation rate variation

The *STE* method assumes no variation in mutation rate. We tested whether there

was a significant variation in mutation rate across loci for each dataset. The distribution of observed and expected divergence for each locus was compared and a goodness-of-fit was determined using a Pearson χ^2 test (Frisse et al 2001, Becquet and Prezworski 2007). Neither the *Papilio* (P = 0.59) nor the human/chimpanzee (P = 0.28) dataset significantly deviated from expectations. Thus, it is not necessary to account for variation in mutation rates in these analyses.

Performance

Performance of the modified *STE* program. We generated 20 simulated datasets based on the human and chimpanzee dataset studied here, under a model of strict allopatry. Each dataset had 14 loci and 20 sampled chromosomes from each population. For simulation parameters of each dataset and locus we used mean θ_1 *and* θ_2 from the observed dataset, and the point estimate of $\rho = 1.05\theta$ and T = 4.5 MYA. A point estimate of θ_A/θ_1 = 2.31 was used to simulate the ancestral population size.

Using the weighted linear regression method of ABC as outlined above, the joint posterior distribution of θ_A/θ_1 was estimated from a prior uniform distribution (0.1, 10). The tolerance, ε, was 0.001 where ε was a proportion of simulated draws of θ_A/θ_1 and 1 X 10^6 simulations were performed. Using the resulting posterior of θ_A/θ_1 as a prior for θ_A, we estimated T per locus from a prior uniform distribution (0,16) N_e generations using ABC with rejection for each dataset. We conditioned

acceptances on |*sim* – *obs*| < ε, for the number of shared and fixed differences and *T*, where ε = 0.05. Simulations were continued until 1000 acceptances were obtained. A joint T for each dataset was inferred using the method described above.

We also compared the performance of *STE* to the *MIMAR* method of Becquet and Przeworski (2007) with same simulated dataset under an allopatric model. Estimated parameter values for θ_1, θ_2 were chosen from uniform priors (0, 0.01) per base pair. θ_A and *T* were also chosen from uniform priors, (0, 0.01) per base pair and (0, 2 X 10^7) generations, respectively. We ran two independent runs for each dataset and ran them until the posterior distributions were similar. 5 X 10^5 burn-in steps were recorded. For each method, a mutation rate of 2 x 10^{-8} is assumed a generation is 20 years.

Results

Ancestral population size

We used *STE2* to estimate the joint multilocus ancestral population size, θ_A, in the two datasets. A joint θ_A has the advantage of more strength than per locus estimates. The MAP estimate of the θ_A for *P. glaucus* and *P. canadensis* is 1.09θ_1 (0.64 – 1.80 95%CI) using ABC with weighted linear regression (Figure

2a). Previously, using STE to estimate divergence times under different θ_A values, Putnam et al. (2007) found that all estimates of $\theta_A \geq 5$ times greater than the current θ were a poor fit to the data. Thus, the method used here allows us to better approximate θ_A.

Using the same method as for the *Papilio* dataset, the chimpanzee and human joint θ_A was estimated to be 2.31 x greater than θ_1 (1.14 – 3.53 95%CI; Figure 2b). This is concordant with previous studies where the ancestral population size of humans and chimpanzees was estimated to be 2-3 times greater than the current human population size (Chen and Li 2001, Yang 2002, Wall 2003).

Per locus divergence time estimates

Per locus *T* was inferred using the rejection-based ABC method in order to compare them to the joint *T* and identify outlier loci with deep divergence times that may be speciation factors. We used the posterior distribution obtained in the previous step for the ancestral population size (θ_A) parameter. For both datasets, per locus divergence time estimates and 95% confidence intervals are listed in Table 2.

MAP estimates of divergence time varied among the 6 *Papilio* loci, ranging from 1.6 to 11.4 N_e generations (Table 2, Figure 3a). Values of *T* are reported in units of N_e and not absolute time because the mutation rate is unknown in *Papilio* and generation time in *P. glaucus* and *P. canadensis* may differ (Hagen 1991). The estimated divergence time of *TH* is 3.84 N_e generations ago (1.25 – 6.22 95% CI). The divergence time estimate of this locus look intermediate compared to the other five loci.

Using 14 loci to examine divergence time estimates per locus between humans and chimpanzees, values of *T* ranged from approximately 3.8 to 9.74 MYA (Table 3, Figure 3b). To convert the divergence time from units of N_e generations to the absolute divergence time we assumed a mutation rate, μ, of 2×10^{-8} per base pair per generation (Nachman and Crowell 2000) and a generation time of 20 years. These estimates span the range of previous divergence time estimates for the two species that were estimated along different genomic regions (Glazko and Nei 2003, Kumar et al 2005, Hobolth et al 2007, Patterson et al 2006). Some studies, however, have used a human/chimpanzee generation time of 25 years (Eyre-Walker and Keightley 1999, Wall 2003), which would push these per locus estimates back to 5.2 to 11.15 MYA.

Test of Allopatry

A purely allopatric model of speciation is tested in each of the datasets using a likelihood ratio test. We ask whether one divergence time (strict allopatry) fits the data better than separate divergence times for each locus (parapatry). In *Papilio* a joint MAP estimate of divergence time is 3.5 N_e generations ago. If we assume the mutation rate is similar to *Drosophila* (1.5 X 10^{-8} per year, Li 1997) and assume there is one generation per year then The fit of an allopatric model is significantly rejected in *Papilo glaucus* and *P. canadensis* (χ^2 = 22.43, d.f. 5, P << 0.001; Table 4). This value is in agreement with the T estimate of 3.2 N_e generations from Putnam et al (2007). Thus, adding the *TH* candidate locus and a more precise estimate of θ_A did not significantly impact the divergence time estimate of the two species compared to the previous study of Z-linked markers.

The human and chimpanzee joint MAP divergence time estimate is 4.5 MYA. This estimate of T is also in agreement with previous multilocus estimates that range between 4-6 MYA (Wall 2003, Glazko and Nei 2003, Kumar et al 2005, Hobolth et al 2007, Patterson et al 2006) when 20 years is the estimated generation time for both species. Our MAP estimate falls on the more recent end of the range, which is often considered the time when all gene flow stopped. Despite the range of divergence time estimates, allopatry is not rejected among these autosomal loci (χ^2 = 4.03, d.f. 14, P = 0.96; Table 4).

Performance

The performance of *STE2* that incorporates an estimation of the ancestral population size and weighted linear regression was assessed. We generated 20 simulated datasets based on the human and chimpanzee data, each with 14 loci, under a model of strict allopatry. *STE2* estimation with the addition of ancestral population size, $\theta_{A,}$ is accurate with parameters similar to our data. Using the stepwise method to first jointly estimate θ_A then *T* conditioning on shared and fixed differences and T_{obs} resulted in low mean bias and a mean square error (Table 5). Performance of the modified *STE* under an allopatric model was compared to *MIMAR* (Becquet and Przeworski 2007) under the same model. The mean bias and square error for θ_A and *T* are comparable between methods (Table 5). In *MIMAR* 95% confidence levels tend to be broader than *STE2*. The comparison, however, is not completely fair because *MIMAR* had two additional free parameters that STE did not (θ_1 and θ_2). Because *MIMAR* uses an MCMC approach, the time to convergence of the posterior distributions of independent runs took approximately 72 hours, much slower than *STE2,* which took approximately 10 hours.

Discussion

Here we used a stepwise ABC method to estimate the ancestral population size and divergence time of sister species with multiple loci. First, θ_A relative to the

current population size was jointly estimated over all loci. The addition of weighted linear regression at this step decreases the computation time, which can be prohibitive with the rejection method when several parameters are estimated and the tolerance is low. Using the posterior distribution of θ_A obtained from the previous step, per locus divergence time estimates, *T*, are next determined with the rejection method. The rejection-based ABC method has the advantage of producing marginal likelihoods for *T* estimates at each locus. Finally, a joint likelihood *T* was obtained and compared to per locus estimates to determine if an allopatric model of speciation can be rejected.

The performance of this method works well for datasets similar to the ones used here. With low bias and error STE is comparable to other methods to estimate divergence times (Hey and Nielsen 2007, Becquet and Przeworski 2007). The *STE2* program can be easily expanded to incorporate migration or test divergence times under different demographic scenarios like bottlenecks. A promising alternative method to estimate speciation parameters, including migration, was recently developed by Becquet and Przeworski (2007). This program, *MIMAR*, combines Bayesian and MCMC methods to estimates the posterior probability of parameters of interest. The advantage of *STE* over this method is substantially less computation time. We compared our *Papilio* results with *STE* to results obtained using *MIMAR* under strict allopatry (Supplementary Methods). Using *MIMAR*, the MAP estimates of *T* and θ_A are close to those reported here (Supplementary Results). Both methods have an upward bias that

will lead to slightly deeper divergence time estimates and ancestral population sizes.

TH as a candidate for reproductive isolation in Papilio

The estimate of the ancestral population size for *P. glaucus* and *P. canadensis* was 1.09θ_1 (Figure 2a). This θ_A is in the range we expected based on testing the goodness of fit of five Z-linked markers to a range θ_A estimates (Putnam et al 2006). *P. glaucus* and *P. canadensis* appear to have a history of large population sizes and little evidence of population structure. The Z-linked locus, *TH*, is part of the *Papilio* pigmentation pathway (Koch et al 1998, Koch et al 2000), and therefore a candidate for being a reproductive isolation factor, as mimicry is only seen in females of *P. glaucus*. If *TH* is important in driving reproductive isolation between the species, it is expected to have a deeper divergence time estimate compared to other markers. Although there were three fixed polymorphisms and no shared polymorphism between the two species, the divergence time estimate was not unusual compared to the other markers (Figure 3a). Another gene in the pigmentation pathway, β-alanyldopamine synthase (BAS), is known to have differential expression in yellow compared to melanic individuals (Koch et al 2000) and may be a more likely candidate as a reproductive isolation factor.

Human and chimpanzee divergence

This is the first report of divergence time between humans and chimpanzees being estimated with a coalescent-based approach the population level. Only recently have chimpanzee population data been made available (Fischer et al 2006). We estimate the ancestral population size of human and chimpanzees to be 2.31 times greater than the current human population size (Figure 2b). This estimate is within the range of previous estimates, which vary from 2 - 10 times larger than the current population size of 10,000 (Chen and Li 2001, Takahata 2001, Yang 2002, Rannala and Yang 2003, Wall 2003, Hobolth et al 2007). In addition to population structure, other factors may be influencing the large ancestral population size compared to the current size in humans and chimpanzees. Differences among loci in mutation rate will increase the variation of T among loci and subsequently increase the joint N_a estimate (Wall 2003, Innan & Watanabe 2006). Using a hidden Markov model, Hobolth et al (2007) found an N_a ~5 x larger than N_e. They were unable to discern whether the large ancestral population size estimate or a lengthy speciation process caused their large variance in divergence time estimates (Hobolth et al 2007). We do not find evidence of mutation rate variation in loci examined in this study. Here we show that the variation in T among loci can still be explained with a θ_A that is only 2.5 times larger than θ_1.

Our joint divergence time estimate for humans and chimpanzees was 4.5 MYA (Figure 2b). A joint MCMC coalescent method that used a ~ 2 million base pair alignment of humans and chimpanzees was 4.1 MYA (Hobolth et al 2007), very

close to our estimate. We find that the likelihood of a divergence time for each locus is not significantly different from one speciation time (strict allopatry). Thus, we are unable to reject allopatry using these loci. Also using noncoding autosomal markers, but employing a maximum-likelihood method, Innan and Watanabe (2006) similarly report that allopatry with no significant evidence of gene flow is a good fit to the human and chimpanzee data. These results are in contrast to Patterson et al (2007) who reject an allopatric model of speciation. Their method found a significantly more recent divergence estimates time across much of the X chromosome than expected when compared to autosomes. X chromosome markers are not used in this study, but may increase the likelihood of rejecting allopatry when this data becomes available in a chimpanzee population.

In conclusion, we have applied the ABC method to estimate divergence times and ancestral population sizes of closely related species. We were able to reject *TH* as a candidate as a reproductive isolation factor in *Papilio* speciation. Although its divergence time estimate compared to other markers was unremarkable, allopatric speciation across the Z chromosome is rejected in favor of a more complex speciation process. Alternatively, using 14 autosomal markers in populations of humans and chimpanzees we are not able to reject a model of simple allopatric speciation. We find this method to be fast and precise, and extensions of this method should not prove to be computationally prohibitive.

Acknowledgements: The authors would like to thank Anna Di Rienzo for Hausa polymorphism data, as well as D. Bachtrog and K. Thornton for providing perl scripts used in analyses. This work was supported by National Science Foundations grants DDIG (DEB- 0710135) and DEB-0717007.

References:

Andolfatto P, Wall JD. (2003) Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster.* Genetics 165:1289–1305.

Barton NH (2006) Evolutionary biology: How did the human species form? Curr Biol 16: R647–R650.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162: 2025–2035.

Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. Nature Reviews Genetics, 5, 251–261.

Becquet C, Przeworski M (2007) A new method to estimate parameters of speciation models, with application to apes. Genome Research 17: 1505-19.

Chen FC, Li WH (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees, Am J Hum Genet 68: 444–456.

Coyne JA, Orr HA (2001) Speciation. Sinauer Associates, Sunderland, MA.

Excoffier L, Estoup A, Cornuet J-M (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. Genetics. Mar;169(3):1727-38.

Eyre-Walker A, Keightley P (1999) High genomic deleterious mutation rates in hominids. Nature 397: 344-347.

Fay J, Wu C (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413.

Fischer A, Pollack J, Thalmann O, Nickel B, Paabo S. (2006) Demographic history and genetic differentiation in apes, Curr. Biol. 16 (11):1133-1138.

Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A (2001) Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. 69: 831-843.

Fu Y-X (1996) New statistical tests of neutrality for DNA samples from a population, Genetics 143:557–570.

Futahashi R, Fujiwara H. (2005) Melanin-synthesis enzymes coregulate stage-specific larval cuticular markings in the swallowtail butterfly, *Papilio xuthus*. Dev. Genes Evol. 215 (10):519-529.

Glazko GV, Nei M (2003) Estimation of divergence times for major lineages of primate species. Mol Biol Evol 20: 424–434.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P (2005). Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations. Genome Res 15: 790–799.

Hagen R, Lederhouse R, Bossart J, Scriber M (1991) *Papilio canadensis* and *P. glaucus* are distinct species. J. Lepidopterist Soc. 45:245-258.

Hey J, Nielsen R. (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc Natl Acad Sci U S A 104(8):2785-90.

Hickerson, MJ, Stahl E, Lessios HA (2006) Test for simultaneous divergence using approximate Bayesian computation. Evolution 60: 2435-2453.
Hobolth A, Christensen OF, Mailund T, Schierup MH (2007). Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genetics, 3, 294-304.

Hudson R (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Hudson R, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159.

Jukes TH, Cantor, CR (1969) Evolution of protein molecules. Pp. 21–132 in H. N. Munro, ed. Mammalian protein metabolism. Academic Press, New York.

Koch PB, Keys DN, Rocheleau T, Aronstein K, Blackburn M, Carroll SB, ffrench-Constant RH (1998) Regulation of dopa decarboxylase expression during color pattern formation in wild-type and melanic tiger swallowtail butterflies. Development 125:2303–2313

Koch PB, Behnecke B, ffrench-Constant RH (2000) The molecular basis of melanism and mimicry in a swallowtail butterfly. Current Biology 10(10):591-4

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002). A high-resolution recombination map of the human genome, Nature Genetics, 31(3), 241-247

Kumar S, Filipski A, Swarna V, Walker A, Hedges SB (2005) Placing confidence limits on the molecular age of the human–chimpanzee divergence. Proc Natl Acad Sci U S A 102: 18842–18847.

Marjoram P, Molitor J, Plagnol V, Tavare S (2003) Markov chain Monte Carlo without likelihoods. Proc Natl Acad Sci U S A. 100(26):15324-8.

Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156:297–304.

Nei M (1987) Molecular Evolutionary Genetics Columbia University Press: New York, NY.

Osada N, Wu C-I (2005) Testing the mode of speciation with genomic data - Examples from the great apes. Genetics 169: 259-264.

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D (2006) Genetic evidence for complex speciation of human and chimpanzees. Nature 441: 1103–1108.

Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA and Paabo S (2005) Fine-scale recombination patterns differ between chimpanzees and humans. Nat. Genet., 37, 429–434

Putnam AS, Scriber JM, Andolfatto P (2007) Discordant divergence times among Z chromosome regions between two ecologically distinct swallowtail butterfly species. Evolution 61(4):912-927.

Rannala B, Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645-1656.

Stephens M, Smith NJ, Donnelly P (2001). A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics 68, 978–989.

Stephens M, Scheet P (2005). Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation. American Journal of Human Genetics, 76:449-462.

Storey JD (2002) A direct approach to false discovery rates. J. Roy. Statist. Soc. B., 64, 479–498.

Takahata N, Lee S-H, Satta Y (2001) Testing multiregionality of modern human origins. Mol. Biol. Evol. 18: 172–183.

Tallmon DA, Beaumont MA, Luikart GH (2004) Effective population size estimation using approximate Bayesian computation. Genetics 167:977-988.

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

Thornton KR, and Andolfatto P. (2006) Approximate Bayesian inference reveals evidence for a recent, severe, bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics, 172:1607-1619.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A (2005). Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc. Natl. Acad. Sci. USA 102, 18508–18513.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. 7:256-276

Wall JD (2003) Estimating ancestral population sizes and divergence times, Genetics 163:395–404.

Yang Z (2002) Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 162: 1811–1823.

Table 1: Summary statistics of observed data used in simulations for inference of joint θ_A. θ_1 and θ_2 represents *P. glaucus* and *P. canadensis,* respectively, in the first dataset and humans and chimpanzees in the second dataset. See Methods for information on parameters.

	L	θ_1[a]	θ_2[a]	ρ[b]	Sh	Fi	*T*
P.glaucus/P.canadensis							
Kettin	1206	0.005	0.007	39.91	0	7	5.33
Ldh	439	0.017	0.006	135.69	0	3	1.52
Period	212	0.006	0.008	47.89	0	3	3.17
TH	588	0.028	0.011	223.49	0	3	1.21
Titin	611	0.011	0.012	87.80	10	1	0.48
Tpi	296	0.024	0.013	191.56	3	0	1.38
Human/Chimpanzee							
Region1	843	0.0012	0.0017	1.06	0	4	4.24
Region2	795	0.0016	0.005	1.34	0	4	1.82
Region3	1144	0.0013	0.0022	1.56	0	16	8.37
Region5	1555	0.0011	0.0009	1.80	0	11	12.60
Region6	934	0.0011	0.0033	1.08	0	9	5.00
Region7	1183	0.0015	0.0012	1.86	0	13	9.96
Region8	775	0.0007	0.0011	0.57	0	4	6.89
Region10	781	0.0023	0.0037	1.89	1	4	1.97
Region11	613	0.0012	0.0005	0.77	0	6	12.41
Region14	738	0.0014	0.0017	1.08	1	2	3.00
Region16	786	0.0025	0.001	1.57	0	11	10.66
Region17	598	0.0021	0.0015	1.33	0	9	8.56
Region21	601	0.0021	0.0015	1.55	0	11	10.93
Region22	705	0.0009	0.0015	0.54	0	4	6.83

[a]Estimate of Watterson's θ per site (Watterson 1975).
[b] Recombination rate per gene

Table 2: *P. glaucus* and *P. canadensis* per locus MAP estimates of divergence time, *T,* and 95% confidence intervals in N_e generations.

Locus	*T* (MAP)	95% CI
Kettin	11.14	9.81-13.29
Ldh	3.52	2.95 - 5.51
Period	6.50	4.21 - 8.68
TH	3.84	1.25 – 5.22
Titin	1.60	0.27 - 3.86
Tpi	2.64	0.95 - 4.11
Joint Z	3.0	1.22 – 5.69

Table 3: Human and chimpanzee per locus divergence time estimates, *T,* and 95% confidence intervals

Locus	*T* (MYA)	95% CI
Region1	4.01	1.33 - 6.92
Region2	7.67	3.01 - 15.16
Region3	7.54	2.47 – 8.61
Region5	4.92	1.13 – 5.47
Region6	9.74	3.38 – 11.13
Region7	4.35	1.62 – 8.04
Region8	4.83	0.91 – 4.46
Region10	4.19	1.24 – 7.11
Region11	4.02	0.77 – 5.63
Region14	4.21	0.54 – 3.50
Region16	5.63	1.49 – 9.72
Region17	6.26	1.47 – 8.63
Region21	4.91	1.17 – 8.29
Region22	4.58	1.25 – 7.96

Table 4: Goodness-of-fit tests for a model of allopatry (one divergence time).

Dataset	Divergence Time[a]	Likelihood	P-value
P. glaucus/P. canadensis	1	-18.15	0.0004
	6	-6.94	
Human/Chimpanzee	1	-28.93	0.96
	14	-26.91	

[a] The models posit one divergence time for all loci versus a unique divergence time for each locus.

Table 5: Performance of *STE* compared to *MIMAR* using 100 simulated datasets. See Methods for parameter details.

Method	θ_A		T	
	Mean bias[a]	Mean squared error	Mean bias	Mean squared error
STE	1.04	0.038	1.10	0.031
MIMAR	1.009	0.012	1.02	0.017

[a] The bias is the parameter estimates divided by the real value.

Supplementary Table 1: Summary statistics of *TH* in Papilio glaucus and P. canadensis.

Locus	Species	Length (silent sites)	S	π[a] (%)	θ[b] (%)	D_{XY}[c] (%)	Shared	Fixed	TajD[d]	FWH[e]
TH	*P. glaucus*	134	10	2.5	2.8	4.3	0	3	0.45	1.07
	P. canadensis		5	1.3	1.1				-0.78	1.24

[a] Average pair-wise diversity per site
[b] Estimate of $\theta = 4N_e\mu$ per site using the number of polymorphic sites.
[c] Average pairwise divergence per site
[d] Tajima's *D* (Tajima 1983)
[e] Fay and Wu's *H* (Fay and Wu 2000)

Supplementary Table 2: Chimpanzee and Human Summary Statistics

Region	Species	Length	S	π^a (%)	$D_{XY}{}^b$ (%)	TajD[c]	FWH[d]
1	*H. sapiens*	843	5	0.05	0.76	-1.45	-1.54
	P. troglodytes schweinfurthii		5	0.17		0.06	0.24
2	*H. sapiens*	795	5	0.24	0.93	1.38	0.06
	P. troglodytes schweinfurthii		14	0.30		-1.45	1.6
3	*H. sapiens*	1144	8	0.07	1.64	-1.30	0.75
	P. troglodytes schweinfurthii		9	0.22		-0.03	1.24
5	*H. sapiens*	1555	6	0.13	1.36	-0.53	-4.82*
	P. troglodytes schweinfurthii		5	0.08		0.72	-1.14
6	*H. sapiens*	934	4	0.11	1.32	0.10	0.79
	P. troglodytes schweinfurthii		11	0.35		0.20	-0.92
7	*H. sapiens*	1183	8	0.10	1.48	-0.92	-3.63
	P. troglodytes schweinfurthii		5	0.10		-0.59	0.80
8	*H. sapiens*	775	2	0.07	0.71	0.12	0.38
	P. troglodytes schweinfurthii		3	0.10		-0.13	-3.76*
10	*H. sapiens*	781	7	0.13	0.89	-1.21	0.6
	P. troglodytes schweinfurthii		10	0.33		-0.37	0.86
11	*H. sapiens*	613	3	0.16	1.14	0.61	-2.13
	P. troglodytes schweinfurthii		1	0.06		0.72	0.01
14	*H. sapiens*	738	4	0.22	0.62	1.55	-2.49
	P. troglodytes schweinfurthii		5	0.19		-0.31	-0.05
15	*H. sapiens*	786	1	0.02	1.58	0.72	-4.09*
	P. troglodytes schweinfurthii		4	0.07		0.17	-1.44
16	*H. sapiens*	598	6	0.28	2.04	0.31	0.37
	P. troglodytes schweinfurthii		2	0.07		-0.68	1.29*
17	*H. sapiens*	601	5	0.12	1.72	-1.22	-0.26
	P. troglodytes schweinfurthii		3	0.15		0.18	0.63
21	*H. sapiens*	705	6	0.24	1.73	0.31	0.37
	P. troglodytes schweinfurthii		2	0.06		-0.68	1.29
22	*H. sapiens*	568	2	0.09	0.94	-0.02	-0.06
	P. troglodytes schweinfurthii		3	0.18		0.55	0.31
Average	*H. sapiens*	841.27	5.0	0.14	1.26	-0.16	-0.49
	P. troglodytes schweinfurthii		5.5	0.16		-0.13	0.17

[a] Average pair-wise diversity per site
[b] Average pairwise divergence per site
[c] Tajima's *D* (Tajima 1983)
[d] Fay and Wu's *H* (Fay and Wu 2000)
Asterisks denote values of *H* that significantly deviates from neutrality based on neutral simulations with recombination (see Methods).

Figure 1: A model of allopatric speciation showing a gene tree (shaded areas) and a species tree (red and blue lines). An ancestral population of size θ_A splits into two species at time, T.

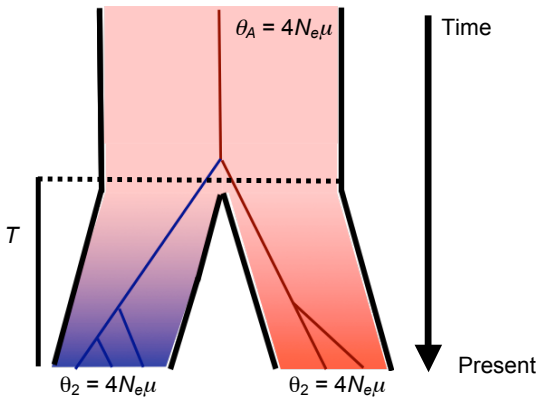


Figure 2: Posterior distributions of jointly estimated ancestral population (N_a) size relative to current population size (N_e) obtained by regression in a) *Papilio glaucus* and *P. canadensis*, and 2) humans and chimpanzees. For parameter details, see Methods.

a) b)

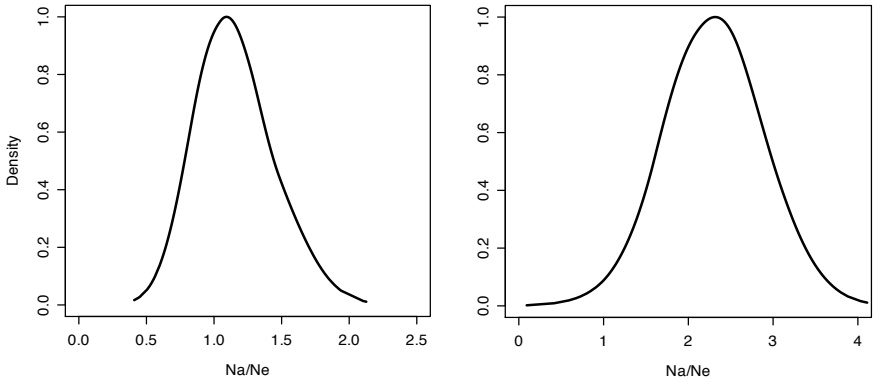
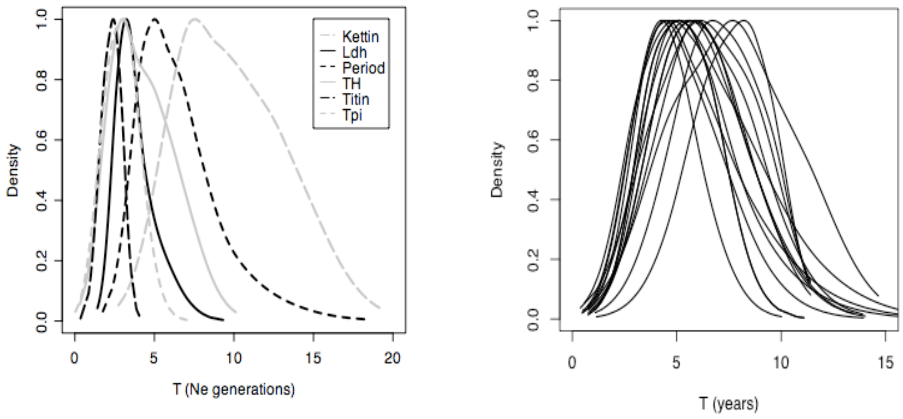
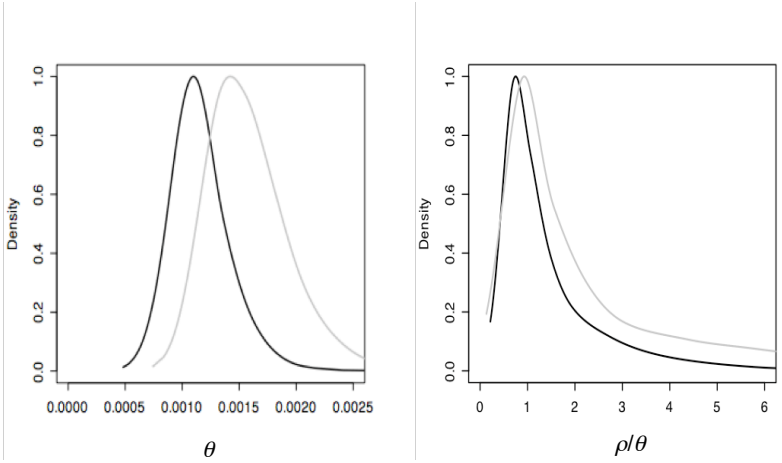


Figure 3: Marginal posterior distributions of per locus divergence time estimates (*T*) obtained by rejection sampling for a) *P. glaucus* and *P. canadensis* and b) human and chimpanzee. For parameter details, see Methods.

a) b)



Supplementary Figure 1: Joint posterior distribution of ρ/θ for humans (black line) and chimpanzees (grey line).



Supplementary Methods and Results:

Comparison to MIMAR *under allopatry*

Methods: We compared our *Papilio* results to the *MIMAR* method of Becquet and Przeworski (2007) under an allopatric model. We set migration to zero. Similar to the methods for estimating performance we estimated parameter values for θ_1, θ_2 were chosen from priors (0, 0.01) per base pair. θ_A and T were also chosen from uninformative priors, (0, 0.01) per base pair and (0, 2 X 10^7) generations, respectively. A mutation rate per base pair is required for *MIMAR*. Because a mutation rate is unknown for *Papilio*, we assumed a mutation rate per generation similar to that of *Drosophila* (1.5 X 10^{-8}), and assumed one generation per year. We recorded 5 X 10^6 total steps and 5 X 10^5 burn-in steps.

Results: The mode of the divergence time distribution is 0.29 MYA (95%CI 0.20 – 0.42). The estimated joint divergence time using STE is 0.35 MYA, using the equation $T_{abs} = T\theta_1/2\mu L$, where μ is the mutation rate per silent site per year (1.5 X 10^{-8}) and L is the number of silent sites. The estimate of the ancestral population size, θ_A, using *MIMAR* ($\theta_A/\theta_1 = 1.23$) is close to the estimate using *STE* ($\theta_A/\theta_1 = 1.09$). The estimates using both methods are quite similar, and the discrepancy may arise from differences in the prior distribution, the assumed mutation rate, or from the MCMC estimation of θ in *MIMAR* compared to point estimates used in *STE*.

Part II.

The effects of demography on selection

Chapter 3.

Influence of demography on detecting positive selection at *Drosophila melanogaster* female reproductive genes

INFLUENCE OF DEMOGRAPHY ON DETECTING POSITIVE SELECTION AT *DROSOPHILA MELANOGASTER* FEMALE REPRODUCTIVE GENES

Andrea S. Putnam, Clinton Edwards, Peter Andolfatto

Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093

Corresponding Author:

Andrea Putnam

Division of Biology

UCSD. MC0116

La Jolla, CA 92093

asputnam@biomail.ucsd.edu

Keywords: adaptive evolution, demography, positive selection, site frequency spectrum, sex-biased genes

ABSTRACT

Male and female reproductive proteins have attracted the attention of evolutionary biologists due to their possible role in reproductive isolation and speciation. In a recent study of a Californian population of *Drosophila melanogaster*, 6 of 9 genes with enriched expression in the female reproductive tract exhibited non-neutral population genetic patterns suggesting they are frequent targets of adaptive evolution (Panhuis and Swanson 2006). However, the Californian population exhibits signatures expected under a recent bottleneck, including a marked reduction in variability, increased linkage disequilibrium and a large variance in polymorphism patterns among loci and bottlenecks are known to produce many false signatures of positive selection We thus surveyed variation at the same candidate genes in a putatively ancestral African population and found that 2 of the 9 genes exhibit signatures of recent positive selection in both populations. In addition, we find that the bottleneck in the California population actually obscures a signature of ongoing positive selection on polymorphic amino acid variants at candidate reproductive genes. These results illustrate that taking demographic history into account can increase the power and accuracy of scans for genes undergoing adaptive evolution.

INTRODUCTION

It has been suggested that reproductive genes are preferential targets for positive selection because interactions between male and female proteins may

be important in driving reproductive isolation (Swanson and Vacuqier 2002). An example of this type of sexual selection is adaptive evolution of genes involved in sperm and egg recognition or sperm competition. As a result, population genetic scans for positive selection frequently focus on genes up-regulated in the reproductive tract, a particularly with male-biased expression, and several studies that have shown that many are rapidly evolving (Civetta and Singh 1995, Singh and Kulathinal 2000, Wu et al 2000, Swanson et al. 2001a, 2001b, Wyckoff et al 2000, Jagadeeshan and Singh 2005, Proschel et al 2006).

For several reasons, *Drosophila* has been an important model system used to study selection on reproductive genes. In addition to morphological and behavioral sexual dimorphism in *Drosophila*, gene expression is also sexually dimorphic. Surveys of sex biased gene expression found that approximately half the expression of *D. melanogaster* genes is sex-dependent (Ranz et al 2003, Parisi et al 2004, Zhang et al 2007). For example, accessory gland proteins in the seminal fluid are known to evolve rapidly in many *Drosophila* species (Tsaur and Wu 1997, Begun et al 2000, Swanson 2001, Betrán and Long 2003, Schully and Hellberg 2006). In mated females, these proteins mediate sperm competition and storage as well as increase the rate of egg laying (Chapman et al 1995, Clark et al 1995) making them candidates for sexual selection and reproductive isolation factors.

There are, however, a paucity of studies that examine sexual selection in female reproductive genes. In search of candidates for positive selection in female

reproductive proteins, Panhuis and Swanson (2006) recently examined 8 genes with nonsynonyomous to synonymous substitution ratios (Ka/Ks) > 0.5 identified from a set of *D. simulans* genes with enriched expression in the female reproductive tract (Swanson et al 2004). An additional candidate gene was also examined based on its high expression in female spermatheca and parovaria tissue (Arbeitman et al 2004). Together, these genes were surveyed for evidence of positive selection in Californian isofemale lines and 6 of the 9 suggested positive selection based on patterns of polymorphism and divergence (Panhuis and Swanson 2006).

However, *Drosophila melanogaster* has a Sub-Saharan African origin and is believed to have only recently colonized temperate habitats (Lachaise et al. 1988). Recent population bottlenecks associated with colonization of these temperate habitats are apparent in reduced levels of nucleotide variability, a skew towards high frequency-derived variants and increased linkage disequilibrium in non-African compared to African populations (Begun et al. 1993; Andolfatto 2001; Ometto et al. 2005; Haddrill et al. 2005; Thornton and Andolfatto 2006; Li and Stephan 2006). This recent bottleneck is expected to be associated with many false positives in genome wide scans for recent positive selection (Thornton and Andolfatto 2006, Thornton and Jensen 2006). To distinguish effects of demography from true signatures of selection at candidate female reproductive genes, we compare polymorphism patterns at the 9 candidate loci in samples from California to the same genes sampled in the putatively ancestral

Zimbabwe population. We also take advantage of a set of 137 randomly chosen genes sequenced in a Zimbabwe population as a control. If selection on these female reproductive genes is not geographically localized, we expect to see similar evidence for selection on these candidates in both populations.

MATERIALS AND METHODS

Candidate loci. The nine female reproductive loci surveyed here are described in Swanson et al. (2004) and Panhuis and Swanson (2006) and show enriched expression in the *D. simulans* female reproductive tract. With the exception of CG17012, these genes were predicted to be likely targets of positive selection because of Ka/Ks values are > 0.5. CG17012 was chosen specifically because it is highly expressed in the female reproductive tract of *D. melanogaster* (Arbeitman et al. 2004). These loci are scattered throughout the genome (Table 1). The 13 *Drosophila melanogaster* fly lines used in this study come from Victoria Falls, Zimbabwe, Africa (collected by B. Ballard, 2002). In order to amplify a single *D. melanogaster* allele from autosomal loci, ~5 virgin females from each *D. melanogaster* line were crossed to ~10 *D. simulans* males from a Madagascar population (collected by B. Ballard 2002). Primers were designed to specifically PCR amplify the *D. melanogaster* allele in all-female F1 offspring of these crosses (Ashburner 1989). To examine divergence, we used orthologous *D. simulans* sequences from Release 1.0 of the *D. simulans* genome assembly (Accession number AAGH01000000). Information on the primers used is listed in

Supplementary Table 1. Sites within insertion-deletion events or ambiguous bases (there were a total of 4 ambiguous bases out of 5712 total bases) in alignments were not considered in analyses.

Genomic DNA was extracted using the Puregene DNA purification kit (Gentra Systems. PCR products were cleaned using Exonuclease I and Shrimp Alkaline Phosphatase. Both strands of PCR product were sequenced using Big-Dye (Version 3, Applied Biosystems) and run on an ABI 3730 capillary sequencer. Sequences were edited and aligned using Sequencher 4.2 (Gene Codes) software. Sequences are deposited in Genbank (Accession numbers _____ - _____).

Control loci. For comparison to candidate loci, similar polymorphism and divergence data for a control set of 137 randomly chosen X-linked loci was used (Andolfatto 2007). This set of control genes was selected randomly from highly recombining regions of the X-chromosome. Sex biased expression for the candidate and 137 control genes was determined data obtained from ovaries and testes (Parisi et al 2004, Gnad and Parsch 2006). For 29 genes there was no available ovaries or testes expression data (Gnad and Parsch 2006).

Summaries of polymorphism and divergence, and tests of neutrality. Panhuis and Swanson (2004) considered all surveyed polymorphisms in analyses. Here we analyzed nonsynonymous and synonymous polymorphism separately except for Fay and Wu's *H* (Fay and Wu 2000), the composite likelihood ratio test (CLR

test; Kim and Stephan 2002), and the selective sweep goodness-of-fit (GOF) test (Jensen et al 2005). There were two discrepancies between our *H* statistic results and those obtained by in Panhuis and Swanson (2006). In our analysis *CG13004* was significant for *H*, but in Panhuis and Swanson (2006) it was not, and *CG5843* was significant for *H* in their analysis and in ours it was not. These discrepancies in *H* values are due to differences in alignments with the outgroup, *D. simulans*. For all other California candidate genes, our values for *H* are similar to those reported in Swanson and Panhuis (2006). To make the two studies comparable, we sub-sampled 12 alleles from the Californian population to match the average sample size of our study.

Levels of neutral variation were summarized using Watterson's estimator, θ (Watterson 1975), and the average pairwise diversity per nucleotide, π (Tajima 1983). Synonymous and nonsynonymous sites were counted using the method of Nei and Gojobori (1986). Divergence between *D. melanogaster* and *D. simulans* was estimated as the average pairwise divergence per site between species, D_{XY} (Nei 1987). To compare the degree of variation in diversity levels among loci for the Zimbabwe and California samples, we used the coefficient of variation (CV = standard deviation/mean) of π. The number of divergent sites (*D*) was estimated as D_{XY} - π with a Jukes-Cantor correction for multiple hits (Li 1997). Two summaries of the distribution of polymorphism frequencies were used. Tajima's *D* (Tajima 1989) is a measure of the standardized difference between π and θ. Fay and Wu's *H* measures the difference between π and θ_H, an

estimator of θ that weights derived variants by the square of their frequencies (Fu 1996; Fay and Wu 2000). For each locus, the level of linkage disequilibrium was assessed using the *ZnS* statistic (Kelly 1997), which is the average pairwise r^2 (Hill and Robertson 1968) among polymorphic sites. The program *msstats* (Thornton 2003, http://molpopgen.org/software/msstats/) was used to calculate *ZnS.*

Comparing levels of polymorphism and divergence at candidate and control genes. A multilocus HKA test (Hudson et al. 1987; Haddrill et al 2005) was used to quantify levels of heterogeneity in levels of polymorphism and divergence at synonymous sites in candidate and control loci compared to neutral expectations. In addition, a two-class version of the multilocus HKA test (Andolfatto 2005) was implemented to compare levels of pooled polymorphism and divergence at Zimbabwe candidate loci the 137 control genes. Code for the multilocus HKA test is available by contacting the authors, and code for the two-locus version of the test (Andolfatto 2005) is available on request to P.A. Empirically the effective population size of the X and autosomes look close to equal in African populations (Andolfatto 2001; Hutter et al. 2007), thus we did not scale diversity on the X to correct for expected differences in chromosome number compared to autosomes (i.e. 3X:4A in a population with equal numbers of males and females). The site frequency spectrum (SFS) of polymorphisms was compared among datasets to examine signatures of selection. Each polymorphism is classified according to its derived frequency from 1 to *n*-1, where *n* is the sample size. The derived

ancestral state was inferred using *D. simulans* as an outgroup and standard parsimony criteria. A test for recent selective sweeps at the candidate loci was performed using a composite-likelihood ratio test (CLRT, Kim and Stephan 2002). For loci where neutrality is rejected, a goodness of fit (GOF) test is then performed to determine if the data is consistent with a selective sweep model (Jensen et al 2005). We attempt to distinguish true sweeps from false positives due to a population bottleneck by simulating the distribution of the GOF statistic under a demographic model. To model demography for the California population, we used “out-of-Africa” bottleneck parameters estimated in Thornton and Andolfatto (2006).

We also used the SFS, pooled across loci, to compare classes of polymorphisms (ie. synonymous and nonsynonymous). For each frequency class we performed a correction for multiple hits to a nucleotide site because multiple hits will lead to more inferred high frequency derived polymorphisms, a signature of positive selection. Ignoring higher order terms, the probability of misinferring the ancestral state $k \sim \kappa(Dxy - \pi)$, where κ = probability of misinference given a back mutation and $(Dxy - \pi)$ is the net divergence per site for a locus. We assume $\kappa \sim 3/8$, which is appropriate for four-fold synonymous sites (Fay and Wu 2000). We model the observed number of polymorphisms in frequency class *i* ($S_{i,obs}$) as reflecting an equilibrium between mutation and back-mutation. Specifically, the observed polymorphisms at frequency *i* in a sample of *n* chromosomes is

$$S_{i,obs} = E(S_i) - E(S_i)(k) + E(S_{n-i})(k) \quad (1)$$

where $E(S_i)$ is the expected number of polymorphisms belonging to class *i*. One can write an analogous equation for the observed number of polymorphisms in class (*n* -*i*) as,

$$S_{n-i,obs} = E(S_{n-i}) - E(S_{n-i})(k) + E(S_i)(k) \quad (2)$$

Equation (2) can be substituted into equation (1) and then solved for $E(S_i)$,

$$E(S_i) = \frac{S_{i,obs}(1-k) - S_{n-i,obs}k}{(1-2k)} \quad (3)$$

After the correction for multiple hits, a χ^2 test was then performed to detect departures from expectations in high and low frequency classes in the California and Zimbabwe datasets relative to neutral expectations. For the Zimbabwe dataset, pooled polymorphism frequencies at candidate loci were also compared to control loci.

To test the fit of various test statistics to the standard neutral model, we carried out neutral coalescent simulations with recombination using the program, *ms* (Hudson 2002). Simulations were generally performed using a point estimate of θ (Watterson 1975). For the HKA test, we used the expected θ for each locus and T calculated using equations in Hudson et al. (1987). *P*-values for test statistics are generally based on 10,000 simulated replicates. The distributions of *P*-values between candidates and control loci for *D*, *H*, were compared among datasets

using a Kolmogorov-Smirnov test as implemented using the R statistical package (http://www.r-project.org).

The McDonald-Kreitman (MK; McDonald and Kreitman 1991) test and its extensions are used to test for recurrent adaptive protein evolution in candidate and control genes. In single locus and pooled MK tests, a two-tailed Fisher's Exact test for independence was used to assess significance. Distributions of *P*-values for candidate and control genes were compared using a Kolmogorov-Smirnov test (as above). For this test, a random subset (n = 12) of California individuals were surveyed in order to make all datasets comparable. The fraction of protein divergence driven to fixation by positive selection, α, can be estimated using an extension of the MK test (Rand and Kahn 1996; Fay et al 2001, Fay et al 2002, Smith and Eyre-Walker 2002). To estimate α, we used the methods of Smith and Eyre-Walker (2002) and Bierne and Eyre-Walker (2004) as implemented in the program DFoE (Eyre-Walker, personal communication). Since many nonsynonymous polymorphisms that segregate in natural populations are likely deleterious, these approaches are expected to underestimate α (Fay et al. 2001; Charlesworth and Eyre-Walker 2006). To minimize this bias, low frequency polymorphisms (frequency < 0.10) were excluded from the calculation (Fay et al 2001, Charlesworth and Eyre-Walker 2006).

RESULTS and DISCUSSION

Evidence for a recent bottleneck in the California population.

Signatures of a recent bottleneck include reduced variation (Maruyama and Fuerst 1985), increased linkage disequilibrium (McVean 2002, Haddrill et al 2005), and increased variance in summary statistics such as the HKA χ^2 (Haddrill et al. 2005), Tajima's *D* and Fay and Wu's *H* (Haddrill et al. 2005). Indeed, candidate loci from the Californian population exhibit many signatures consistent with a recent bottleneck (see for e.g. Haddrill et al. 2005). In particular, candidate genes in California have reduced variability and increased levels of linkage disequilibrium (measured by ZnS), and increased heterogeneity in diversity (as measured by CV(π) and the HKA χ^2) relative to the same loci in the Zimbabwe population (Table 1). In addition, the variances of Tajima's *D* and Fay and Wu's *H* are larger for the Californian population (Table 1), as expected after a recent bottleneck (Haddrill et al. 2005). While these differences between the Californian samples could be due to geographically localized selection (in California), we think this explanation is less parsimonious than a simple bottleneck, which is also apparent at randomly chosen loci in non-African populations (Baudry et al. 2004; Ometto et al. 2004; Haddrill et al 2005). These bottleneck signatures will likely lead to false rejections of the neutral model (Thornton and Andolfatto 2006; Jensen and Thornton 2006) and potentially obscure true signals of positive selection among the candidate genes.

Assessing evidence for selection at candidate loci

Which if any of the candidate genes are targets of recent adaptation? We propose to first address this question by comparing locus-by-locus tests of neutrality on the candidate loci in the two populations. In our analysis of the California sample, *CG17012* was significantly different from neutral expectations by a McDonald-Kreitman (MK) test (McDonald and Kreitman 1991), however no individual loci reject the MK test in the Zimbabwe population. We further found that 5 of the 9 candidate loci in the Californian population significantly reject neutrality by either Tajima's *D*, Fay and Wu's *H* or the CLR test (Table 2). At individual candidate loci in the Zimbabwe population, values of Tajima's *D* do not significantly deviate from neutral expectations. However, moderately significant values of *H* suggest an excess of high-frequency derived polymorphisms at 2 of the 9 candidate loci (*CG5106* and *CG17108*) in the Zimbabwe population (Table 2). One of these two loci (*CG17108*) rejects the neutral model by a CLR test (*P* = 0.01) and appears to be consistent with a recent selective sweep model (*P* = 0.47, GOF test, Table 2). Intriguingly this same gene is also the only candidate in the California population to fit a selective sweep model when a bottleneck is accounted for (*P* =0.95, GOF test, Table 2),suggesting it may truly be a target of recent adaptation. There is little information known about the ontology of *CG17108* other than it appears to have acetyl-CoA carboxylase activity (Ashburner et al 2000), which is important in fat metabolism.

While it appears that at least some are targets of recent selection, is there any evidence that candidate female reproductive genes are preferential targets for

adaptation? To address this question, we compared the 9 candidate genes surveyed in Zimbabwe to the 137 control loci surveyed in the same population (Andolfatto 2007). Sixteen of these 137 control loci (12%) and 2 of the 9 Zimbabwe candidates (22%) reject neutrality (P = 0.3 by a Fisher's Exact test). We also failed to find a significant difference in the distribution of P-values for D, H and the CLR test for the Zimbabwe candidate and control sets (P = 0.81, P = 0.46, P= 0.55, respectively, Kolmogorov-Smirnov test, Figure 1). In addition, comparing the distribution of MK test P-values for the candidate and control loci failed to detect a difference in their distributions (P =0.96, Kolmogorov-Smirnov test; Figure 1). Finally, we tested for a difference in levels of the level of polymorphism relative to divergence in candidates and controls using a two-class HKA test (χ^2 = 1.73, P = 0.19). These results together suggest that, while we do detect recent selection at two candidates, candidates overall do not look particularly unusual relative to the genome background.

To evaluate this further, we investigated the extent of adaptive protein evolution in candidate versus control loci using α, an index that measures the fraction of adaptive protein divergence (Rand and Kann 1996; Fay et al 2001, Fay et al. 2002, Smith and Eyre-Walker 2002). Using two methods of calculating α (Table 3), we found that estimates of α (46-62%) were similar in candidates and controls and to previous estimates of α in *D. melanogaster-D. simulans* comparisons (Andolfatto 2007).

Evidence for ongoing positive selection on female reproductive genes in the site frequency spectrum.

The tests for selection and estimates of adaptive divergence above suggest that the Zimbabwe female reproductive candidate genes largely do not look different than the genomic background. To try a different approach, we examined the polymorphism site frequency spectrum (SFS), which can be used to tease apart signatures of negative and positive selection (Nielsen 2005). In particular, negative selection on nonsynonymous sites will skew the SFS toward low frequency polymorphisms. Alternatively, positive selection will skew the SFS toward high frequency derived polymorphisms that are on their way to fixation in the population. Using the SFS to evaluate evidence for selection, however, can be misleading because demography (Przeworski 2002, Lazzaro and Clark 2003, Haddrill et al 2005) and genetic hitchhiking (Fay and Wu 2000, Kim and Stephan 2000, Przeworski 2002) can cause shifts in SFS toward both high and low frequency polymorphisms. To avoid these problems, we compare two classes of polymorphic sites, synonymous and nonsynonymous. When evaluating patterns at nonsynonymous sites, synonymous sites are used as a control since demography and hitchhiking are expected to affect both classes of sites similarly.

To look for signatures of direct selection on nonsynonymous polymorphisms, we compared the SFS of nonsynonymous polymorphisms in the Zimbabwe samples for candidate and control loci to synonymous sites (Figure 2). At nonsynonymous sites for control loci in the Zimbabwe population, there is a significant excess of

rare nonsynonymous polymorphisms relative to synonymous sites ($P = 3^{-7}$, Fisher's Exact test) consistent with the action of purifying selection on amino acid polymorphisms.

Intriguingly, there is also a significant excess of high frequency derived nonsynonymous polymorphisms in the Zimbabwe candidate genes relative to synonymous polymorphisms (P=0.01, Fisher's Exact test, Figure 2). Two loci at which we detected positive selection in locus-by-locus tests, *CG5106* and *CG17108,* did not contribute disproportionately to the total number of high frequency derived nonsynonymous polymorphisms (i.e. 38%), and this significant difference persists when these loci are excluded (P=0.001, Fisher's Exact test, Figure 2), The same excess of high frequency derived nonsynonymous polymorphisms is not apparent in control loci (P=0.1, Fisher's Exact test, Figure 2), and in fact there is a significant difference in the SFS for candidate and control loci in the Zimbabwe sample (χ^2 =12.2, P=0.002, Figure 2). Together, these patterns suggest that not only are a significant fraction of amino acid polymorphisms in candidate loci experiencing positive selection, but candidate loci also appear to be preferential targets for positive selection by this analysis.

Despite the apparent excess of low frequency nonsynonymous polymorphisms compared to synonymous polymorphisms in the Californian sample of candidate genes (Figure 2), this difference is actually not significant (P = 0.14, Fisher's Exact test). We also found that there are not more high frequency derived

nonsynonymous polymorphisms compared to synonymous sites in the Californian sample (P = 0.48, Fisher's Exact test). We propose that signatures of selection on candidate genes in the Californian population may be more difficult to see because there are fewer polymorphisms and polymorphism frequencies may have been recently perturbed by a recent bottleneck. Thus, surveying the Zimbabwe population may have afforded us more statistical power to detect selection.

Analyzing sex bias in the control and candidate genes

Our investigation of candidate female reproductive genes, following the study of Swanson and Panhuis (2006), has affinities with recent studies investigating rates of adaptation in the context of sex-differences in gene expression (Proeschel et al. 2006; Sawyer et al. 2007). In particular, Proeschel et al (2006) documented higher levels of adaptive divergence in genes with male- and female-biased expression relative to genes with unbiased expression (but see Sawyer et al. 2007). Though our set of candidate genes is targeted towards female reproduction, and one gene (*CG17012*) was chosen specifically because it is highly expressed in the female reproductive tract, in fact only 4 of these 9 candidates have female-biased expression according to the SEBIDA database (Gnad and Parsch 2006). Interestingly, both candidate genes we detected as targets of recent positive selection in Zimbabwe in locus-by-locus tests (*CG5106* and *CG17108)* actually have male-biased expression.

Because rates of adaptation seem to correlate with sex-specific expression pattern, and encouraged by our analysis of the SFS comparing candidates and controls, we revisited our combined dataset in the context of sex-biased expression (Gnad and Parsch 2006). We separated the 137 control and 9 candidate genes from Zimbabwe into male biased (N =20), female biased (N = 23) and unbiased (N = 70) expression classes based on data from *D. melanogaster* ovaries and testes (Parisi et al 2004, Gnad and Parsch 2006). Thirty-three surveyed genes did not have available expression data.

In line with previous studies (Pröeschel et al 2006), we confirm that the fraction of positively selected amino acid replacements between species, α, is greater in male- and female-biased genes compared to unbiased genes (Figure 3). However, only the maximum-likelihood estimate of α for male-biased genes is outside of the 95% confidence interval for unbiased genes (P = 0.022, by a likelihood ratio test). Interestingly, we found that the distributions of P-values for Fay and Wu's H are significantly different between male and female biased classes (P = 0.02, Kolmogorov-Smirnov test, Supplementary Figure 1). This is likely because the Fay and Wu's H values were generally more negative overall in the male-biased (mean H = -0.55) compared to female-biased genes (mean H = -0.07).

To examine this further, we considered the SFS for each of these three classes of genes. There is a significant excess of rare nonsynonymous polymorphisms relative to synonymous sites for male (P = 0.05, Fisher's Exact test, Figure 4),

female (P = 0.01) and non- sex biased P = (0.004) genes. This result is consistent with the action of purifying selection on amino acid polymorphisms. At high frequency nonsynonymous polymorphisms, only female biased genes were significantly different from synonymous (*P* = 0.003). Although high frequency polymorphisms in male biased genes did not differ significantly between synonymous and nonsynonymous sites, at synonymous high frequency sites the male biased genes were significantly different from neutral (*P* =0.01), female (*P* = 0.01) and unbiased sites (*P* = 0.03). This excess of high frequency polymorphisms is not seen at other expression classes. Together, the SFS results and the significant male biased Fay and Wu's *H P*-values compared to female biased are consistent with the notion that genes with male-biased expression may be more frequent targets of adaptation (Proeschel et al. 2006).

Conclusions

We compared polymorphism and divergence patterns at 9 candidate female reproductive genes surveyed in a putatively ancestral African and recently derived Californian population. Though we confirmed that the a large fraction of these candidates depart from neutral expectations in the Californian sample, we propose that many of these departures from neutrality likely result from a recent bottleneck based on caparisons between Zimbabwe and Californian samples. Despite a large impact of demography in the Californian population, we did detect evidence for recent positive selection at 2 of the 9 candidate loci in the Zimbabwe sample, one of which showing strong support for a recent selective

sweep in both populations. Given evidence for recent positive selection at some candidates, we then asked whether the candidates are preferential targets for adaptation compared to a control set of 137 loci surveyed in the same Zimbabwe population. We found that the candidate female reproductive genes indeed seem to be enriched for positively selected amino acid polymorphisms, and that this signature of positive selection is most easily seen in the relatively demographically stable Zimbabwe population. These results, in combination with results for locus-by-locus tests for selection suggest that taking the demographic history of populations into account may increase the power and accuracy of tests for selection.

It is also important to note that our candidate 'female' reproductive genes that appear to be under selection actually have male biased expresion. Tests of neutrality and the site frequency spectrum of the control and candidate genes partitioned into sex biased classes revealed that indeed, male biased genes appear to be under greater positive selection that female and non- sex biased genes. These results point to genes with male biased expression driving the signatures for positive selection.

AKNOWLEDGEMENTS

The authors appreciate comments and discussions with Jeff Jensen, Kevin

Thornton, Fedya Kondrashov, and Doris Bachtrog.

REFERENCES

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

Andolfatto P (2001) Adaptive hitchhiking effects on genome variability. Curr Opin Genet Dev 11(6): 635-641.

Andolfatto P, Przeworski M (2001) Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. Genetics 158: 657 665.

Andolfatto P (2005) Adaptive evolution of non-coding DNA in Drosophila. Nature 437: 1149-1153.

Andolfatto P (2007) Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome. Gen Res 17: 1755-62.

Arbeitman MN, Fleming A A, Siegal ML, Null BH, Baker BS (2004) A genomic analysis of *Drosophila* somatic sexual differentiation and its regulation. Development 131: 2007-2021.

Ashburner M (1989) *Drosophila*: A Laboratory Handbook. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Ashburner M, Ball CA, Blake, JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. Nature Genetics 25(1): 25-9.

Begun DJ, Aquadro CF (1993) African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature 365: 548-550.

Begun DJ, Whitley P, Todd BL, Waldrip-Dail HM, Clark AG (2000) Molecular population genetics of male accessory gland proteins in *Drosophila*. Genetics

156: 1879–1888.

Betrán E, Long M (2003) *Dntf-2r*, a young *Drosophila* retroposed gene with specific male expression under positive Darwinian selection Genetics 164: 977-988.

Bierne N, Eyre-Walker A (2004) The genomic rate of adaptive amino acid substitution in *Drosophila*. Mol Biol Evol 21: 1350 - 1360.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140: 783–796

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22(3): 231-8.

Chapman T, Liddle LF, JKalb JM, Wolfner MF, Partridge L (1995) Cost of mating in *Drosophila* melanogaster females is mediated by male accessory gland products. Nature 373: 241–244.

Charlesworth B (1996) Background selection and patterns of genetic diversity in *Drosophila melanogaster* Genet Res 68: 131-149.

Charlesworth J, Eyre-Walker A (2006) The rate of adaptive evolution in enteric bacteria. Mol Biol Evol 23: 71348–1356.

Civetta A, Singh RS (1995) High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. J Mol Evol 41: 1085–1095.

Clark AG, Aguadé M, Prout T, Harshman LG, Langley CH (1995) Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila* melanogaster. Genetics 139:189–201

Comeron J, Kreitman M (2000) The correlation between intron length and recombination in *Drosophila*. Dynamic equilibrium between mutational and selective forces. Genetics 156: 1175-1190.

Fay J, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.

Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human

genome. Genetics 158: 1227-1234.

Fay JC, Wyckoff GJ, Wu CI (2002) Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature 415: 1024-1026.

Fu YX (1996) New statistical tests of neutrality for DNA samples from a population. Genetics, 143: 557–570.

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147: 915–925.

Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D (2003) Demography and natural selection have shaped genetic variation in Drosophila melanogaster: a multi-locus approach. Genetics 165: 1269–1278.

Gnad F, Parsch J (2006) Sebida: a database for the functional and evolutionary analysis of genes with sex-biased expression. Bioinformatics 22: 2577-2579.

Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, Charlesworth B, Keightley PD (2007) Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. Nature 445(7123): 82-5.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P (2005) Multilocus pattern of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations. Genome Res 15:790–799.

Haerty W, Jagadeeshan WS, Kulathinal RJ, Wong A, Ram KR, Sirot LK, Levesque L, Artieri CG, Wolfner MF, Civetta A, Singh RS (2007) Evolution in the fast lane: rapidly evolving sex-and reproduction-related genes in *Drosophila* species. Genetics 177:1321–1335.

Halligan DL, Eyre-Walker A, Andolfatto P, Keightley PD (2004) Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. Genome Res 14(2): 273-279.

Hey J, Harris E (1999) Population bottlenecks and patterns of human polymorphism. Mol. Biol. Evol. 16:1423-1426

Hill WG, Robertson A, (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38: 226–231.

Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153–159.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of

genetic variation. Bioinformatics 18: 337-338.

Jagadeeshan S, Singh RS (2005) Rapidly evolving genes of *Drosophila*: differing levels of selective pressure in testis, ovary, and head tissues between sibling species. Mol Biol Evol 22: 1793–801.

Jensen JD, Kim, Bauer DuMont V, Aquadro CF, Bustamante CD (2005) Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170: 1401-1410.

Jukes TH, Cantor C (1969) Evolution of protein molecules. In: *Mammalian protein metabolism* (ed. MN Munro), Academic Press, New York.

Kauer MO, Dieringer D, Schlotterer C (2003) A microsatellite variability screen for positive selection associated with the "Out of Africa" habitat expansion of *Drosophila melanogaster*. Genetics 165: 1137–1148.

Kelly JK (1997) A test of neutrality based on interlocus associations. Genetics 146: 1197-1206.

Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res 16: 980–989.

Kim Y (2004) Effect of strong directional selection on weakly selected mutations at linked sites: implication for synonymous codon usage. Mol Biol Evol 21: 286-94.

Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765-777.

Lachaise D, Cairou ML, David JR, Lemeunier F, Tsacas L, Ashburner M (1988) Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol Biol 22: 159-226.

Lazzaro BP, Clark AG (2003) Molecular population genetics of inducible antibacterial peptide genes in *Drosophila melanogaster*. Mol Biol Evol 20: 914–923.

Li H (1997) *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Li H, Stephan W (2006) Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet. 2: e166.

Maruyama T, Fuerst PA (1985) Population bottlenecks and non-equilibrium

models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. Genetics 111: 675 689 .

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351: 652-654.

McVean GA, Vieira J (2001) Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. Genetics 157: 245-257.

McVean GA (2002) A genealogical interpretation of linkage disequilibrium. Genetics 162(2): 987-991.

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York.

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418 - 426.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nature Reviews Genetics 8(11): 857.

Ometto L, Glinka S, De Lorenzo D, Stephan W (2005) Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol Biol Evol 22: 2119–2130.

Orengo DJ, Aguadé M. (2004) Detecting the footprint of positive selection in a European population of *Drosophila melanogaster*: multilocus pattern of variation and distance to coding regions. Genetics 167:1759–1766

Panhuis TM, Swanson WJ (2006). Molecular evolution and population genetic analysis of candidate female reproductive genes in *Drosophila*. Genetics 173 (4): 2039-2047.

Parisi M, Nuttall R, Edwards P, Minor J, Naiman D, Lü J, Doctolero M, Vainer M, Chan C, Malley J, Eastman S, Oliver B (2004) A survey of ovary-, testis-, and soma-biased gene expression in *Drosophila melanogaster* adults. Genome Biol. 5: R40.

Pröschel M, Zhang Z, Parsch J. (2006) Widespread adaptive evolution of *Drosophila* genes with sex-biased expression. Genetics 174: 893–900.

Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 160: 1179-1189.

Przeworski M, Wall JD, Andolfatto P (2001) Recombination and the frequency spectrum in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. 18: 291-298.

Rand DM, Kann LM (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. Mol Biol Evol 13(6): 735-48.

Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL (2003) Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. Science 300: 1742-1745.

Sawyer S, Parsch J, Zhang Z, Hartl D (2007) Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. Proc. Natl. Acad. Sci. 104: 6504–6510.

Schully SD, Hellberg ME (2006) Positive selection on nucleotide substitutions and indels in accessory gland proteins of the *Drosophila pseudoobscura* subgroup. J Mol Evol 62: 793–802.

Singh RS, Kulathinal RJ (2000) Sex gene pool evolution and speciation: a new paradigm. Genes Genet. Syst. 75: 119-130.

Smith NG, Eyre-Walker A (2002) Adaptive protein evolution in *Drosophila*. Nature 415: 1022-1024.

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF (2001a) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc Natl Acad Sci USA 98: 7375-7379.

Swanson WJ, Yang Z, Wolfner MF, Aquadro CF (2001b) Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc Natl Acad Sci USA 98:2509–14.

Swanson WJ, Vacquier VD. (2002) The rapid evolution of reproductive proteins. Nat Rev Genet 3: 137–44.

Swanson WJ, Wong A, Wolfner MF, Aquadro CF (2004) Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. Genetics 168: 1457-1465.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations.

Genetics 105:437-460.

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

Thornton K (2003) Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics 19: 2325–2327.

Thornton KR, Andolfatto P (2006) Approximate Bayesian inference reveals evidence for a recent, severe, bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics 172: 1607-1619.

Thornton K, Bachtrog D, Andolfatto P (2006) X chromosomes and autosomes evolve at similar rates in *Drosophila*: No evidence for faster-X protein evolution. Genome Res16(4): 498 – 504.

Thornton K, Jensen J (2007) Controlling the false-positive rate in multilocus genome scans for selection. Genetics 175: 737–750.

Tsaur S-C, Wu CI (1997) Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of *Drosophila*. Mol Biol Evol 14: 544–9.

Vigouroux, Y, McMullen M, Hittinger CT, Houchins K, Schulz L, Kresovich S, Matsuoka Y, Doebley J (2002) Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. Proc Nat Acad Sci USA 99: 9650–9655.

Watterson G (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7: 256-276.

Wu C-I (2000) Genetics of species differentiation: What is unknown and what will be unknowable? Evol. Biol. 32: 239 - 248.

Wyckoff GJ, Wang W, Wu CI (2000) Rapid evolution of male reproductive genes in the descent of man. Nature 403, 304–309.

Zhang Y, Sturgill D, Parisi M, Kumar S, Oliver B (2007) Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. Nature 450: 233–237

Figures and Tables

Table 1: Tests of neutrality using the Zimbabwe (Zimb) dataset and a subset of 12 individuals from the California (Calif) dataset. P-values are in parentheses: average (Ave) and variance (Var) Tajima's D (1983), Fay and Wu's H (2000), and HKA χ^2 statistic (Hudson et al 1987).Significance was based on simulations under the neutral model with recombination. See *Methods* for simulation parameter estimates.

Population	Ave *D*	Ave *H*	Var *D*	Var *H*	HKA	CV
California Candidate	-0.58 (0.14)	-0.54 (0.09)	0.35 (0.17)	1.02 (0.014)	19.11 (0.04)	0.87
Zimbabwe Candidate	-0.63 (0.06)	-0.34 (0.15)	0.22 (0.13)	1.13 (<0.001)	7.14 (0.75)	0.70
Control (Zimbabwe)	-0.28 (0.10)	-0.55 (0.08)	0.45 (0.94)	3.29 (<0.001)	22.0 (0.82)	0.54

Table 2. *P*-values for tests of neutrality and selective sweeps in California and Zimbabwe candidate genes. Significance was based on simulations under the neutral model with recombination. See *Methods* for simulation parameter estimates.

Locus	*D*[a]		*H*[b]		LR[c]		GOF[d]	
	Zimb	Calif	Zimb	Calif	Zimb	Calif	Zimb	Calif
CG5273	0.13	0.01*	0.38	0.76	0.52	0.44	N/A	N/A
CG13004	0.47	0.34	0.10	0.01*	0.67	0.02*	N/A	0.02*
CG8453	0.10	0.58	0.74	0.06	0.43	0.55	N/A	N/A
CG9897	0.85	0.31	0.51	0.05*	0.65	0.06	N/A	N/A
CG10200	0.39	0.12	0.33	0.10	0.08	0.68	N/A	N/A
CG17012	0.40	0.72	0.19	0.07	0.50	0.42	N/A	N/A
CG17108	0.19	0.04*	0.05*	0.02*	0.01*	0.002*	0.47	0.95
CG5106	0.72	0.04*	0.04*	0.004*	0.74	0.20	N/A	N/A
CG5976	0.80	0.32	0.53	0.97	0.52	0.11	N/A	N/A

*Significant after a Bonferroni correction for multiple tests.
[a] Tajima's D (1983). [b] Fay and Wu's H (2000) [c]Composite likelihood ratio test (Kim and Stephan 2005) [d] Goodness of fit test (GOF) for selective sweeps (Jensen et al. 2005) that account for bottleneck (Thornton and Andolfatto 2006).

Table 3: Estimates of the fraction of adaptive amino acid divergence, α, in candidate genes from Zimbabwe (Zimb) and and 137 control genes from Zimbabwe. 95% confidence intervals in parentheses.

Polymorphisms	SEW[a]		BEW[b]	
	Zimb	Control	Zimb	Control
All	0.49 (0.24 - 0.66)	0.54 (0.44 - 0.63)	0.42 (0.10 - 0.63)	0.40 (0.30 - 0.48)
p >0.1	0.48 (0.15 - 0.74)	0.62 (0.49 - 0.72)	0.50 (0.16 - 0.70)	0.52 (0.40 - 0.59)
p >0.2	0.46 (0.18 -0.71)	0.43 (0.24 - 0.59)	0.35 (-0.35 - 0.66)	0.40 (0.22- 0.59)

[a] (Smith and Eyre-Walker 2002); [b] (Bierne and Eyre-Walker 2004).
[c] A subset of 12 California individuals used for analyses to make all the datasets comparable.

Figure 1a-d: Tests of neutrality. Distribution of *P*-values for a) Fay and Wu's *H* and, b) Tajima's *D*, c) CLR test, and d) MK test in Zimbabwe candidate, California candidate, and Zimbabwe control genes.

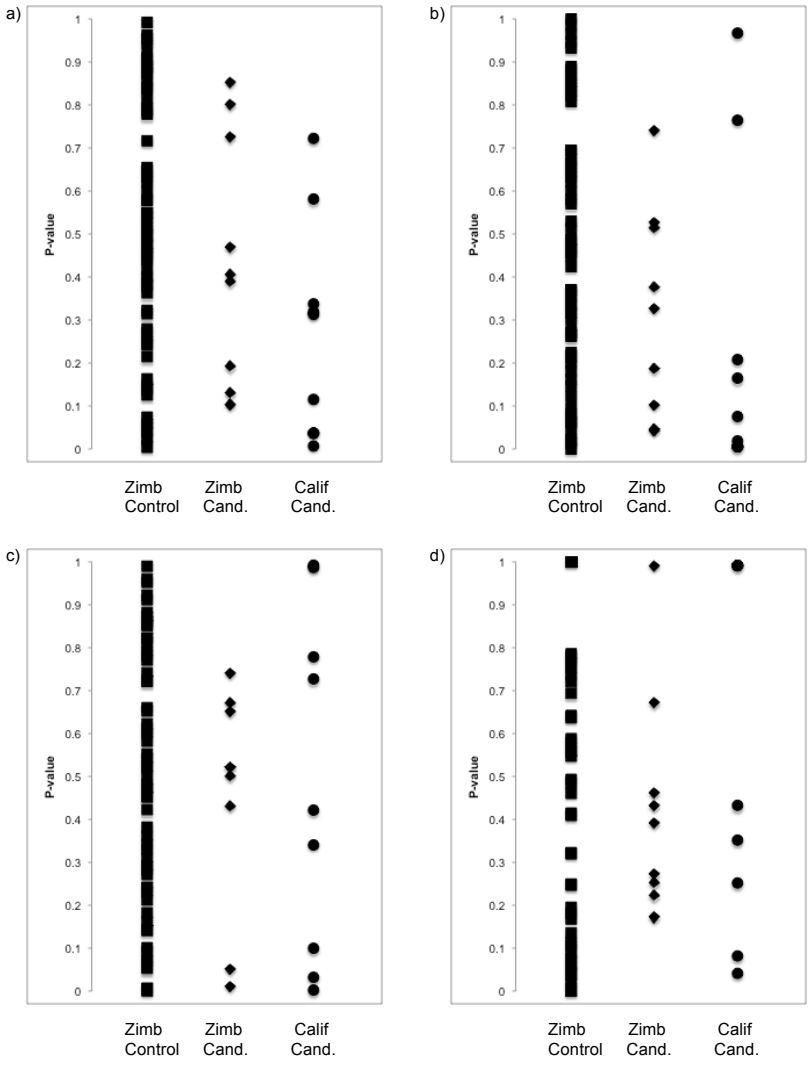


Figure 2a-c: Distribution of polymorphisms. a) synonymous polymorphisms, b) replacement polymorphisms, c) 4-fold synonymous sites. Synonymous and replacement polymorphisms were divided into three classes: low, common, and high frequency. A correction for multiple hits was employed for each dataset.

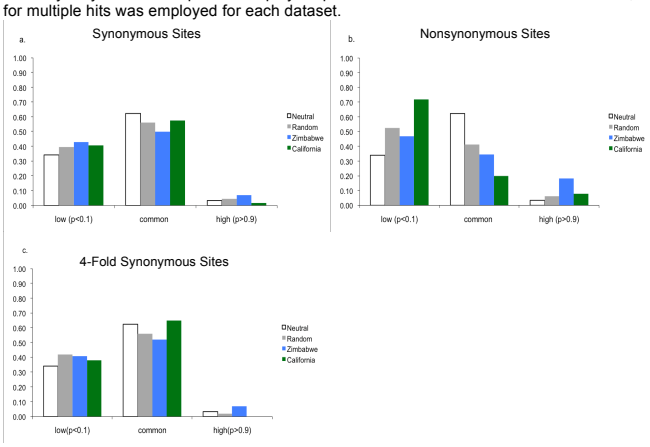


Figure 3: Estimates of the fraction of adaptive amino acid divergence, α, using the method of Bierne and Eyre-Walker (2004). Results given for sex-biased and non-sex biased genes, pooled from Zimbabwe candidate and 137 control genes. Error bars represent 95% confidence intervals.

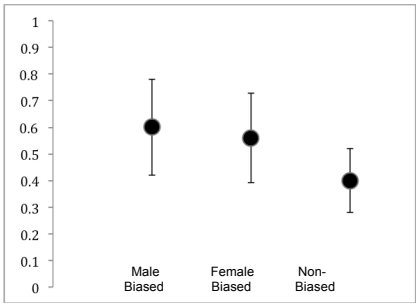
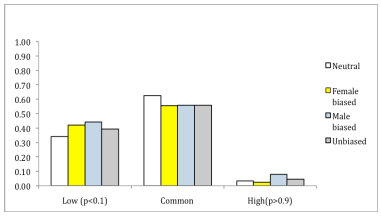
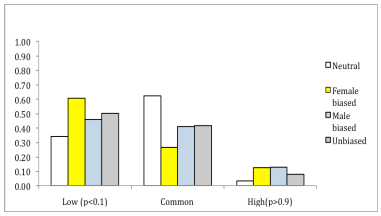


Figure 4: Distribution of polymorphisms. a) synonymous polymorphisms, b) replacement polymorphisms. Synonymous and replacement polymorphisms were divided into three classes: low, common, and high frequency. A correction for multiple hits was employed for each dataset (see Methods).

a)



b)



Supplementary Table 1: Cytological positions, sex-biased expression, and recombination rate of candidate loci.

Candidate	Chromosome	Position	M/F expression[a]	Rec[b]	Rec[c]
CG5273	X	1C4	0.85	0.36	0
CG13004	X	15A11	1.13	2.27	2.10
CG8453	2R	48E7	2.54	1.00	0.99
CG9897	2R	59C1	0.45	1.31	0.62
CG10200	2R	51C3-4	1	1.31	1.14
CG17012	2L	22D2	0.41	2.35	1.70
CG17108	2L	32A4	1.94	1.87	1.89
CG5106	3R	86D8	22.71	0.7	0.56
CG5976	3L	77C4	0.74	0.20	0.08

[a]Average ratio of male to female expression (Gnad and Parsch 2001), [b]Recombination rate (cM/Mb) estimated from Charlesworth (1996), [c]Recombination rate (cM/Mb) estimated from Comeron and Kreitman (2000)

Supplementary Table 2: See Excel file.

Supplementary Table 4: Polymorphism and divergence.

Locus	Polymorphism[a]				Divergence				P-value[b]	
	Synonymous		Nonsynonymous		Synonymous		Nonsynonymous			
Candidate	Zimbabwe	California[c]	Zimbabwe	California	Zimbabwe	California	Zimbabwe	California	Zimbabwe	California
CG5273	1	1	0	0	17	37	6	16	0.99	0.99
CG13004	5	4	3	0	27	22	34	21	0.27	0.08
CG8453	3	2	1	1	12	56	10	19	0.43	0.43
CG9897	13	1	6	0	15	20	15	17	0.17	0.99
CG10200	3	1	1	0	9	18	14	55	0.22	0.25
CG17012	3	4	9	1	23	27	49	60	0.46	0.03*
CG17108	14	1	3	0	27	41	12	15	0.25	0.99
CG5106	10	2	1	1	24	32	2	4	0.67	0.35
CG5976	1	0	1	0	21	19	5	7	0.39	N/A
Candidate (Total)	53	16	25	3	175	272	147	214	0.009*	0.01*
Candidate[d] (Total)	51	15	24	3	137	216	126	191	0.003*	0.009*
Control (Total)	1016		242		2285		1672		0.001*	

[a] Only segregating sites with a frequency >0.1 are reported.
[b] McDonald-Kreitman test (1991). Probabilities are from a two-tailed Fisher's exact test. Significance ($P < 0.05$) is indicated with an asterisk.
[c] A subset of 12 California individuals are reported here to make the number of individuals comparable among datasets.
[d] Total number of polymorphisms when two loci with the lowest recombination rates (CG5976 and CG5273) are removed.

Supplementary Figure 1: Distribution of *P*-values for Fay and Wu's *H* in Zimbabwe female, male, and non-biased expression classes of pooled Zimbabwe candidate and control genes.