

The genomic consequences of adaptive divergence and reproductive isolation between species of manakins

T. L. PARCHMAN,^{*1} Z. GOMPERT,^{*1} M. J. BRAUN,[†] R. T. BRUMFIELD,[‡] D. B. MCDONALD,[§] J. A. C. UY,[¶] E. D. JARVIS,^{**} B. A. SCHLINGER^{††} and C. A. BUERKLE^{*}

^{*}Department of Botany, University of Wyoming, Laramie, WY 82071, USA, [†]National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20560, USA, [‡]Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA, [§]Department of Zoology and Physiology, University of Wyoming, Laramie, WY 82071, USA, [¶]Department of Biology, University of Miami, Miami, FL 33146, USA, ^{**}Department of Neurobiology, Duke University Medical Center, Durham, NC 27710, USA, ^{††}Department of Integrative Biology and Physiology, University of California–Los Angeles, Los Angeles, CA 90095, USA

Abstract

The processes of adaptation and speciation are expected to shape genomic variation within and between diverging species. Here we analyze genomic heterogeneity of genetic differentiation and introgression in a hybrid zone between two bird species (*Manacus candei* and *M. vitellinus*) using 59 100 SNPs, a whole genome assembly, and Bayesian models. Measures of genetic differentiation (F_{ST}) and introgression (genomic cline center [α] and rate [β]) were highly heterogeneous among loci. We identified thousands of loci with elevated parameter estimates, some of which are likely to be associated with variation in fitness in *Manacus* populations. To analyze the genomic organization of differentiation and introgression, we mapped SNPs onto a draft assembly of the *M. vitellinus* genome. Estimates of F_{ST} , α , and β were autocorrelated at very short physical distances (< 100 bp), but much less so beyond this. In addition, average statistical associations (linkage disequilibrium) between SNPs were generally low and were not higher in admixed populations than in populations of the parental species. Although they did not occur with a constant probability across the genome, loci with elevated F_{ST} , α , and β were not strongly co-localized in the genome. Contrary to verbal models that predict clustering of loci involved in adaptation and isolation in discrete genomic regions, these results are consistent with the hypothesis that genetic regions involved in adaptive divergence and reproductive isolation are scattered throughout the genome. We also found that many loci were characterized by both exceptional genetic differentiation and introgression, consistent with the hypothesis that loci involved in isolation are also often characterized by a history of divergent selection. However, the concordance between isolation and differentiation was only partial, indicating a complex architecture and history of loci involved in isolation.

Keywords: genomic cline, hybrid zone, *Manacus*, next-generation sequencing, population genomics, speciation

Received 1 October 2012; revision received 26 November 2012; accepted 29 November 2012

Introduction

Knowledge of the genetic architecture of adaptive divergence and reproductive isolation is fundamental to our understanding of the speciation process. Mounting

empirical data support the genic view of speciation whereby species boundaries can be porous and adaptation and isolation are properties of individual genetic loci that often affect variation in only small genomic regions (Wu 2001; Nosil & Feder 2012). In support of this view, recent studies have repeatedly documented genome-wide heterogeneity in locus-specific estimates of genetic differentiation (Turner *et al.* 2005; Nosil *et al.* 2009; Hohenlohe *et al.* 2010). Moreover, researchers

Correspondence: Thomas Parchman, Fax: (307) 766 2851;

E-mail: tparchma@uwyo.edu

¹These authors contributed equally.

have begun to identify individual variants that contribute directly to adaptation or isolation (e.g., Mihola *et al.* 2009; Tang & Presgraves 2009; Barr & Fishman 2010; Nosil & Schluter 2011). Despite this progress, relatively little is known about the number, effect size and genomic distribution of loci that contribute to adaptation or reproductive isolation. Likewise, despite evidence that divergent natural selection can contribute directly to reproductive isolation and to the origin of incompatibilities (e.g., Dodd 1989; Nosil 2004; Funk *et al.* 2006), we have limited information on the evolutionary history of loci that contribute to reproductive isolation (but see, Tang & Presgraves 2009).

One influential hypothesis regarding divergence during speciation is that genetic differences contributing to reproductive isolation accumulate in extended 'genomic islands' of reduced gene flow and elevated genetic differentiation (Turner *et al.* 2005; Harr 2006; Feder & Nosil 2010; Via 2012). The process of divergence hitchhiking (DH), where divergent selection on a locus reduces effective gene flow in the surrounding genomic region, has been described as a potential cause of extensive genomic islands (Feder & Nosil 2010; Via 2012). Such genomic islands could be important for speciation with gene flow, if they contain coinherited blocks of associated loci that contribute to isolation, and would be more likely to arise in regions of abnormally low recombination (incl. chromosomal inversions or centromeric regions; Feder *et al.* 2003; Navarro & Barton 2003; Hoffmann & Rieseberg 2008; Noor & Bennett 2009). Alternatively, many loci, unlinked and scattered throughout the genome, could underlie reproductive isolation. This hypothesis is supported by mathematical and simulation models suggesting that genomically extensive islands of elevated differentiation between lineages are highly unlikely, except under restrictive circumstances (Barton & De Cara 2009; Feder & Nosil 2010; Feder *et al.* 2012). While each hypothesis has received limited qualitative empirical support (Lawniczak *et al.* 2010; Hohenlohe *et al.* 2010; Nadeau *et al.* 2011; Strasburg *et al.* 2012), the recent accessibility of population genomic data should begin to facilitate more quantitative analyses of the genomic distribution of divergence during speciation (Lawniczak *et al.* 2010; Gompert *et al.* 2012b).

Natural selection can lead to exceptional differentiation in regions of the genome that harbor loci responsible for divergent adaptations (Beaumont & Balding 2004; Pritchard *et al.* 2010). Moreover, reproductive isolation between lineages can be maintained by natural selection if immigrants or recombinants have reduced fitness (Nosil *et al.* 2002; Rundle & Nosil 2005), and can readily arise in diverging lineages through the evolution of incompatibilities (Orr 2001; Turelli *et al.* 2001). Thus, if overlapping sets of loci contributed to adaptation and

reproductive isolation, then the effects of these processes on genetic variation could be localized in the same portions of the genome. Despite the possibility that genomic regions of exceptional differentiation and those underlying reproductive isolation can coincide, there are many reasons they might differ. Genomic regions of pronounced genetic differentiation can evolve by stochastic, neutral processes, particularly if species diverge with little or no gene flow; these include: genome-wide variation in effective population size, mutation rate, frequency of gene conversion, or recombination rate (Lynch 2007; Noor & Bennett 2009). Neutral evolutionary processes can also cause incompatibilities to evolve (Gavrilets *et al.* 1998; Lynch & Force 2000; Fierst & Hansen 2010). Divergence while populations are geographically isolated increases the likelihood of discordance between exceptionally differentiated genetic regions and the loci that confer reproductive isolation. This is because reproductive isolation upon secondary contact could be decoupled from historical divergent selection as a result of spatial or temporal variation in the environment or the genomic composition of populations.

Whether exceptionally differentiated genetic regions harbor loci that contribute to reproductive isolation is an empirical question that can be answered by studying hybrid zones. Specifically, admixture and introgression in hybrid zones can be quantified to identify genetic regions affecting hybrid fitness or assortative mating, which are major components of reproductive isolation (Szymura & Barton 1986; Rieseberg *et al.* 1999; Gompert *et al.* 2012b). Introgression (the movement of alleles from one gene pool into another by admixture) can be quantified by measuring the geographic distribution of alleles across a hybrid zone (geographic introgression; e.g., Barton & Hewitt 1985, 1989; Payseur *et al.* 2004) or the distribution of alleles in different genomic backgrounds (genomic introgression; e.g., Anderson 1949; Rieseberg *et al.* 1999; Gompert & Buerkle 2009). Loci that cause reproductive isolation are expected to reside in genetic regions with specific patterns of introgression. For example, the geographic range of introgression should be reduced for genetic regions containing alleles that decrease hybrid fitness (Barton 1983; Barton & Hewitt 1985; Szymura & Barton 1986). Likewise, within an admixed population genetic regions responsible for reproductive isolation are expected to contain relatively few alleles from one parental species, or alleles from each parental species should be confined to hybrids with different genomic backgrounds (Rieseberg *et al.* 1999; Lexer *et al.* 2007; Gompert *et al.* 2012a). As a result, hybrid fitness and assortative mating can effect the pattern of introgression at individual loci, causing them to introgress differently than neutral loci (Gompert *et al.* 2012a, b).

In this paper we used DNA sequence data to quantify heterogeneity in the genome-wide patterns of genetic differentiation and introgression in manakins (*Manacus*). Manakins are lek-mating birds that occur in forests of South and Central America, known for their spectacular courtship displays and colorful male plumage. A narrow hybrid zone occurs in Panama between the ranges of the white collared manakin (*M. candei*) and the golden collared manakin (*M. vitellinus*; Fig. 1; Parsons *et al.* 1993). These sexually dimorphic species are distinguishable based on morphological and genetic characters, most strikingly in the color of the male collar plumage, a sexually selected trait (Brumfield *et al.* 2001; McDonald *et al.* 2001). Previous studies revealed discordance among geographical clines at molecular markers and plumage traits (Parsons *et al.* 1993), and differential introgression of the *M. vitellinus* yellow collar across the hybrid zone driven by sexual selection (Parsons *et al.* 1993; Brumfield *et al.* 2001; McDonald *et al.* 2001). We generated a population genomic-level data set and used Bayesian models to quantify patterns of genetic differentiation and introgression for tens of thousands of genetic polymorphisms in *M. candei*, *M. vitellinus*, and their hybrids. To quantify the genomic distribution of differentiation and isolation, we mapped

the sequenced polymorphisms onto a draft of the *M. vitellinus* genome. Finally, we tested for a positive correlation between locus-specific genetic differentiation and introgression, to evaluate the hypothesis that exceptionally differentiated genetic regions harbor loci that affect hybrid fitness or contribute to assortative mating between *M. candei* and *M. vitellinus*.

Methods

Sequence data collection and assembly for SNP analyses

We obtained DNA sequence data from 48 *M. candei* (three localities), 52 *M. vitellinus* (three localities), and 104 putatively admixed individuals (five localities; Fig. 1, Table 1). DNA was extracted from muscle or liver tissue from vouchered specimens stored at the National Museum of Natural History of the Smithsonian Institution and the Louisiana State University Museum of Natural Science. We produced highly multiplexed reduced complexity libraries for sequencing on the Illumina platform following the protocol described in Parchman *et al.* (2012), a GBS (genotyping by sequencing) style approach. We digested genomic DNA with two restriction endonucleases (*EcoRI*

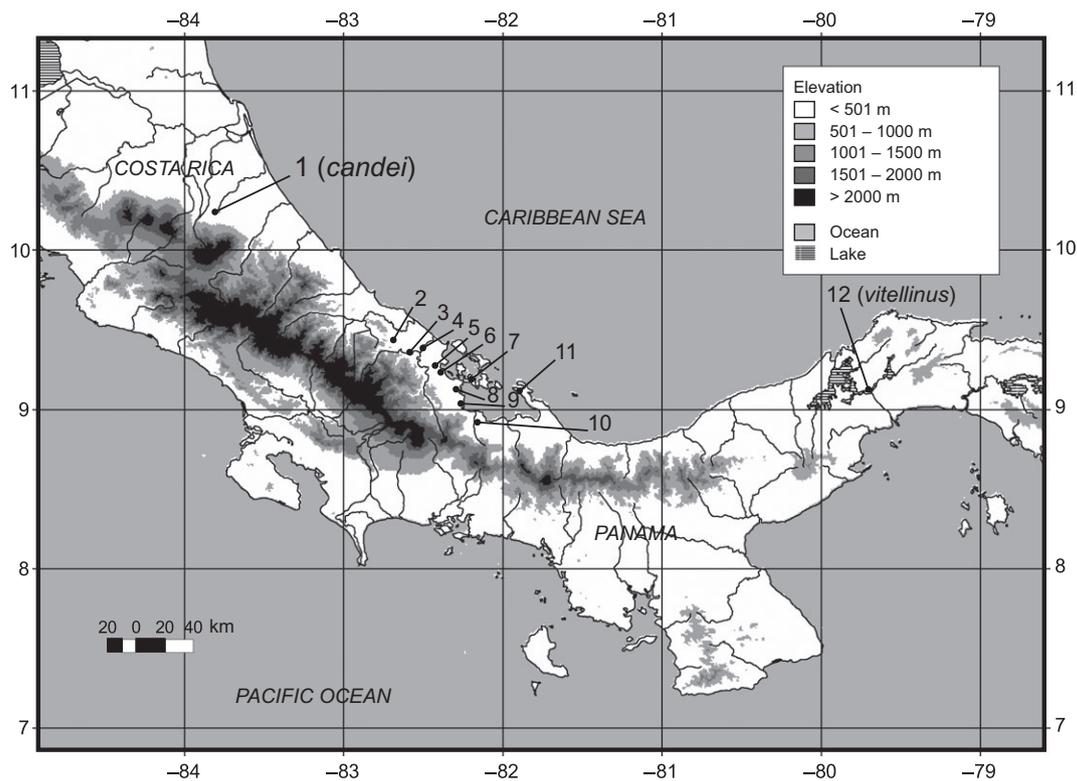


Fig. 1 A map of sampling localities for *M. candei* (sites 1–3), *M. vitellinus* (sites 10–12), and admixed populations (sites 4–9) sampled for this study. The whole genome sequencing project at BGI utilized an *M. Vitellinus* individual from population 12. Population numbers correspond to those listed in Table 1.

Table 1 Population locations and number of individuals of *M. candei*, *M. vitellinus*, and putatively admixed individuals sampled for this study. Numbers corresponding to population designations are as in Brumfield *et al.* (2001). Full details on each vouchered specimen are given in Table S1, Supporting information

Population	Location	Designation	<i>N</i>
1	Limon, Costa Rica (C)	<i>candei</i>	6
2	Rio Sixaola, Panama (C)	<i>candei</i>	22
3	Rio Teribe, Panama (C)	<i>candei</i>	20
4	Rio Changuinola, Panama (VC)	<i>candei/vitellinus</i>	20
5	Rio Oeste, Panama (VC)	<i>candei/vitellinus</i>	21
6	Quebrada Pastores, Panama (VC)	<i>candei/vitellinus</i>	4
7	Tierra Oscura, Panama (VC)	<i>candei/vitellinus</i>	12
8	Rio Uyama, Panama (VC)	<i>candei/vitellinus</i>	21
9	Rio Robalo, Panama (VC)	<i>candei/vitellinus</i>	26
10	Chiriqui Grande, Panama (V)	<i>vitellinus</i>	19
11	Valiente Peninsula, Panama (V)	<i>vitellinus</i>	16
12	Soberania, Panama (V)	<i>vitellinus</i>	17

and *MseI*) and ligated double-stranded, adaptor oligonucleotides to the digested fragments. These oligonucleotides contained the priming sites for Illumina sequencing and included 10 bp barcodes for the unique identification of the individual associated with each sequencing read. We used PCR to amplify barcode-adapted fragments using two replicate reactions for each individual, which were then pooled. Amplicons for sets of up to 96 individuals (204 individuals were spread across 2.2 sets or sequencing libraries) with unique barcodes were pooled. We separated pooled amplified libraries on a 2% agarose gel and excised fragments between approximately 350 and 500 bp in length. We purified these fragments using Qiagen's Qiaquick Gel Extraction kit (Qiagen Inc.). Sequencing of the libraries was performed by the National Center for Genome Research (Santa Fe, NM, USA) using the Illumina GAIIx platform. Each of the pooled libraries was sequenced in a separate lane, and 108 base reads were produced.

We conducted a two step assembly procedure that first used a *de novo* assembly on a subset of GBS reads to produce an artificial reference. We then assembled the full set of reads onto the reference using SeqMan

xng 1.0.3.3 (DNASTAR). We used custom Perl scripts together with samtools and bcftools (Li *et al.* 2009) to identify variant sites in the assembled sequence data and determine the number of reads supporting each alternative nucleotide state for each individual and locus. The data from the called variant sites were parsed and placed in files containing information on the haplotype counts for each individual at each genetic region, and a file containing counts of the number of reads for each SNP in each individual. We then removed all genetic regions where individuals appeared to have more than two haplotypes, and discarded any variable site where the observed allele counts from apparently heterozygous individuals were very unlikely given a binomial distribution with $P = 0.5$.

Quantifying genetic differentiation and introgression

We used a Bayesian model to estimate population allele frequencies for each of the SNPs identified above based on the observed sequence data (the model is described in Gompert *et al.* 2012a). We treated the genotype at a locus and the population allele frequency as unknown model parameters, and estimated genotype probabilities and allele frequencies separately for each of the 12 sampled localities (populations). We obtained parameter posterior probabilities using Markov chain Monte Carlo (MCMC). Each analysis consisted of a single chain iterated for 20 000 steps where samples were recorded every fourth step. We used principal component analysis (PCA) to summarize population genetic structure for the samples of the pure species and putative hybrids. We used the estimated genotype probabilities for two of three genotypes (the heterozygous genotype and one homozygous genotype) at each locus as variables for PCA. We performed the PCA in R (R Development Core Team 2012) using prcomp.

We used a Bayesian implementation of the F-model to quantify genome-wide genetic differentiation between *M. candei* (sites 1–2) and *M. vitellinus* (sites 10–12) (Gompert *et al.* 2012a). This provides a metric of genetic differentiation that is analogous to F_{ST} under several evolutionary models (Balding & Nichols 1995; Nicholson *et al.* 2002; Falush *et al.* 2003), and in this paper we equate this metric with F_{ST} . This approach models uncertainty in F_{ST} arising from evolutionary and statistical sampling, allows information sharing among loci, and treats F_{ST} as an evolutionary parameter rather than a summary statistic of allele frequencies (Gompert *et al.* 2012b). In particular, the model accounts for the genotypic uncertainty arising from variable and limited sequencing coverage across loci. We estimated posterior probabilities of F_{ST} for each locus and a metric of F_{ST} for the genome-wide average F_{ST} using 25 000 MCMC steps where every 10th step was saved and the first 1,000 steps were discarded as

burn-in. We designated outlier loci with higher than expected levels of genetic differentiation between *M. candei* and *M. vitellinus* as loci with $F_{ST} >$ the 0.95 quantile for the genome F_{ST} (Gompert & Buerkle 2011b; Gompert *et al.* 2012b).

We used the Bayesian genomic cline model (Gompert & Buerkle 2011a; Gompert *et al.* 2012b; Gompert & Buerkle 2012) to quantify genome-wide variation in introgression among putatively admixed *Manacus*. The model includes two cline parameters that determine the probability that an individual with hybrid index $H = h$ inherited a gene copy at locus $I = i$ from *M. candei* (denoted ϕ ; the probability of *M. vitellinus* ancestry is $1 - \phi$), where the average probability of *M. candei* ancestry across all loci is equal to an individual's hybrid index. The cline parameter α specifies an increase (positive α) or decrease (negative α) in the probability of *M. candei* ancestry relative to the base expectation of $\phi = h$, whereas the genomic cline parameter β specifies an increase (positive β) or decrease (negative β) in the rate of transition from low to high probability of *M. candei* ancestry as a function of hybrid index (Gompert & Buerkle 2011a). The cline parameter β is a measure of the average, ancestry-based, pairwise linkage disequilibrium between a locus and all other loci. Simulations indicate that selection against specific hybrid genotypes affects both cline parameters (Gompert *et al.* 2012b). For example, extreme values of α can reflect single-locus underdominance or Dobzhansky-Muller incompatibilities (DMIs; Gompert *et al.* 2012a, b). Because underdominance leads to a deficit of heterozygotes, the frequency of one of the homozygotes will rise necessarily, leading to an excess of one species' ancestry at a locus relative to the base expectation based on hybrid index (Gompert *et al.* 2012b). Extreme values of β are likely when there is population structure in the hybrid zone, or with strong selection against hybrids and high levels of gene flow from the parental species (Gompert *et al.* 2012a, b). We estimated marginal posterior probability distributions for hybrid indices and cline parameters (α and β) using MCMC. We ran five independent chains for 50 000 steps each and recorded samples from the posterior distribution every 20th step following a 30 000 step burn-in. We combined the output of the two chains after inspecting the MCMC output to assess convergence to the stationary distribution. We tested for an association between F_{ST} and genomic cline parameters α and β using Pearson's product moment correlation coefficient.

Quantifying genomic heterogeneity

A draft assembly of the *M. vitellinus* genome was produced at the Beijing Genome Institute (BGI) as part of a larger avian phylogenomics project (further details of

genome sequencing will be in a future publication). The genome was sequenced with a whole-genome shot gun strategy on an Illumina HiSeq 2000 instrument. DNA extracted from an unvouchered male individual was used to construct eleven paired-end libraries with various insert sizes ranging from $0.2 - 4 \times 10^5$ bp. We aligned the consensus sequences that formed the artificial reference for our GBS assembly onto the largest 659 scaffolds (all scaffolds longer than 50 000 bases; covering 959×10^6 bp) from the draft *M. vitellinus* genome using SeqMan xng. We calculated Moran's I (Moran 1950; Epperson 2003) at various physical distances (from 10^2 to 10^6 bases) to measure genomic autocorrelation for F_{ST} and genomic cline parameters for the mapped genetic regions. Moran's I is given by:

$$I_{(k1,k2)} = \frac{n_k \sum_i \sum_{i'} r_{ii'} \{(k1, k2)\} \theta_i \theta_{i'}}{R_{(k1,k2)} \sum_i \theta_i^2}. \quad (1)$$

In Eqn. 1, the $r_{ii'}$ are binary variables that take on a value of one if the physical distance between locus i and i' is $k1 < r_{ii'} \leq k2$, and otherwise take on a value of 0. $R_{(k1,k2)}$ is the sum of all $r_{ii'}$ and n_k is the total number of loci. θ is the parameter of interest (i.e., F_{ST} , α , or β). We calculated $I_{(k1,k2)}$.

We further analysed the genomic distribution of parameter estimates by examining the distribution of loci that had outlier estimates for F_{ST} , α , and β across the scaffolds they mapped to. Specifically, we tested whether the distribution of loci per scaffold better fit a binomial distribution with a single probability of occurrence for the whole genome (the number of outliers per total bases) or a distribution where occurrence of outlier loci was scaffold-specific. Our goal was to determine whether there was statistical evidence that loci with exceptional parameter estimates were clustered or distributed homogeneously across the genome. We generated likelihood estimates for both constant probability and scaffold-specific models for loci with exceptional parameter estimates (F_{ST} outliers, α 95% CI not including zero, and β 95% CI not including zero). We used a likelihood ratio test to compare the log-likelihoods of these models, and determined significance with a one-tailed χ^2 test statistic.

We estimated Burrow's composite measure (Δ) of Hardy-Weinberg and linkage disequilibrium (LD) between each of the 59 100 SNP loci (Weir 1979) in order to examine the independence of markers and the strength of overall LD among them. We also estimated Δ separately in each population to compare average LD across these markers in parental populations of *M. candei*, *M. vitellinus*, and in admixed populations. The Δ metric does not require phased data, does not assume Hardy-Weinberg equilibrium at either locus, and provides a joint measure of intralocus and interlocus

disequilibria (Weir 1979). Because of uncertainty in genotypic state for each individual, we used a Monte Carlo procedure to sample genotypes for individuals and calculated the composite linkage disequilibrium measure (Δ_{ij}) for each locus pair. We sampled genotypes for each individual and locus according to the estimated genotype probabilities obtained previously. We then obtained our estimate for Δ_{ij} by taking the mean of the calculated values over 1000 repeated samples of individual genotypes generated with the Monte Carlo procedure. Using estimates of LD between all pairs of loci that mapped onto the *M. vitellinus* genome, we examined the relationship between LD and physical distance on the largest scaffolds.

Results

Genetic differentiation and introgression

Illumina sequencing of our reduced complexity GBS libraries produced 49.5 million, 108 base pair (bp) DNA sequences. An initial *de novo* assembly of a subset of 16 million sequences placed 9 941 228 reads into 315 533 contigs, which were used to build an artificial reference genome. The reference based assembly utilizing all reads placed 39 383 248 reads onto the artificial reference, for an average coverage depth of $130 \times$ per sequenced region ($0.63 \times$ per bird per sequenced region). After assembly and variant parsing, these data contained 88,560 single nucleotide polymorphisms (SNPs) distributed among 74 283 short (~ 92 bp) contigs. We discarded loci with minor allele frequencies < 0.1 as they contain little information about ancestry and divergence, and because low F_{ST} estimates for these loci could confound analyses of the correlation between genetic divergence and introgression (Gompert *et al.* 2012b). We retained 59 100 SNPs that we used in most analyses. Mean estimates of Burrow's composite measure of intra-locus and inter-locus disequilibria (Δ) were less than 0.01 in each population (Table 2), indicating that most sampled SNPs are largely independent.

Principal component analysis of estimated posterior probabilities of multilocus genotypes and estimates of genome-level F_{ST} demonstrated substantial genetic differentiation between parental *M. candei* and *M. vitellinus* populations ($F_{ST} = 0.259$, 95% credible interval [CI] 0.256–0.261), but also a wide range of weaker differentiation among populations that had genetic compositions intermediate between the parental species (Fig. 2). In agreement with previous studies (Parsons *et al.* 1993; Brumfield *et al.* 2001; Yuri *et al.* 2009), the principal genomic transition from *candeii* to *vitellinus* occurred in the geographic region of populations 8–10, whereas the greatest change in male plumage traits occurred from

Table 2 Quantiles of pairwise linkage disequilibrium (Δ) estimates for 59 100 loci for individual populations of parental *M. candei*, *M. vitellinus* and admixed populations. Δ was not estimated for populations 1 and 6 because of small sample sizes

Species	Population	5%	25%	50%	75%	95%
<i>M. candei</i>	2C	0.000	0.001	0.001	0.002	0.011
	3C	0.000	0.001	0.001	0.003	0.013
<i>M. candei</i> × <i>M. vitellinus</i>	4VC	0.000	0.001	0.002	0.003	0.011
	5VC	0.000	0.001	0.002	0.003	0.010
	7VC	0.000	0.002	0.004	0.007	0.013
	8VC	0.000	0.001	0.002	0.005	0.019
<i>M. vitellinus</i>	9VC	0.000	0.001	0.004	0.011	0.024
	10V	0.000	0.001	0.003	0.005	0.013
	11V	0.000	0.001	0.003	0.006	0.017
	12V	0.000	0.001	0.003	0.009	0.031

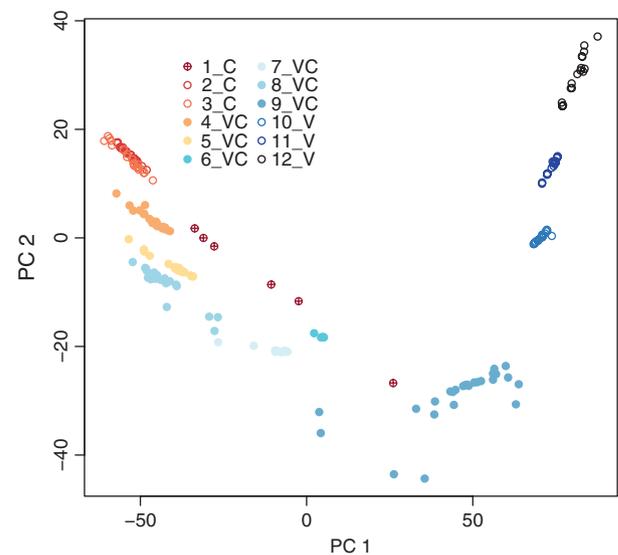


Fig. 2 Statistical summary of population genetic structure based on principal component axes one (PC1), and two (PC2). These axes explain 38.4% (PC1), and 3.9% (PC2) of the genetic variation among individuals, based on Bayesian estimates of genotypes. Population numbers are as in Figure 1 and Table 1. Individual members of the populations of parental species were used to estimate parental allele frequencies and are represented by open circles. Individuals from the admixed populations for which genomic cline analysis was conducted are represented by filled symbols. Six individuals from a sparsely sampled population (1_C) are indicated by crossed open symbols; there was considerable uncertainty about allele frequencies in this population and it was not included in other analyses.

population 3–4, which contain predominantly *candeii* ancestry. These analyses indicated an abundance of hybrid individuals with admixed genomes relative to *M. candei* and *M. vitellinus* (Fig. 2). Bayesian estimates of hybrid index, a measure of the proportion of an indi-

vidual's genome inherited from *M. candei*, for the admixed individuals ranged from 0.05 to 0.95 but were skewed toward *M. candei* ancestry (Fig. 3b), in agreement with Brumfield *et al.* (2001) and Yuri *et al.* (2009). Nonetheless, the wide range of hybrid indices provided an informative basis for the analysis of variable introgression across the hybrid zone (Fig. 3b).

Locus-specific estimates of genetic differentiation (F_{ST}) between *M. candei* and *M. vitellinus* were highly heterogeneous (Fig. 3a) and ranged from 0.09 to 0.61 (mean 0.26). Outlier loci are expected to reside in regions of the genome affected by divergent selection more often than non-outlier loci (Beaumont & Nichols 1996; Gompert & Buerkle 2011b). Estimates of F_{ST} for 1007 loci exceeded the 95th quantile of the estimated genome-wide distribution of F_{ST} (a Normal $[\mu, \tau]$ prior distribution for $\log[F_{ST_i}/(1-F_{ST_i})]$) and were classified as statistical outlier loci (Fig. 3a).

Estimates of genomic cline parameters α (min = -2.17, max = 1.25) and β (min = -2.13, max = 1.09) were highly variable across the genome (Fig. 3c,d), and introgression at many loci differed significantly from the genome-wide average. Specifically, we detected excess *M. candei* ancestry for 623 loci (lower bound of 95% CI for $\alpha > 0$) and excess *M. vitellinus* ancestry for 514 loci (upper bound of 95% CI for $\alpha < 0$). We also detected loci with significantly elevated estimates of genomic cline rate (203 loci with lower bound of 95% CI for $\beta > 0$, 220 loci with upper bound of 95% CI for $\beta < 0$). These loci exhibited exceptional rates of transition between high and low probabilities of *M. candei* ancestry given hybrid index. The estimated variation in locus-specific introgression is biologically significant. For individuals with intermediate

hybrid indices, the probability of locus-specific *M. candei* ancestry was nearly one for some loci (e.g., loci with high positive α) and nearly zero for other loci (e.g., loci with high negative α and β ; Fig. 3c,d).

Genetic differentiation between *M. candei* and *M. vitellinus* was correlated with introgression in admixed individuals (Fig. 3e). There was a positive correlation between F_{ST} and the absolute values of both α ($r = 0.17$, $P < 2.2 \times 10^{-16}$) and β ($r = 0.36$, $P < 2.2 \times 10^{-16}$). 43 loci were F_{ST} outliers with extreme α estimates, and 117 loci were F_{ST} outliers and had extreme β estimates. These loci thus exhibit patterns consistent with being affected by divergent selection between the parental species and also being associated with hybrid fitness or assortative mating in the hybrid zone. The level of genetic divergence between the parental species at a locus dictates how informative the locus can be about ancestry, which could lead to an overall spurious correlation between exceptional introgression and F_{ST} . However, this is unlikely to apply to this study, because analyses of simulations of neutral admixture indicate that such a correlation is most likely when $F_{ST} < 0.1$ Gompert *et al.* (2012a), whereas $F_{ST} > 0.2$ between the *Manacus* species.

Genomic heterogeneity

Sequencing of the male *M. vitellinus* multiple size insert libraries on an Illumina HiSeq instrument produced 171 Gb of raw data. After filtering, 132 Gb of sequence data were retained for assembly, which was accomplished *de novo* with the software SOAPdenovo (Li *et al.* 2010). A final draft assembly of 1.16 Gb was generated after

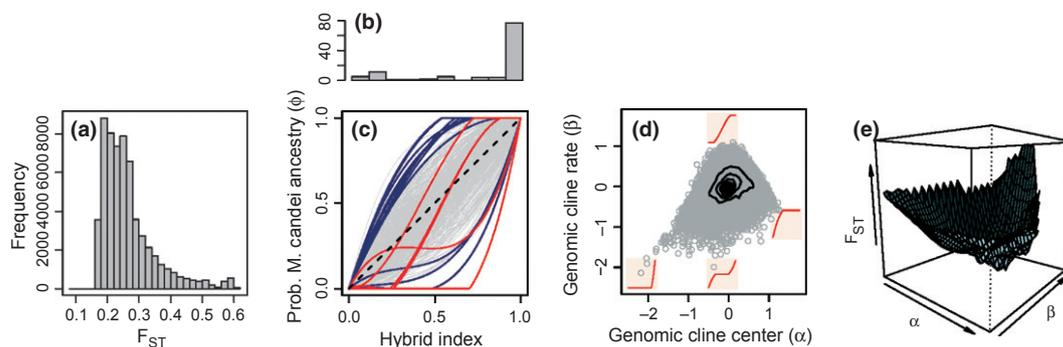


Fig. 3 (a) Histogram of locus-specific estimates of F_{ST} between parental *M. candei* and *M. vitellinus* populations. (b) Histogram of hybrid indices for the 110 putatively admixed individuals (hybrid index of pure *M. candei* = 1.0, and pure *M. vitellinus* = 0.0). (c) Estimated genomic clines for 1000 representative loci. Each blue (95% CI for α does not include 0), red (95% CI for β does not include 0) or gray (95% CI for α and β includes 0) line represents the genomic cline for a single locus. The dashed black line gives $\phi = h$. (d) Variable introgression in hybrid populations based on genomic cline center (α) and genomic cline rate (β). Each point represents one of the 59 100 loci. Contour lines depict the joint density of estimated cline parameters. In red are illustrations of example genomic clines that are representative of loci in the corresponding region of the α - β parameter space. (e) Three dimensional surface of the relationship between α , β , and F_{ST} , as estimated with a loess interpolation algorithm. The highest estimates of F_{ST} occur at the most extreme values for genomic cline rate and center parameters.

gaps were filled. The total contig size was 1.13 Gb and the total scaffold size was 1.16 Gb, while the contig N50 and scaffold N50 were 34.26 Kb and 2.54 Mb. We aligned all GBS sequences containing the SNPs analyzed above to the largest 659 scaffolds (all scaffolds longer than 50 000 bases; covering 959×10^6 bp) from the draft *M. vitellinus* genome. Given the size of other bird genomes Warren *et al.* (2010), this set of scaffolds could represent as much as 80% of the *M. vitellinus* genome. 27 095 of the 59 100 GBS variable sites aligned to scaffolds, a good percentage given the stringent criteria used. These mapped at a density of 138 GBS fragments per 10^6 bp (or 191 SNPs per 10^6 bp). There was a tight linear relationship between scaffold length and the number of SNPs mapped onto scaffolds ($r^2 = 0.98$, $P < 2.2 \times 10^{-16}$; intercept = 0.058, slope = 0.0001), indicating that the GBS library contained fragments that were uniformly distributed across the draft genome (Fig. 4). In addition, GC content in all scaffolds in the draft *M. vitellinus* genome (0.4104) was essentially identical to the GC content of the GBS fragments (0.4098), suggesting that GBS fragments were similar in composition to sequences in the whole genome assembly. Less than 10% of the GBS sequences aligned to protein coding sequences from the *M. vitellinus* genome (based on draft annotation), indicating, as expected, that the majority of the GBS reads reside in non-coding regions. Locus-specific estimates of the population genetic parameters were moderately autocorrelated at very short distances (< 100bp) along the scaffolds (F_{ST} : Moran's $I = 0.36$, α : $I = 0.32$, and β : $I = 0.35$; Fig. 5). Autocorrelations for F_{ST} and genomic cline parameters decayed at a similar rapid rate with physical distance ($I \approx 0.05$ or less at 5000 bp), but remained above zero for loci that mapped to the same scaffold. As expected, we did not detect autocorrelation of parameter estimates among loci on different scaffolds ($I \approx 0$).

Loci with extreme parameters were widely distributed across the draft genome assembly: 166 (25.2%) of the scaffolds had F_{ST} outlier loci, 240 (36.4%) had loci with exceptional α estimates, 115 (17.5%) had loci with exceptional β estimates, and overall 45.4% of the scaffolds contained exceptional loci. F_{ST} outlier loci better fit a model with different probabilities of occurrence on different scaffolds, rather than a model where outlier occurrence was characterized by a common probability on all scaffolds ($P < 0.0001$, $df = 658$), suggesting some aggregation of loci with exceptional F_{ST} estimates. We did not detect a statistically significant difference between constant probability and scaffold-specific probability models for loci with exceptional estimates of α (upper bound of 95% CI $\alpha < 0$, lower bound of 95% CI $\alpha > 0$) or β (upper bound of 95% CI $\beta < 0$, lower bound of 95% CI $\beta > 0$; P ranged from 0.989 to 1). Thus,

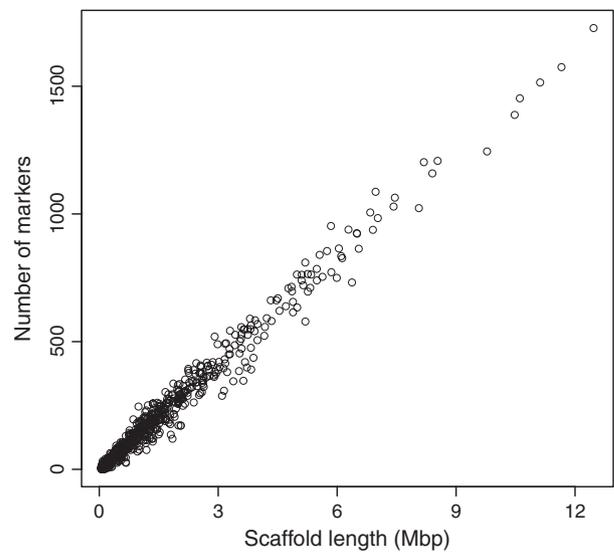


Fig. 4 The number of sequencing reads that mapped to scaffolds in the draft genome was a tight linear function of scaffold size (number of markers = $0.058 + 0.0001 \times$ scaffold length; $r^2 = 0.981$, $P < 2.2 \times 10^{-16}$). This relationship suggests uniform coverage, rather than a clustering of reads from the GBS library into a particular fraction of the *M. vitellinus* genome.

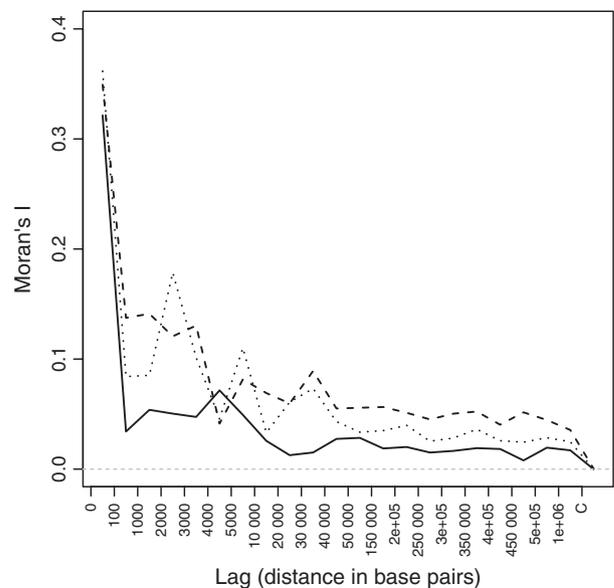


Fig. 5 Correlogram of genomic autocorrelation of parameter estimates. Solid lines show Moran's I for α (solid line), β (dashed line), and F_{ST} (dotted line) at different genetic distances. 'C' is the maximum distance between two variable sites on the same scaffold; loci on different scaffolds are assigned a distance greater than 'C'. Genomic autocorrelation approaches zero (dashed grey line) for loci on different scaffolds.

our analyses provide evidence that loci with extreme parameter estimates are widespread across the genome, although highly differentiated loci (i.e., F_{ST} outliers) do

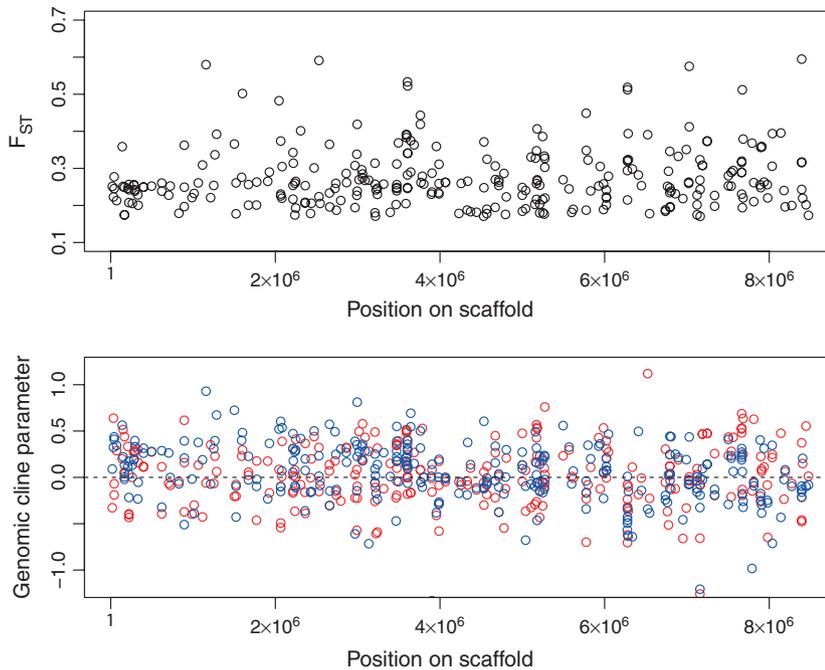


Fig. 6 Parameter estimates for F_{ST} (black, top panel), α (blue, bottom panel), β (red, bottom panel) by their physical position across one of the largest scaffolds in the *M. vitellinus* draft genome assembly.

not occur with equal frequency across all regions of the genome (Fig. 6).

Discussion

We detected considerable variation in genetic differentiation and introgression across the genome, with estimates of each differing from the genome-average or genome-wide expectations at hundreds of loci. By mapping the sequenced regions onto a draft of the *M. vitellinus* genome we detected pervasive, fine-scale genomic heterogeneity in levels of differentiation and introgression. The frequency of highly differentiated loci was heterogeneous across the genome, but genetic regions with extreme introgression or high levels of genetic differentiation were widespread rather than being colocalized in discrete genomic regions. Although multiple evolutionary processes affect patterns of differentiation and introgression, theory suggests that functional variants affecting local adaptation or hybrid fitness should be more prevalent in regions of the genome characterized by elevated genetic differentiation and non-neutral patterns of introgression. Thus, the observed pattern of genomically widespread variation in differentiation and introgression is consistent with the hypothesis that many functional variants contribute to isolation and with selection being variable in space and time. We detected a positive correlation between locus-specific estimates of F_{ST} and genomic cline parameters. This observation further supports the hypothesis that the

genetic basis of isolation is complex, and the hypothesis that the genetics of local adaptation and reproductive isolation are entwined. Nonetheless, differentiation and introgression were only partially concordant. We discuss multiple, complementary explanations for this partial concordance below.

Genomic heterogeneity of genetic differentiation and isolation

We found considerable heterogeneity across the genome in genetic differentiation and introgression (Figs 5 and 6), including hundreds of loci with exceptional, non-neutral patterns of differentiation or introgression. Rather than strong co-localization of loci with exceptional variation in a few, discrete genomic regions, we found these loci to be scattered across hundreds of genomic regions and surrounded by rapidly decaying LD. Our results build on comparable recent studies that have identified narrow regions of elevated differentiation that are widely dispersed across the genome (Burke *et al.* 2010; Lawniczak *et al.* 2010; Strasburg *et al.* 2012; Ellegren *et al.* 2012). Such empirical findings are consistent with the view of a porous genome, where the genomic consequences of adaptation and isolation are narrow and restricted. In the literature there is considerable interest in the consequences of *in situ* divergent selection for linked (divergence hitchhiking; Feder & Nosil 2010; Feder *et al.* 2012; Via 2012). Empirical and theoretical studies in this area are concerned with

divergence in the presence of on-going gene flow, and indicate that the efficacy of divergence hitchhiking is heavily conditional on levels of migration, effective population size, and the strength of selection (Feder & Nosil 2010; Yeaman & Whitlock 2011; Feder *et al.* 2012). Although it is unclear how divergence hitchhiking expectations might apply to taxa that have differentiated during periods of allopatry (as in our study), our results agree with theory suggesting that strong co-localization of genetic regions involved in speciation are highly unlikely, except under restrictive circumstances (Barton & De Cara 2009; Feder & Nosil 2010; Yeaman & Whitlock 2011; Feder *et al.* 2012). When studies have provided strong evidence for elevated divergence across extended chromosomal regions, this commonly occurs in areas of suppressed recombination (centromeric regions or inversions; Feder *et al.* 2003; Hoffmann & Rieseberg 2008; Noor & Bennett 2009; White *et al.* 2010; Michel *et al.* 2010). Part of the difficulty with evaluating the applicability of the “genomic islands” metaphor is that, as noted by Nosil & Feder (2012), it has generally been used for qualitative discussion of heterogeneous genomic divergence, rather than with quantitative models and statistics to test covariation among linked loci and to tie patterns of variation to processes. Our results highlight the need for more quantitative analysis of how genetic differentiation and introgression vary across the genome (with high density, mapped markers), and for refined predictions for the genomic consequences of divergent selection and isolation during speciation.

The genomic patterns of heterogeneous differentiation and introgression we detected would be expected if adaptive phenotypic differences or barriers to gene flow are caused by many widely dispersed alleles with small individual effects. Theory suggests that this type of genetic architecture is expected when populations adapt to gradually changing environments (Kopp & Hermisson 2009) or when selection acts on standing genetic variation (Hermisson & Pennings 2005; Pritchard *et al.* 2010). Conversely, intermediate and large effect alleles should contribute more to reproductive isolation when populations adapt to a stable optima by fixing novel mutations (Gillespie 1984; Orr 1998), particularly if populations diverge with gene flow (Yeaman & Whitlock 2011). Mounting empirical evidence also indicates that large effect alleles can contribute to reproductive isolation (Bradshaw & Schemske 2003; Colosimo *et al.* 2005; Linnen *et al.* 2009; Joron *et al.* 2011). But most of these studies investigate discrete traits, and the genetic architecture of quantitative traits likely differs (Rockman 2012). Thus, whereas it is possible that large effect alleles contribute to adaptive differences in *Manacus* and that we simply failed to detect these alleles, the genomic data could easily be

explained if fitness differences are caused by a suite of quantitative traits and arise from polygenic adaptation involving substantial standing variation.

Estimates of LD (Δ) in populations of the parental species were roughly equal to those in hybrid populations (Table 2). The absence of admixture LD and the rapid decay of autocorrelation among genomic cline parameter estimates is perhaps surprising given expectations for hybrid zones (Buerkle & Rieseberg 2008). This pattern would be predicted after extensive recombination in advanced generation hybrids and a lack of recurrent migration from the parental species into the zone. A diminished amount of parental gene flow into the hybrid zone is plausible because of small rivers in the region that limit dispersal (Brumfield *et al.* 2001). Admixed populations have the potential to be useful for the mapping of traits in natural populations, due to admixture LD in earlier generation hybrids (Buerkle & Rieseberg 2008; Lexer *et al.* 2007). However, the rapid decay of LD across the physical map and low levels of genomic autocorrelation in the *Manacus* hybrid zone mean that a high density of markers would be required for comprehensive mapping of causal variants associated with phenotypes and fitness.

Divergent selection and reproductive isolation

Heterogeneous genetic differentiation and introgression can be caused by selection or stochastic evolutionary processes, particularly if population sizes are small. Divergent selection is most likely to generate regions of elevated genetic differentiation when strong selection favors a new derived allele or occurs in a genetic region with reduced recombination (Maynard-Smith & Haigh 1974). The effect of selection on introgression depends on the rate of gene flow from parental species to the hybrid zone and the form of selection (Endler 1977; Gompert *et al.* 2012b). Elevated estimates of α can arise from multiple forms of selection. For example, simulations indicate that selection against interspecific heterozygotes (i.e., underdominance) or reduced hybrid fitness caused by DMIs can purge alleles inherited from one species from the admixed population (i.e., cause elevated positive or negative α). This pattern can also arise from introgression of a universally favored allele, but would not contribute to a positive correlation between α and F_{ST} . These same forms of selection can also confine alleles inherited from each species to hybrids with different genomic compositions (i.e., cause elevated positive β), but this effect is minimal unless there is considerable gene flow from parental species, selection in the hybrid zone is very strong, or there is population structure within the hybrid zone (Gompert & Buerkle 2011a; Gompert *et al.* 2012b). The high

positive estimates of β observed in the *Manacus* hybrid zone were likely caused by population structure in the hybrid zone, perhaps coupled with variation in selection among admixed populations, as pairwise linkage disequilibria were quite low within each admixed population (Table 2). Thus, the large number of loci with exceptional differentiation or introgression is consistent with the hypothesis that the genetic basis of divergent adaptation and reproductive isolation is complex, and influenced by many, widely dispersed genomic regions, rather than a few, co-localized loci of major effect. But, at this stage we cannot reject the alternative hypothesis that selection is generally weak in the hybrid zone and that the observed genomic heterogeneity is the result of genetic drift.

We found that genomic regions exhibiting substantial genetic differentiation and exceptional introgression were partially concordant. There was a strong correlation between F_{ST} and the absolute values of both α and β , and F_{ST} outlier loci also commonly had extreme genomic cline rate and center parameters (Fig. 3e). These results are consistent with the hypothesis that historical divergent selection is associated with reproductive isolation and predicts hybrid fitness, which could be caused by DMIs or underdominance. These results are in agreement with a large body of literature linking reproductive isolation with adaptive divergence (Funk *et al.* 2006; Schluter 2009; Nosil & Feder 2012), and with a recent population genomic analysis of a *Lycaeides* hybrid zone (Gompert *et al.* 2012a). In *Lycaeides* butterflies, locus-specific genetic differentiation between parental species was correlated with locus-specific introgression (genomic cline parameter α), and patterns of differentiation and introgression were consistent with the hypothesis that reduced hybrid fitness is caused by differential adaptation to habitat or host plant, or loci involved in DMIs.

Nonetheless, introgression for the majority of highly differentiated loci conformed with average, genome-wide introgression. Similarly, we observed extreme introgression in the hybrid zone for many weakly differentiated loci. These results suggest that reproductive isolation is not simply a consequence of adaptive divergence between the parental species. Such a complex relationship between the genomic consequences of reproductive isolation and adaptation might stem from allopatric divergence between the parental species, when a large portion of the genome diverged through diverse selection pressures and other evolutionary processes before secondary contact (Gompert *et al.* 2012a). There are many ways in which reproductive isolation can evolve by processes other than divergent selection. For example, DMIs can arise from genetic drift (Gavrilets *et al.* 1998), biased gene conversion (Fierst & Hansen 2010), or the stochastic deactivation of gene duplicates in allopatry (Lynch & Force 2000). Alternatively, variation in genetic differentiation and introgression could be

caused primarily by drift, which could operate semi-independently in allopatric and admixed populations. This discordance and concordance in patterns of differentiation and introgression highlights complexity in the processes giving rise to reproductive isolation.

Previous studies of this hybrid zone based on geographical clines at molecular markers and plumage traits indicated differential introgression of sexually selected traits (Parsons *et al.* 1993; Brumfield *et al.* 2001; McDonald *et al.*, 2001). In particular, the yellow collar of *M. vitellinus* has introgressed into the *M. candei* genomic background. The yellow plumage of *M. vitellinus* males has been associated with increased fitness due to sexual selection (Stein & Uy 2006) and also shown to be involved in dominance interactions in mating leks in the hybrid zone (McDonald *et al.*, 2001). While hybrid indices indicate a predominance of *M. candei* ancestry in the hybrid zone, these populations are fixed for the yellow collar of *M. vitellinus*. Loci with extremely high, negative values for genomic cline center (α) estimates have introgression of *M. vitellinus* ancestry into a *M. candei* background (Fig. 3d), and could be candidates tagging genomic regions associated with the yellow collar trait. Although a subset of these high, negative α loci assemble to the draft *Manacus* genome and are near or inside annotated genes, gene annotations involve unrelated function, or are too crude to reliably identify candidate plumage loci. Future genome-wide association mapping of plumage variation in large samples of individuals, together with continued refinement of *M. vitellinus* genome annotations have the potential to advance our understanding of the phenotypic and genetic basis of adaptation and reproductive isolation in these taxa.

Acknowledgements

This research was funded through the Small Grants Program of the National Museum of Natural History to M. Braun and NSF grants IOS-1137624 to A. Uy and DEB-1146265 to R. T. Brumfield. C. W. Benkman, P. Nosil, and R. Safran provided helpful comments on previous drafts of this manuscript. We thank the "International Avian Phylogenomic Community" for early access to the *M. vitellinus* genome data, and T. Yuri for preparing Fig. 1. We also thank T. Parsons, S. Olson, T. Glenn and J. Blake for contributions to the fieldwork.

References

- Anderson E (1949) *Introgressive Hybridization*. John Wiley, New York.
- Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.
- Barr CM, Fishman L (2010) The nuclear component of a cytonuclear hybrid incompatibility in *Mimulus* maps to a cluster of entaticopeptide repeat genes. *Genetics*, **184**, 455–465.

- Barton NH (1983) Multilocus clines. *Evolution*, **37**, 454–471.
- Barton NH, De Cara MAR (2009) The evolution of strong reproductive isolation. *Evolution*, **63**, 1171–1190.
- Barton NH, Hewitt GM (1985) Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.
- Barton NH, Hewitt GM (1989) Adaptation, speciation and hybrid zones. *Nature*, **341**, 497–503.
- Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.
- Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society B—Biological Sciences*, **263**, 1619–1626.
- Bradshaw H, Schemske D (2003) Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers. *Nature*, **426**, 176–178.
- Brumfield RT, Jernigan RW, McDonald DB, Braun MJ (2001) Evolutionary implications of divergent clines in an avian (Manacus: Aves) hybrid zone. *Evolution*, **55**, 2070–2087.
- Buerkle CA, Rieseberg LH (2008) The rate of genome stabilization in homoploid hybrid species. *Evolution*, **62**, 266–275.
- Burke MK, Dunham JP, Shahrestani P, Thornton KR, Rose MR, Long AD (2010) Genome wide analysis of a long-term evolution experiment with *Drosophila*. *Nature*, **467**, 587–590.
- Colosimo PF, Hosemann KE, Balabhadra S et al. (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, **307**, 1928–1933.
- Dodd DMB (1989) Reproductive isolation as a consequence of adaptive divergence in *Drosophila pseudoobscura*. *Evolution*, **43**, 1308–1311.
- Ellegren H, Smeds L, Burri R et al. (2012) The genomic landscape of species divergence in ficedula flycatchers. *Nature*, **491**, 756–760.
- Ender JA (1977) *Geographic Variation, Speciation, and Clines*. Princeton University Press, Princeton, NJ.
- Epperson BK (2003) *Geographical Genetics. Monographs in Population Biology*. Princeton University Press, Princeton, NJ.
- Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.
- Feder JL, Gejji R, Yeaman S, Nosil P (2012) Establishment of new mutations under divergence and genome hitchhiking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 461–474.
- Feder JL, Nosil P (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, **64**, 1729–1747.
- Feder JL, Roethele JB, Filchak K, Niedbalski J, Romero-Severson J (2003) Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly, *Rhagoletis pomonella*. *Genetics*, **163**, 939–953.
- Fierst JL, Hansen TF (2010) Genetic architecture and postzygotic reproductive isolation: evolution of Bateson-Dobzhansky-Muller incompatibilities in a polygenic model. *Evolution*, **64**, 675–693.
- Funk DJ, Nosil P, Etges W (2006) Ecological divergence exhibits consistently positive associations with reproductive isolation across disparate taxa. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 3209–3213.
- Gavrilets S, Li H, Vose MD (1998) Rapid parapatric speciation on holey adaptive landscapes. *Proceedings of the Royal Society B—Biological Sciences*, **265**, 1483–1489.
- Gillepsie J (1984) Molecular evolution over the mutational landscape. *Evolution*, **38**, 1116–1129.
- Gompert Z, Buerkle CA (2009) A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. *Molecular Ecology*, **18**, 1207–1224.
- Gompert Z, Buerkle CA (2011a) Bayesian estimation of genomic clines. *Molecular Ecology*, **20**, 2111–2127.
- Gompert Z, Buerkle CA (2011b) A hierarchical Bayesian model for next-generation population genomics. *Genetics*, **187**, 903–917.
- Gompert Z, Buerkle CA (2012) bgc: Software for bayesian estimation of genomic clines. *Molecular Ecology Resources*, **12**, 1168–1176.
- Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, A BC (2012a) Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution*, **66**, 2167–2181.
- Gompert Z, Parchman TL, Buerkle CA (2012b) Genomics of isolation in hybrids. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 439–450.
- Harr B (2006) Genomic islands of differentiation between house mouse subspecies. *Genome Research*, **16**, 730–737.
- Hermisson J, Pennings PS (2005) Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. *Genetics*, **169**, 2335–2352.
- Hoffmann AA, Rieseberg LH (2008) Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? *Annual Review of Ecology, Evolution, and Systematics*, **39**, 21–42.
- Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.
- Joron M, Frezal L, Jones RT, et al. (2011) Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, **477**, 203–206.
- Kopp M, Hermisson J (2009) The genetic basis of phenotypic adaptation ii: the distribution of adaptive substitutions in the moving optimum model. *Genetics*, **183**, 453–476.
- Lawniczak MKN, Emrich SJ, Holloway AK et al. (2010) Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science*, **330**, 512–514.
- Lexer C, Buerkle CA, Joseph JA, Heinze B, Fay MF (2007) Admixture in European *Populus* hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differences. *Heredity*, **98**, 74–84.
- Li H, Handsaker B, Wysoker A et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
- Li R, Zhu H, Ruan J et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, **20**, 265–272.
- Linnen CR, Kingsley EP, Jensen JD, Hoekstra HE (2009) On the origin and spread of an adaptive allele in deer mice. *Science*, **325**, 1095–1098.
- Lynch M (2007) *The Origins of Genome Architecture*. Sinauer Associates, Sunderland MA.
- Lynch M, Force A (2000) The origin of interspecific genomic incompatibility via gene duplication. *American Naturalist*, **156**, 590–605.
- Maynard-Smith J, Haigh J (1974) Hitch-hiking effect of a favorable gene. *Genetical Research*, **23**, 23–35.

- McDonald DB, Clay RP, Brumfield RT, Braun MJ (2001) Sexual selection on plumage and behavior in an avian hybrid zone: Experimental tests of male-male interactions. *Evolution*, **55**, 1443–1451.
- Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P, Feder JL (2010) Widespread genomic divergence during sympatric speciation. *Proceedings of National Academy of Sciences*, **107**.
- Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J (2009) A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science*, **323**, 373–375.
- Moran P (1950) Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.
- Nadeau NJ, Whibley A, Jones RT et al. (2011) Evidence for genomic islands of divergence among hybridizing *Heliconius* butterflies obtained by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B–Biological Sciences*, **367**, 343–353.
- Navarro A, Barton NH (2003) Chromosomal speciation and molecular divergence–accelerated evolution in rearranged chromosomes. *Science*, **300**, 321–324.
- Nicholson G, Smith AV, Jonsson F, Gustafsson O, Stefansson K, Donnelly P (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society Series B–Methodological*, **64**, 695–715.
- Noor MAF, Bennett SM (2009) Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, **103**, 439–444.
- Nosil P (2004) Reproductive isolation caused by visual predation on migrants between divergent environments. *Proceedings of the Royal Society of London Series B–Biological Sciences*, **271**, 1521–1528.
- Nosil P, Crespi BJ, Sandoval CP (2002) Host-plant adaptation drives the parallel evolution of reproductive isolation. *Nature*, **417**, 440–443.
- Nosil P, Feder JL (2012) Genomic divergence during speciation: causes and consequences. *Philosophical Transactions of the Royal Society B–Biological Sciences*, **367**, 332–342.
- Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.
- Nosil P, Schluter D (2011) The genes underlying the process of speciation. *Trends in Ecology and Evolution*, **26**, 160–167.
- Orr HA (1998) The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution*, **52**, 935–949.
- Orr HA (2001) The genetics of species differences. *Trends in Ecology and Evolution*, **16**, 343–350.
- Parchman TL, Gompert Z, Mudge J, Schilkey F, Benkman CW, Buerkle CA (2012) Genomewide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, **21**, 2991–3005.
- Parsons TJ, Olson SL, Braun MJ (1993) Unidirectional spread of secondary sexual plumage traits across an avian hybrid zone. *Science*, **260**, 1643–1646.
- Payseur BA, Krenz JG, Nachman MW (2004) Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. *Evolution*, **58**, 2064–2078.
- Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: Hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, **20**, R208–R215.
- R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rieseberg LH, Whitton J, Gardner K (1999) Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, **152**, 713–727.
- Rockman MV (2012) The qtn program and the alleles that matter for evolution: All that's gold does not glitter. *Evolution*, **66**, 1–17.
- Rundle H, Nosil P (2005) Ecological speciation. *Ecology Letters*, **8**, 336–352.
- Schluter D (2009) Evidence for ecological speciation and its alternative. *Science*, **323**, 737–741.
- Stein AC, Uy JAC (2006) Plumage brightness predicts male mating success in the lekking golden-collared manakin, *manacus vitellinus*. *Behavioral Ecology*, **17**, 41–47.
- Strasburg JL, Sherman NA, Wright KM, Moyle LC, Willis JH, Rieseberg LH (2012) What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philosophical Transactions of the Royal Society B–Biological Sciences*, **367**, 364–373.
- Szymura JM, Barton NH (1986) Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina orientalis* and *B. orientalis*, near Cracow in southern Poland. *Evolution*, **40**, 1141–1159.
- Tang S, Presgraves DC (2009) Evolution of the *Drosophila* nuclear pore complex results in multiple hybrid incompatibilities. *Science*, **323**, 779–782.
- Turelli M, Barton NH, Coyne JA (2001) Theory and speciation. *Trends in Ecology and Evolution*, **16**, 330–343.
- Turner T, Hahn M, Nuzhdin S (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, **3**, 1572–1578.
- Via S (2012) Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society B–Biological Sciences*, **367**, 451–460.
- Warren WC, Clayton DF, Ellegren H et al. (2010) The genome of a songbird. *Nature*, **464**, 757–762.
- Weir BS (1979) Inferences about linkage disequilibrium. *Biometrics*, **35**, 235–254.
- White B, Cheng C, Simard F, Costantini C, Besansky N (2010) Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. *Molecular Ecology*, **19**, 925–939.
- Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.
- Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration selection balance. *Evolution*, **65**, 1897–1911.
- Yuri T, Jernigan RW, Brumfield RT, Bhagabati NK, Braun MJ (2009) The effect of marker choice on estimated levels of introgression across an avian (pipridae: *Manacus*) hybrid zone. *Molecular Ecology*, **18**, 4888–4903.

T.L.P., Z.G. and C.A.B. designed research, generated and analyzed data, and wrote the paper (T.L.P. and Z.G. contributed equally). M.J.B., D.B.M., J.A.C.U. and R.T.B. collected specimens in the field and developed the study system. G.Z. led the genome sequencing at BGI, resulting in the initial draft of the *M. vitellinus* genome, with assistance from E.J. and B.S.

Data accessibility

- Final DNA sequence assembly: doi:10.5061/dryad.85078
- File containing the locus information for each SNP, and the count of the number of reads representing each SNP state in each individual: doi:10.5061/dryad.85078
- *M. vitellinus* draft genome assembly with scaffold sequences: doi:10.5061/dryad.85078

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Full collection and accession information on each individual vouchered specimen used in this study.