MOLECULAR ECOLOGY

Molecular Ecology (2016) doi: 10.1111/mec.13645

SPECIAL ISSUE: MOLECULAR MECHANISMS OF ADAPTATION AND SPECIATION: INTEGRATING GENOMIC AND MOLECULAR APPROACHES

Geographic cline analysis as a tool for studying genome-wide variation: a case study of pollinator-mediated divergence in a monkeyflower

SEAN STANKOWSKI,* JAMES M. SOBEL† and MATTHEW A. STREISFELD*

Institute of Ecology and Evolution, 335 Pacific Hall 5289, University of Oregon, Eugene, OR 97403-5289, USA, †Department of Biological Sciences, Binghamton University, PO Box 6000, Binghamton, NY 13902, USA

Abstract

A major goal of speciation research is to reveal the genomic signatures that accompany the speciation process. Genome scans are routinely used to explore genome-wide variation and identify highly differentiated loci that may contribute to ecological divergence, but they do not incorporate spatial, phenotypic or environmental data that might enhance outlier detection. Geographic cline analysis provides a potential framework for integrating diverse forms of data in a spatially explicit framework, but has not been used to study genome-wide patterns of divergence. Aided by a first-draft genome assembly, we combined an F_{CT} scan and geographic cline analysis to characterize patterns of genome-wide divergence between divergent pollination ecotypes of *Mimulus aurantiacus*. F_{CT} analysis of 58 872 SNPs generated via RAD-seq revealed little ecotypic differentiation (mean F_{CT} = 0.041), although a small number of loci were moderately-to-highly diverged. Consistent with our previous results from the gene *MaMyb2*, which contributes to differences in flower colour, 130 loci have cline shapes that recapitulate the spatial pattern of trait divergence, suggesting that they may reside in or near the genomic regions that contribute to pollinator isolation. In the narrow hybrid zone between the ecotypes, extensive admixture among individuals and low linkage disequilibrium between markers indicate that most outlier loci are scattered throughout the genome, rather than being restricted to one or a few divergent regions. In addition to revealing the genomic consequences of ecological divergence in this system, we discuss how geographic cline analysis is a powerful but under-utilized framework for studying genome-wide patterns of divergence.

Keywords: ecotype formation, hybrid zone, *Mimulus aurantiacus*, pollinator isolation, reproductive isolation

Received 15 January 2016; revision received 7 April 2016; accepted 8 April 2016

Introduction

The last decade has seen a resurgence of interest in the role that ecological differences play in the origin of new species (Rundle & Nosil 2005; Mallet 2008; Schluter 2009; Sobel *et al.* 2010; Nosil 2012). Despite long-standing debate, most researchers have now embraced the notion that ecologically based divergent selection can generate barriers to gene flow, especially early in speciation when other forms of isolation are absent (Nosil 2012). Moreover, recent efforts have advanced the field of speciation research beyond discussions about the nature of species or the plausibility of different geographic modes of speciation, towards a mechanistic understanding of the different factors that contribute to the evolution of reproductive isolation (Butlin *et al.* 2008, 2012; Mallet 2008; Sobel *et al.* 2010; Nosil 2012; Seehausen *et al.* 2014). Broad access to high-throughput sequencing technologies has enabled key questions about speciation

Correspondence: Sean Stankowski, Fax: +(541) 346 4532; E-mail: sstankow@uoregon.edu

to be placed in a genomic context (Hohenlohe *et al.* 2010; Seehausen *et al.* 2014), and many methods commonly employed in population genetic studies have been applied to genome-wide data (e.g. Hohenlohe *et al.* 2010).

A particularly active area of research surrounds the patterns of genome-wide divergence that accompany the speciation process (Butlin *et al.* 2012; Nosil 2012; Seehausen *et al.* 2014). Outlier scans that estimate population genetic statistics at thousands to millions of loci have revealed highly heterogeneous patterns of genome-wide differentiation across a diverse array of taxa (Turner *et al.* 2005; Harr 2006; Hohenlohe *et al.* 2010; Martin *et al.* 2013; Renaut *et al.* 2013; Soria-Carrasco *et al.* 2014; Burri *et al.* 2015; Roesti *et al.* 2015; Twyford & Friedman 2015). In many cases, only a small fraction of loci are highly diverged between taxa, while the majority of the genome shows relatively low differentiation (Seehausen *et al.* 2014). Although the interpretation of this pattern is not as straightforward as once thought (Ralph & Coop 2010; Cruickshank & Hahn 2014; Burri *et al.* 2015), a common explanation is that these highly differentiated 'outlier loci' are associated with genomic regions that contribute to divergent selection and the barrier to gene flow between populations (Seehausen *et al.* 2014).

Given the relative ease with which genome-wide variation can be assayed, it is not surprising that genome scans have become a common first step in identifying candidate genomic regions that are associated with ecological divergence and speciation (Seehausen *et al.* 2014). However, traditional genome scans have a number of limitations that are likely to reduce their efficacy in identifying ecologically important loci. For example, in most methods, there is no way to integrate detailed spatial patterns of phenotypic or environmental variation into the analysis (but see Coop *et al.* 2010; Günther & Coop 2013; Gautier 2015). Rather, individuals are typically assigned to *a priori* groups based on one or more phenotypic or ecological criteria. However, morphological and ecological characteristics are often distributed continuously across geographic gradients (Endler 1977). Thus, any categorization into discrete groups may fail to capture the important variation of interest. Consequently, genome-wide analyses should ideally incorporate these diverse data in a spatially explicit framework. An additional limitation of most genome scans is that they do not take advantage of natural zones of admixture between divergently adapted populations. However, when they are present, hybrid zones can provide unique opportunities to study patterns of segregation and recombination among loci, which may reveal details about the genomic architecture of reproductive isolation (Barton & Gale 1993; Jiggins & Mallet 2000). For example, patterns of admixture and linkage disequilibrium in hybrid zones can establish how divergent loci are distributed throughout the genome, which may provide insight into how selection acts to maintain ecological differences (Barton & Gale 1993; Jiggins & Mallet 2000).

Geographic cline analysis provides a potentially powerful framework for integrating multiple forms of data into studies of genome-wide variation. Cline analysis involves fitting geographic cline models to allele frequency and/or quantitative data (e.g. phenotypic and environmental data), and has long been used as a tool for inferring the nature and strength of barriers to gene flow between hybridizing taxa (Barton & Hewitt 1985; Szymura & Barton 1986; Barton & Gale 1993; Gay *et al.* 2008). Despite the power of this approach for estimating an array of parameters, cline analysis has never been used as a tool to characterize patterns of genome-wide variation. One possible impediment is that the existing theoretical framework developed in the 1980s and 1990s requires just a handful of differentially fixed (or nearly diagnostic), unlinked loci (Barton & Hewitt 1985; Szymura & Barton 1986; Barton & Gale 1993). Indeed, recent studies have continued the tradition of removing markers with substantial allele sharing from data sets prior to analysis (e.g. Larson *et al.* 2014; Baldassarre *et al.* 2014; Lafontaine *et al.* 2015). However, some divergence histories, including adaptation from standing variation, predict substantial allele sharing at loci that are linked to divergently selected causal variants (Barrett & Schluter 2008). While these loci may not emerge as candidates in traditional genome scans, cline analysis should be able to detect spatial gradients in allele frequency differences, even if the differences are modest. Moreover, because quantitative data can be included in this framework (Bridle *et al.* 2001; Stankowski 2013), the shapes of molecular marker clines can be directly compared to clines in phenotypic traits or environmental variables, providing further support for their association with ecological differences.

In this study, we use an outlier scan and cline analysis as we begin to characterize patterns of genome-wide divergence between pollination ecotypes of the perennial shrub, *Mimulus aurantiacus*. In San Diego County, California, there is a sharp geographic transition between red- and yellow-flowered ecotypes of *M. aurantiacus* (Streisfeld & Kohn 2005). These ecotypes are extremely closely related to each other, with the western, red ecotype having evolved recently from an ancestral yellow-flowered form that is currently distributed in the east (Stankowski & Streisfeld 2015). Despite their recent shared evolutionary history, the ecotypes show striking phenotypic differences in their flowers (Fig. 1; Waayers 1996; Tulig 2000; Streisfeld & Kohn 2005). In

addition to pronounced variation in flower colour, the ecotypes differ extensively in floral size, shape and reproductive organ placement, which led Grant (1981, 1993a,b) to hypothesize that divergence was driven primarily by selection to maximize visitation and pollen transfer by alternate pollinators.

Recent studies indicate that these phenotypic differences are maintained by divergent selection acting on floral traits despite ongoing gene flow. First, hummingbird and hawkmoth pollinators demonstrate opposing preferences and constancy for flowers of the red and yellow ecotypes, respectively, which generates strong but incomplete pollinator isolation between them (Streisfeld & Kohn 2007; Handelman & Kohn 2014; Sobel & Streisfeld 2015). A *cis*-regulatory mutation in the gene *MaMyb2* is primarily responsible for the transition from yellow to red flowers, and patterns of molecular variation in the gene show clear evidence for recent divergent selection (Streisfeld *et al.* 2013; Stankowski & Streisfeld 2015). In addition, six floral traits show sharp geographic clines across San Diego County (Stankowski *et al.* 2015). These clines are all positioned in the same geographic location, suggesting that the traits are differentiated due to a common selective agent. Therefore, we expect the loci underlying the floral traits to recapitulate the spatial position of trait divergence seen across San Diego County (Stankowski *et al.* 2015).

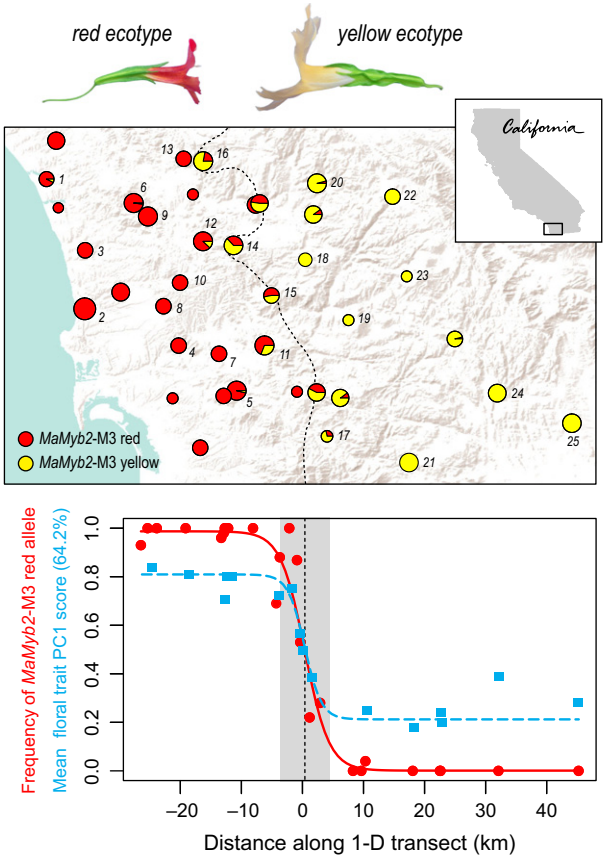

Fig. 1 Sampling locations and clinal variation in *MaMyb2*-M3 and floral traits between the red and yellow ecotypes. a) The pie charts show the frequencies of *red* and *yellow* alleles at the *MaMyb2*-M3 marker across 39 locations. The dashed line is the contour where the alternative alleles are predicted to be at a frequency of 0.5. The 25 numbered sites are the locations selected for RAD-seq. b) One-dimensional (1D) clines in *MaMyb2*-M3 allele frequency and mean floral trait PC1 score between the ecotypes. The allele frequency cline is based on allele frequencies from the 25 focus populations, while the floral trait PC1 cline (six phenotypic traits) is based on data from 16 locations (see Stankowski *et al.* 2015 for details).

In contrast to the sharp transition in floral traits and the steep cline at *MaMyb2*, a recent study of more than 5000 single nucleotide polymorphisms (SNPs) revealed a gradual spatial pattern of divergence across the region (Stankowski *et al.* 2015). In addition, analysis of floral trait variation in the narrow hybrid zone revealed an abundance of hybrid phenotypes that are indicative of extensive admixture between the divergent forms (Stankowski *et al.* 2015). These data support a long history of gene flow between the red- and yellow-flowered ecotypes and indicate that selection is responsible for maintaining floral trait differences across most of the ranges of these ecotypes. Further, endogenous postmating barriers are effectively absent between the ecotypes, indicating that gene flow is not impeded by intrinsic hybrid unfitness (Sobel & Streisfeld 2015).

Here, we generated an initial draft genome assembly for *M. aurantiacus* and identified more than 50 000 SNP markers using restriction-associated digest sequencing (RAD-seq). After revealing limited genome-wide divergence between the ecotypes, we then fit a geographic cline model to each of the 427 most differentiated markers. To our knowledge, cline models have never been fit to this many loci. We find that estimates of cline shape and position are likely to provide a more informative picture of ecological divergence than measures of allele frequency differentiation (F_{CT}). Specifically, we identified 130 loci whose cline shapes recapitulate the shape of the cline in floral trait divergence between the ecotypes, suggesting that they may reside in or near the genomic regions that contribute to pollinator isolation. Patterns of admixture and linkage disequilibrium in the hybrid zone reveal that these loci are not restricted to one or a few genomic regions, which suggests that the loci contributing to local adaptation are broadly distributed throughout the genome. We end with a discussion of the features that make geographic cline analysis a powerful but under-utilized approach for studying genome-wide variation, and highlight future directions for the further integration of cline analysis into the field of speciation genomics.

Methods

Genome sequencing and assembly

We sequenced and assembled a draft genome for *M. aurantiacus* (diploid; 10 chromosomes) using Illumina-based shotgun sequencing. We used a protocol outlined in Sobel & Streisfeld (2015) to isolate total genomic DNA from a greenhouse-grown individual of the red-flowered ecotype (Site UCSD; Table S1, Supporting information). We generated a single sequencing library by sonic shearing 1 μg of DNA, selecting the 400–600 bp size fraction and annealing paired-end T-overhang adapters to the repaired fragment ends (see supplement for a detailed protocol). After PCR enrichment of the library, 100-bp paired-end sequencing was carried out in a single lane on the Illumina HiSeq 2000 at the University of Oregon's Genomics Core Facility.

Initial processing of the raw reads was accomplished using the STACKS pipeline v. 1.12 (Catchen *et al.* 2013). The *process_shortreads* program was used with default settings to discard reads with uncalled or low-quality bases. The program *kmer_filter* was used to remove rare and abundant sequences over a range of different kmer sizes and abundance thresholds. After removing rare kmers that appeared only once and abundant kmers that were present more than 150 000 times, we used a kmer size of 69 to generate the final draft assembly using the software package VELVET v. 1.2.10 (Zerbino & Birney 2008). Contigs of a minimum size of 100 bp were retained, and summary statistics were calculated with custom scripts. Finally, as an assessment of the completeness of the gene space in our assembly, we used the *CEGMA* pipeline (Parra *et al.* 2007) to estimate the proportion of a set of 248 core eukaryotic genes (CEGs) that were completely or partially assembled. The proportion of CEGs present in an assembly has been shown to be correlated with the total proportion of assembled gene space, and thus serves as a good predictor of assembly completeness (Parra *et al.* 2007).

Samples, RAD sequencing methods and F_{CT} analysis

We identified SNPs by sequencing restriction-site-associated DNA tags (RAD-seq) generated from 298 individuals sampled from 25 locations across the range of both ecotypes and the hybrid zone (mean individuals per site = 12; range 4–18) in San Diego County, California (locations 1–25 in Fig. 1; Table S1, Supporting information). These included 11 sites within the range of the red-flowered ecotype, 8 sites within the range of the yellow-flowered ecotype, and 6 sites located in the narrow transition zone where hybrid phenotypes have been observed (Stankowski *et al.* 2015) (Table S1, Supporting information). Samples from 16 of these populations were sequenced as part of a previous study that examined phenotypic divergence and population structure between the ecotypes (Stankowski *et al.* 2015). DNA isolation and sequencing libraries were prepared using *Pst*I from an additional nine populations, following the methods described in Etter *et al.* (2011), Sobel & Streisfeld (2015) and Stankowski *et al.* (2015). Single-end 100-bp sequencing was conducted on an Illumina HiSeq 2000 at the University of Oregon's Genomics Core Facility.

We processed the raw sequencing reads, identified SNPs and called genotypes using the STACKS pipeline v. 1.29 (Catchen *et al.* 2013). Reads were filtered based on quality, and errors in the barcode sequence or RAD site were corrected using the *process_radtags* script in STACKS. Individual reads were aligned to the *M. aurantiacus* genome (described herein) using BOWTIE2 v. 2.2.5 (Langmead & Salzberg 2012), with the *very_sensitive* settings. We then identified SNPs using the *ref_map.pl* function of STACKS, with two identical raw reads required to create a stack and two mismatches allowed when processing the catalogue. SNP identification and genotype calls were conducted using the maximum-likelihood model implemented in STACKS, with alpha set to 0.01 (Hohenlohe *et al.* 2010, 2012; Catchen *et al.* 2011). We performed several independent runs in STACKS using a range of parameters for stack building and genotype calling, and all provided qualitatively similar results. To include a SNP in the final data set, we required it to be present in at least 90% of all individuals and in a minimum of 8 copies across the entire data set (i.e. minor allele frequency > 0.015).

We performed a locus-by-locus analysis of molecular variance (AMOVA) in ARLEQUIN v3.5 (Excoffier *et al.* 2005) to determine the extent and pattern of genome-wide divergence between the red and yellow ecotypes. After accounting for variation partitioned between populations within the ecotypes and within populations, we obtained the fixation index F_{CT} between the ecotypes for each SNP marker. We arbitrarily defined markers in the top 1% of the F_{CT} distribution as 'outlier loci', and used these in subsequent analyses.

Calculation of one-dimensional transect

We used one-dimensional (1D) cline analysis to explore spatial variation in allele frequencies for each outlier locus. Although the ecotypes are distributed over a broad two-dimensional landscape, the transition between them occurs in a primarily east–west direction. To allow 1D clines to be fitted to our data, we collapsed the two-dimensional sampling locations onto a 1D transect. We used empirical Bayesian kriging, a geostatistical interpolation method, to generate a prediction

surface of geographic variation in allele frequencies at the *MaMyb2*-M3 marker, which is tightly linked to the *cis*-regulatory mutation that is primarily responsible for the transition from yellow to red flowers in *M. aurantiacus* (Streisfeld *et al.* 2013; Stankowski & Streisfeld 2015). Allele frequency data for 30 sample sites have been used in previous studies to generate a 1D transect (Streisfeld *et al.* 2013; Stankowski *et al.* 2015). In this study, we included allele frequency data from nine additional locations that were sampled in spring of 2014 and genotyped according to Streisfeld *et al.* (2013) (Table S1, Supporting information).

We generated the prediction surface in ARCMAP *v.* 10.2 (Esri) (see Data S1, Supporting information for details), and determined the position of the cline centre in two dimensions by extracting the linear contour where the frequencies of both alleles are predicted to be equal (*i.e.* 0.5). This location was set to position '0'. We then obtained 1D coordinates for each of the 25 focal populations by calculating their minimum straight-line distance from the two-dimensional cline centre, resulting in sites to the west of the two-dimensional cline centre having negative 1D distance values and sites to the east having positive values.

The cline model and cline fitting procedure

With the 1D transect established, we fitted a cline model to the allele frequency data for each outlier SNP (top 1% of the F_{CT} distribution) using maximum likelihood (ML). As SNPs located within 95 bp on either side of the same restriction enzyme cut site (*PstI*) show very similar patterns of divergence (see Results), we only fitted a cline to one randomly selected SNP in each 190-bp RAD locus (RAD tags are sequenced 95 bp in each direction from each *PstI* cut site).

Several alternative models have been proposed to describe clines in molecular markers (Szymura & Barton 1986, 1991; Gay *et al.* 2008; Derryberry *et al.* 2014). While these models vary in complexity and allow different numbers of parameters to be estimated, all are different formulations or extensions of the sigmoidal *tanh* cline described in Szymura & Barton (1986, 1991). We used a *tanh* cline model to obtain 5 parameters that describe the spatial position, rate and extent of allele frequency changes across the 1D transect. These parameters, which are either estimated during the fit or derived from the ML solution, are described in greater detail in Fig. 2. The first parameter, ΔP, provides an estimate of the total change in the allele frequency difference across the transect. Like F_{CT}, ΔP can range between 0 (no difference in allele frequency across the cline) and 1 (alternative alleles fixed in each tail). However, unlike F_{CT} analysis, where the difference in allele frequency is estimated between discrete groups, ΔP is estimated from the tails of a continuous function. Thus, depending upon how the allele frequencies vary over space, estimates of F_{CT} and ΔP may differ markedly from one another. The second and third parameters, P_{max} and P_{min}, describe the frequency of allele *i* in the high and low tails of the cline, respectively, which allows us to explore the spatial pattern of allele frequency variation in more detail. For example, both alleles at a locus may be present at appreciable frequencies on one side of the cline, but one of the alleles could be fixed on the other side of the cline. In this case, the marker would show a moderate ΔP, but P_{max} and P_{min} indicate different levels of allele sharing on each side of the cline.

The fourth and fifth parameters are the cline centre (*c*) and cline slope (*m*) (Fig. 2). Assuming that a cline is at equilibrium, and that alternative alleles are maintained in different areas due to selection across an ecological gradient, the cline centre indicates the position where the direction of selection acting on an allele changes (Endler 1977; Barton & Gale 1993; Kruuk *et al.* 1999). The cline slope (*m*) indicates the rate of change in allele frequency at the maximum gradient of the function and provides a relative indication of the strength of selection acting on a locus, with cline slope increasing as the strength of selection increases (Barton & Gale 1993). Traditionally, the comparison of cline width (*w*) has been

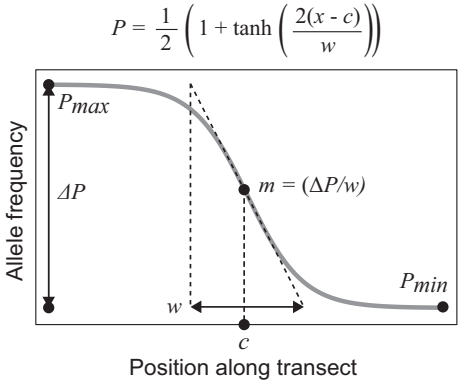

Fig. 2 The sigmoid cline model. The hyperbolic *tanh* function enables us to estimate two parameters that describe cline shape, and derive four additional parameters from the ML solution. The estimated parameters are the cline centre (*c*), which is the geographic position of the maximum gradient of the cline function, and the cline width (*w*). The derived parameters are P_{max}, the frequency of the focal allele in the high tail; P_{min}, the frequency of the focal allele in the low tail; and ΔP, the total change in the allele frequency across the transect, calculated as $P_{max} - P_{min}$. Finally, the cline slope, *m*, defined as the maximum gradient of the sigmoid function, is calculated as $\Delta P/w$.

used to infer variation in the strength of selection acting among loci (Barton & Hewitt 1985; Barton & Gale 1993). However, w is a function of ΔP and m, which means that for a given value of m, cline width decreases as ΔP decreases. In studies of genome-wide variation, we expect considerable variation in ΔP, which complicates comparisons of w among loci. Thus, when a set of markers shows considerable variation in allele sharing, cline slope has a more straightforward interpretation.

Cline fitting was conducted using *Analyse* v.1.3 (Barton & Baird 1995). For each 'outlier locus' (top 1% of the F_{CT} distribution), and for the *MaMyb2*-M3 marker, we fit a cline to the allele frequency data from the 25 sample sites, arbitrarily using the allele that was most common in the red ecotype as the focal allele. To ensure that the likelihood surface was thoroughly explored, we conducted two independent runs, each consisting of 10 000 iterations, with different starting parameters and random seeds. Each fit was visually inspected for quality. Because we were interested in identifying markers with cline shapes that were similar to floral traits, we refitted a cline to the floral trait data for the 16 populations published in Stankowski *et al.* (2015) using the new 1D transect. Rather than fitting a cline to each trait, we conducted a principal components analysis on the trait data for each individual, and scaled the data between 0 and 1 as required by the software. We then calculated the mean PC1 score for each site, and fitted a one-dimensional cline as described in Stankowski *et al.* (2015).

Given that linked loci often show similar patterns of divergence due to the effects of linked selection, we also tested whether markers in close genomic proximity have similar cline shapes. Because our genome assembly consists primarily of short scaffolds (see Results), we were unable to perform a detailed analysis of cline parameters across large chromosomal regions. Instead, for each of the five cline parameters (P_{min}, P_{max}, ΔP, c and m), we used a regression analysis to test for a relationship in the value of the parameter between all pairs of SNPs found on the same genome scaffold. We tested the significance of the relationship by comparing the observed r^2 value to a null distribution of values generated using 1 000 000 random permutations of the data using custom scripts in *R*, as described in Stankowski *et al.* (2015). Because the effects of linked selection are expected to decrease with increasing physical distance between loci, we repeated this analysis including only pairs of SNP <5 kb apart, and pairs of SNPs >5 kb apart.

Associations between outlier loci in the hybrid zone

We have shown previously that floral trait associations are greatly reduced in the hybrid zone between the ecotypes, which is consistent with ongoing gene flow between the ecotypes (Stankowski *et al.* 2015). This extensive hybridization provides us with an excellent opportunity to determine how the outlier loci are distributed throughout the genome. For example, if the loci are restricted to one or a few genomic regions, we expect the associations of alleles among loci to be maintained despite ongoing gene flow. However, if loci are spread throughout the genome, we expect the associations in the hybrid zone to be dramatically reduced as they are broken up by segregation and recombination.

We used two methods to assess the strength and pattern of the associations among the outlier loci. First, we used STRUCTURE v. 2.3.4 (Pritchard *et al.* 2000) to infer patterns of admixture based on the outlier loci and the *MaMyb2*-M3 marker, using the settings outlined in Stankowski *et al.* (2015). We then compared the distributions of admixture scores from individuals in the hybrid zone (n = 61) relative to the pure red- and yellow-flowered ecotypes (n = 215). Rather than examining data from all six hybrid populations, we restricted the analysis to the four locations where floral trait variation was previously examined (sites 11, 13, 14 and 15 in Fig. 1). This previous study revealed an abundance of intermediate and transgressive floral morphologies in the hybrid zone, suggesting extensive admixture between the ecotypes (Stankowski *et al.* 2015). If this is the case, and associations between outlier loci are being broken up by extensive hybridization, we expect a broad, unimodal distribution of hybrid index scores centred approximately intermediate of the admixture scores for the pure red- and yellow-flowered ecotypes.

Second, we examined patterns of linkage disequilibrium (LD) among the outlier loci inside versus outside the hybrid zone. Significant reductions of LD in the hybrid zone could suggest that associations between divergent markers have been reduced following gene flow. Although patterns of LD in hybrid zones reflect the interaction of diverse evolutionary processes, large reductions between divergent loci suggest that they are distributed in different genomic regions. To test for reduced LD, we first generated a theoretical expectation for LD in the absence of interecotype gene flow by pooling individuals from pure red and yellow ecotype sites into a single 'population'. Using the R package *LDheatmap* (Shin *et al.* 2006), we then calculated D' between all pairs of outlier SNPs in this pure 'population' and from samples in the hybrid zone. We then used the *heatmap2* function in the *R* package *gplots* to cluster sets of markers with similar D' estimates in each of the pairwise matrices. We then qualitatively evaluated whether the number of groups of loci that showed tight LD with one another in the hybrid zone was reduced relative to the pooled pure 'population'. We also tested for quantitative reductions in the mean

estimates of D' within the hybrid zone relative to the mean estimate obtained from our pooled population using a permutation test (100 000 permutations). However, because physical linkage and selection can influence associations between alleles among loci, we also tested for differences in mean LD inside and outside the hybrid zone for (i) pairs of markers located on different genome scaffolds, (ii) all pairs of linked loci located on the same genome scaffold, (iii) pairs of linked loci <5 kb apart, (iv) pairs of linked loci >5 kb apart and (v) all pairwise combinations that included the *MaMyb2*-M3 marker, which is in the divergently selected flower colour locus *MaMyb2*.

Results

Draft genome assembly for Mimulus aurantiacus

We obtained 176 451 402 raw paired reads from a single lane of Illumina HiSeq PE100 sequencing of one red ecotype individual, which is equivalent to an average of 118× coverage of the estimated 297 Mbp genome size (Murovec & Bohanek 2013). The final draft assembly consisted of 23 129 scaffolds larger than 500 bp and totalled 223.8 Mbp (74% of the estimated genome size). The N50 scaffold length was 31 153 bp, and the largest scaffold was 209 453 bp. *CEGMA* analysis identified partial or complete sequences for 240 of the 248 CEGs (97%), with 208 of them being completely assembled.

F_{CT} analysis reveals low genome-wide divergence between the ecotypes

To explore genome-wide divergence between the ecotypes, we sequenced RAD tags from 298 individuals, aligned the reads to our draft assembly and identified 58 872 SNPs that met our filtering requirements. A locus-by-locus analysis of F_{CT} for these markers revealed a highly skewed distribution of genetic divergence between the ecotypes, with most loci showing little or no differentiation (Fig. 3a). Estimates of F_{CT} ranged from −0.062 to 0.853, with a mean interecotype differentiation of 0.041 (SD 0.075). The top 1% of the F_{CT} distribution (n = 589) spanned approximately half of the total range of values among RAD markers, with a minimum value of 0.358. These 'outlier' SNPs showed moderate differentiation between the ecotypes (mean 0.451, SD 0.084), but none of the markers were as highly differentiated as the *MaMyb2*-M3 marker (F_{CT} = 0.98).

The 589 SNPs in the top 1% of the F_{CT} distribution are located in 426 distinct RAD loci. We expected that SNPs found in the same locus would show very similar levels of interecotype differentiation. Indeed, a pairwise regression of F_{CT} among pairs of these 163 SNPs was

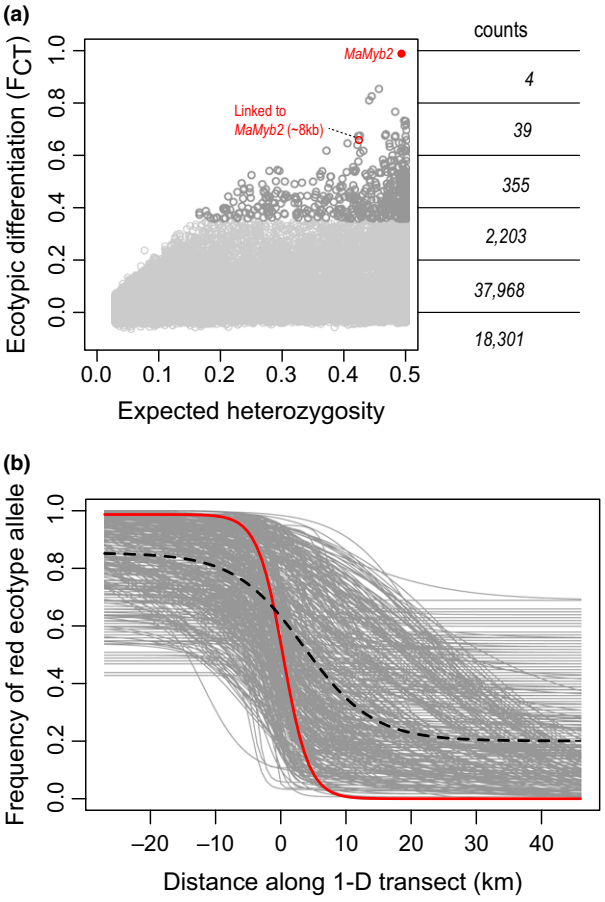

Fig. 3 Differentiation of SNP markers between the red and yellow ecotypes. (a) Plot of F_{CT} on expected heterozygosity for 58 872 RAD markers and the *MaMyb2*-M3 marker. SNPs in the top 1% of the F_{CT} distribution are coloured dark grey (n = 589). The counts show the density of markers in six bins. The *MaMyb2*-M3 marker and a SNP located approximately 8 kb from *MaMyb2* are highlighted. (b) Geographic clines for the top 1% of the F_{CT} distribution, including only a single SNP marker per RAD cut site (n = 426). The red line shows the cline for the *MaMyb2*-M3 marker; the dashed line is the average cline based on the mean parameter estimates across the 426 markers.

highly significant and explained more than 90% of the variation in interecotype differentiation (r^2 = 0.91; permutation test $P = 9.99 \times 10^{-7}$; Fig. S1, Supporting information). Therefore, we conducted further analyses based on data from a single randomly selected SNP from each of the 426 unique 190-bp RAD loci represented in the top 1% of the F_{CT} distribution, as well as a marker in the *MaMyb2* gene (*MaMyb2*-M3).

Geographic cline analysis reveals extensive variation in the spatial pattern of divergence among loci

Consistent with previous results and the role of *MaMyb2* in flower colour evolution (Streisfeld *et al.*

2013; Stankowski *et al.* 2015), the cline in the floral traits and *MaMyb2*-M3 marker had very similar shapes (Fig. 1). In contrast, we observed a diverse array of cline shapes for the 426 'outlier loci' obtained from RAD sequencing (Fig. 3b). The average 1D cline (calculated from the mean of all parameter estimates) was shallower than the cline for the *MaMyb2*-M3 marker, in terms of both the total change in allele frequency across the 1D transect ($\Delta P_{mean} = 0.66$ vs. $\Delta P_{MaMyb2} = 0.99$) and cline slope ($m_{mean} = 0.032$ vs. $m_{MaMyb2} = 0.125$). The average cline centre was shifted approximately 5 km to the east of the centre for the floral trait cline.

Given the difficulty of drawing conclusions from visual analysis of so many clines, we examined the distributions of the parameters that describe cline shape (Fig. 4). As with F_{CT}, the three parameters involving allele frequency change in the tails (ΔP, P_{max} and P_{mim}) revealed considerable allele sharing between the ecotypes (Fig. 4). However, examination of the cline parameters provides additional information about the spatial patterns of this allele sharing that is not possible from estimates of F_{CT}. Specifically, although 77% of markers were at or near fixation in at least one tail of the cline (40% had $P_{max} > 0.90$ in the left tail and 37% had $P_{min} < 0.10$ in the right tail; Fig. 4a), 75% of the 427 markers had a $\Delta P < 0.8$ (Fig. 4b). Thus, despite most markers being near fixation on one side of the cline, both alleles were often at appreciable frequency in the opposite tail.

Patterns of variation in the remaining two parameters, cline centre (c) and slope (m), revealed a subset of loci with cline shapes that recapitulated the pattern and scale of floral trait divergence across San Diego County. First, despite broad variation in the estimates of c (range −13 to 30 km), 60% of markers had ML estimates of cline centre that coincided with the narrow phenotypic transition zone between the ecotypes, as defined by the width of the floral trait cline (w for the mean floral trait PC1 score = −3.5 to 3.5 km; Fig. 4c). Cline slope varied more than 40-fold among loci, with estimates of m ranging from 0.008 to 0.344 (Fig. 4d). One-third of markers had slopes that were greater than the average for all 427 markers (mean $m = 0.053$), including a RAD locus located approximately 8 kb from

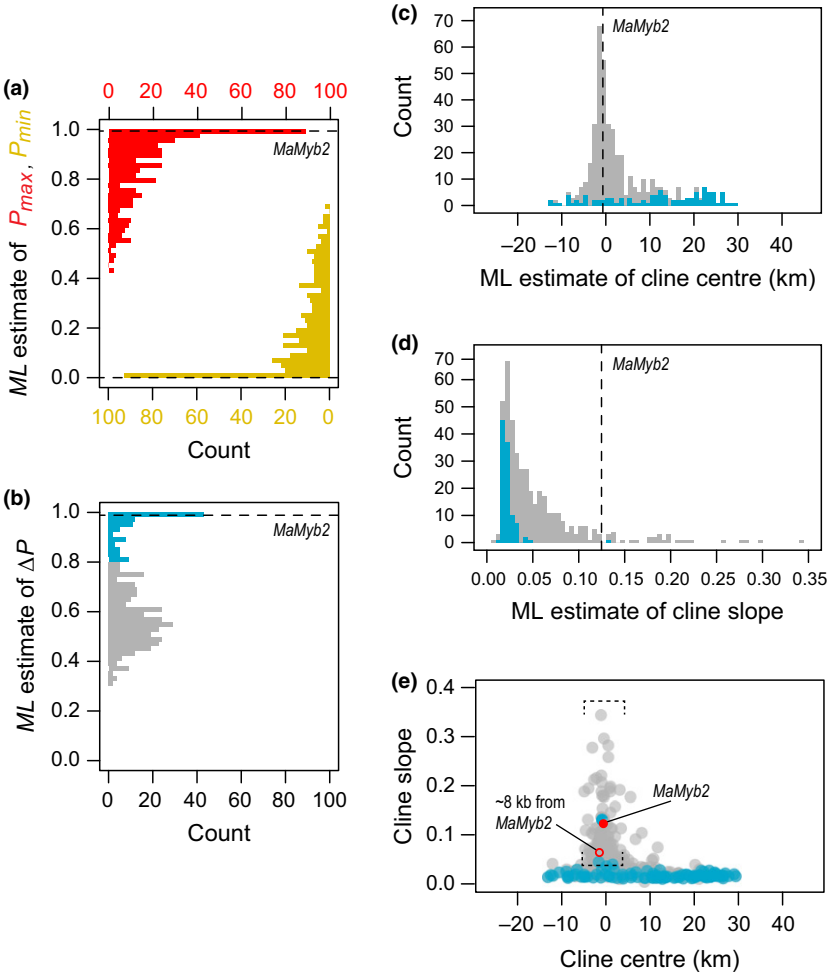

Fig. 4 Distributions of ML cline parameters for the 426 RAD outlier loci. (a) Distributions of allele frequencies for *Pmax* (red) and *Pmin* (yellow) in the left and right tails of each cline. (b) Change in allele frequency across the cline (ΔP). (c) Distribution of cline centre, (d) cline slope and (e) the relationship between cline centre and cline slope. The dashed black lines in plots a–d show the ML estimates for the *MaMyb2*-M3 cline. In panels b–e, the blue bars and points are for markers where ΔP is > 0.8, and the grey bars and points show the distribution for markers where $\Delta P < 0.8$. In panel e, points within the brackets have centres that coincide with the geographic transition in floral traits, and have cline slopes above the average for all 427 loci; the filled and open red circles represent the *MaMyb2*-M3 marker and linked RAD marker, respectively.

the flower colour gene *MaMyb2*. Fifty-six markers had slopes that were greater than the slope for the *MaMyb2*-M3 marker ($m = 0.125$). Finally, we observed a striking relationship between cline centre and cline slope (Fig. 4e), with the sharpest clines coinciding with the geographic position of the cline centre in floral traits. Specifically, 130 marker clines had cline centres that coincided with the floral trait cline and whose slopes were elevated above the average for the 427 markers ($m_{mean} = 0.053$). This included both the *MaMyb2*-M3 marker and the RAD marker located approximately 8 kb from *MaMyb2*.

Curiously, our cline analysis also revealed that markers showing the largest differences in allele frequency in both tails ($\Delta P > 0.8$) tended to have cline shapes that were discordant from the spatial pattern of trait divergence between the ecotypes (Fig. 4d). Specifically, these markers tended to have the shallowest slopes, and cline centres that were shifted to the east of the *MaMyb2*-M3 cline (Fig. S2, Supporting information). Rather, the markers that showed cline shapes that recapitulated the spatial transition in the floral traits tended to show moderate differences in allele frequency across the transect ($\Delta P < 0.8$) (Fig. S2, Supporting information).

SNPs in close genomic proximity have similar cline shapes

Using our draft genome assembly, we tested whether SNPs in the same genomic regions had similar cline shapes. In support of this hypothesis, a permutation-based regression including 97 pairs of loci from the 64 genomic scaffolds containing more than one outlier SNP (mean distance between SNPs = 9.7 kb, s.d. = 14.3 kb; maximum distance = 74.3 kb) explained 40% of the variation in ΔP ($r^2 = 0.400$, $P = 9.99 \times 10^{-7}$), 51% and 37% of the variation in P_{max} and P_{min}, respectively (P_{max}: $r^2 = 0.509$, $P = 9.99 \times 10^{-7}$; P_{min}: $r^2 = 0.373$, $P = 9.99 \times 10^{-7}$), 51% of the variation in cline centre (*c*: $r^2 = 0.505$, $P = 9.99 \times 10^{-7}$) and 35% of the variation in cline slope (*m:* $r^2 = 0.353$; $P = 4.99 \times 10^{-6}$; Fig. S3, Supporting information).

Because the effects of linked selection are expected to be greater for tightly linked loci than for those located farther apart, we binned pairs of SNPs from the same scaffold that were located <5 kb apart ($n = 50$), and >5 kb apart ($n = 47$). As predicted, SNPs in close proximity showed very similar estimates of all five cline parameters (r^2 range 0.62 to 0.78, $P = 9.99 \times 10^{-7}$), while pairs of SNPs >5 kb apart showed far less similarity in the parameters that describe cline shape (ΔP: $r^2 = 0.150$, $P = 0.023$; P_{max}: $r^2 = 0.07$, $P = 0.023$; P_{min}: $r^2 = 0.21$, $P = 0.007$; *c* $r^2 = 0.321$, $P = 0.002$; *m:* $r^2 = 0.072$; $P = 0.087$).

F_{CT} is a poor predictor of variation in cline shape parameters

We used linear regression to test for a relationship between ecotypic differentiation (F_{CT}) and the ML estimates of each cline parameter obtained for the 427 markers. In general, F_{CT} was a poor predictor of the parameters that describe cline shape. Although highly significant, only 3.4% of the variation in ΔP was explained by the estimates of F_{CT} ($r^2 = 0.0346$, $P < 0.0001$). For the other cline parameters, the relationships were even weaker. Estimates of F_{CT} explained only 1.4% of the variation in cline centre ($r^2 = 0.0137$, $P = 0.015$), and 1.9% of the variation in cline slope ($r^2 = 0.0193$, $P = 0.0041$).

Associations among outlier loci are reduced in the face of gene flow

The observed patterns of admixture and linkage disequilibrium in the hybrid zone suggest these outlier loci are scattered broadly throughout the genome. Based on admixture scores, individuals from pure red- and yellow-flowered sample sites were generally assigned into alternative clusters with high probability (Fig. 5a). A few individuals were clear outliers in each distribution, suggesting they were either hybrids or pure individuals of the alternative ecotype. In contrast, the distribution of hybrid index scores in the hybrid zone spanned nearly the full range of values and had an approximately unimodal shape centred intermediate of the pure ecotypes. This extensive admixture indicates that the alternative alleles among these outlier loci are often inherited independently of one another in hybrid offspring.

Similarly, we observed significantly reduced linkage disequilibrium (LD) in the hybrid zone relative to the level expected if the ecotypes coexisted without gene flow between them (Fig. 5). Pairwise LD calculated in the hybrid zone was significantly reduced compared to the pooled 'population' containing pure red- and yellow-flowered individuals (mean $D'_{RY} = 0.60$, mean $D'_{HZ} = 0.31$; $P = 9.999 \times 10^{-5}$). In contrast to the large clusters of markers that showed high LD in the pooled population, much smaller clusters were detected in the hybrid zone (Fig. 5b). We also observed significantly reduced LD in the hybrid zone between markers on different genome scaffolds (mean $D'_{RY} = 0.60$, mean $D'_{HZ} = 0.32$; $P = 9.999 \times 10^{-5}$), as well as for pairwise comparisons that included the *MaMyb2*-M3 marker (mean $D'_{RY} = 0.73$, mean $D'_{HZ} = 0.26$; $P = 9.999 \times 10^{-5}$) (Fig. 5c). In contrast, we observed only marginally lower LD in the hybrid zone between linked markers that are <74 kb apart (mean $D'_{RY} = 0.92$, mean

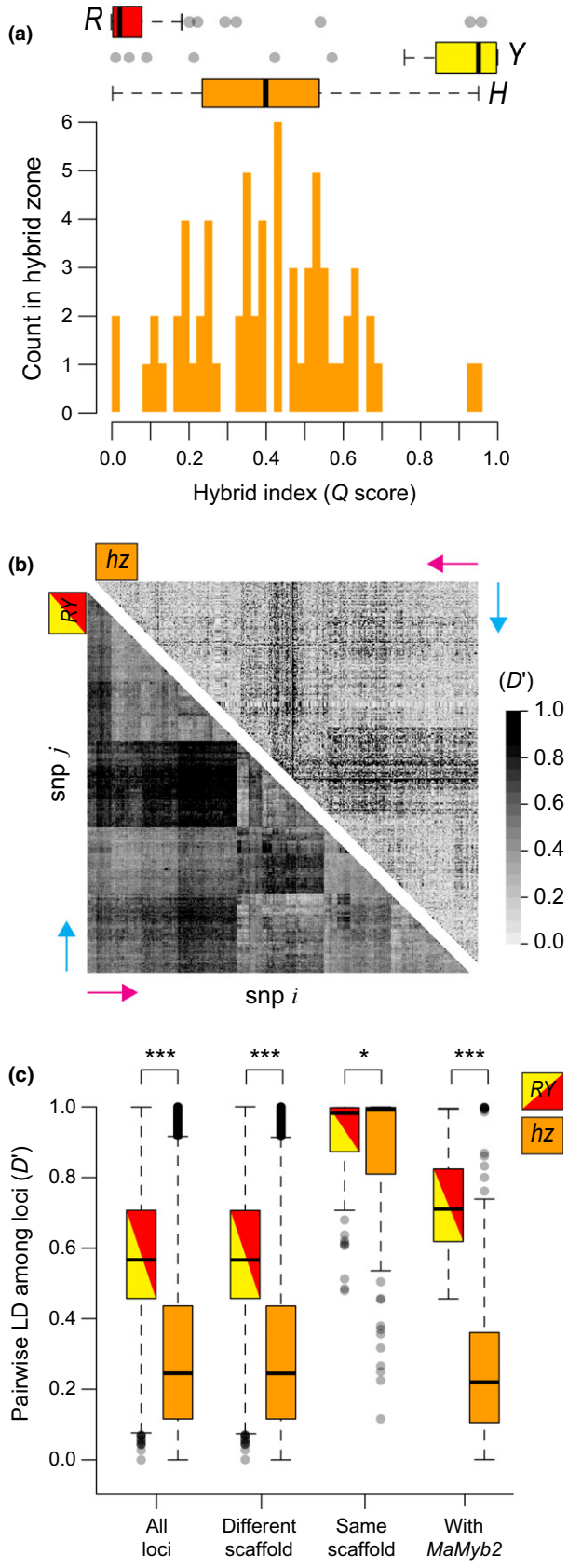

Fig. 5 Pattern of admixture and linkage disequilibrium in the hybrid zone. (a) Distributions of hybrid index scores (*Structure* *Q* score) inside and outside the hybrid zone. The boxplots show the distributions of *Q* scores for the red (*R*) and yellow (*Y*) ecotypes, and four sample sites inside the hybrid zone (*hz*). The histogram shows a detailed view of the distribution of *Q* scores inside the hybrid zone. (b) Heatmaps of linkage disequilibrium (*D*′) between all pairs of outlier loci ($n = 427$) inside the hybrid zone (*hz*) and when the pure red and yellow ecotypes are pooled together (*RY*). The order of markers is the same in both matrices; the coloured arrows assist comparing the matrices, as they are oriented differently. c) Boxplots showing the distribution of linkage disequilibrium (*D*′) for all outlier loci, loci on different genome scaffolds, loci on the same genome scaffold, and for comparisons that include the *MaMyb2*-M3 marker. Asterisks indicate the level of significance (based on permutation tests) of mean differences in pairwise *D*′ for comparisons between the hybrid zone and the pooled 'population' containing pure red and yellow ecotype individuals (* $P = 0.02$; *** $P < 9.999 \times 10^{-5}$).

$D'_{HZ} = 0.86$; $P = 0.0215$). Further analysis revealed that this modest but significant decrease in LD between linked loci is due primarily to reduced *D*′ between pairs of SNPs that are separated by more than 5 kb ($D'_{RY} = 0.88$, $D'_{HZ} = 0.71$; $P = 0.008$), as there was no significant difference in mean pairwise LD between the hybrid population and the pooled red -yellow population for pairs of SNPs that were <5 kb apart ($D'_{RY} = 0.96$, $D'_{HZ} = 0.94$; $P = 0.302$).

Discussion

In this study, we combine an F_{CT} scan and geographic cline analysis to reveal the genomic signatures of pollinator-mediated divergence between red and yellow ecotypes of *Mimulus aurantiacus*. Overall, our results reveal low genome-wide divergence between the ecotypes, further supporting the conclusion that these taxa are at an early stage of divergence (Sobel & Streisfeld 2015; Stankowski *et al.* 2015). By contrast, the markers with the steepest clines closely align with the spatial transition in floral traits, suggesting that these loci may reside in or near the genomic regions that contribute to pollinator isolation. Moreover, by taking advantage of the natural hybrid zone between the ecotypes, our data indicate that gene flow and recombination have been extensive, suggesting that the outlier loci are not concentrated in one or a few regions of the genome. In addition to elucidating the genomic consequences of pollinator-mediated reproductive isolation in this system, we end by discussing the utility of cline analysis as a spatially explicit framework for future studies of genome-wide variation.

Genome-wide divergence between the ecotypes

Consistent with a recent origin of the red ecotype from an ancestral yellow-flowered population (Stankowski & Streisfeld 2015), our F_{CT} scan revealed very limited genome-wide differentiation between the ecotypes (mean estimate of $F_{CT} = 0.046$). Such low levels of differentiation are predicted in population pairs that are at a very early stage in the speciation process (Feder *et al.* 2012; Nosil 2012; Seehausen *et al.* 2014). Indeed, our estimate is similar in magnitude to other closely related ecotypes where divergence occurred recently despite gene flow, including apple and hawthorn races of *Rhagoletis pomonella* ($F_{ST} = 0.035$; Egan *et al.* 2015), wave and crab ecotypes of the intertidal snail *Littorina saxatilis* ($F_{ST} = 0.027$; Butlin *et al.* 2014) and normal and dwarf forms of the lake whitefish *Coregonus clupeaformis* ($F_{ST} = 0.046$; Hebert *et al.* 2013).

In addition to revealing the overall level of genome-wide divergence between the ecotypes, our primary goal was to identify loci associated with pollinator isolation in this system. Recent theoretical and empirical studies suggest that regions of the genome that contribute to ecological divergence with gene flow should show elevated differentiation relative to selectively neutral regions (Feder *et al.* 2012; Nosil 2012; Seehausen *et al.* 2014). However, depending on the strength and timing of selection and the local recombination rate, genome scans based on reduced representation approaches that rely on tight linkage to selected sites may often be underpowered. While we cannot rule this possibility out, the presence of a highly differentiated RAD marker (F_{CT} =0.66) that is on the same scaffold and ~8 kb away from *MaMyb2* suggests that it has become diverged due to the effects of linked selection on flower colour. However, traditional genome scans may have a limited capacity for revealing associations between molecular and phenotypic divergence, particularly if trait variation is continuously distributed in space.

As a consequence, we employed cline analysis to further test whether the outlier loci are diverged due to spatially varying selection. While the point estimates of F_{CT} for most of these loci are modest, the estimates of the allele frequencies in each tail (P_{min} and P_{max}) reveal a complex spatial pattern of allele sharing on both ends of the cline. Almost 80% of the markers are at or near fixation for one allele in one ecotype, while both alleles are present at appreciable frequencies in the other ecotype. This pattern could result from selective sweeps on novel alleles in one of the ecotypes followed by dispersal of that allele to the other ecotype (Pritchard *et al.* 2010), or from selection on standing variation in only one of the ecotypes (Barrett & Schluter 2008). While additional data will be necessary to distinguish between these hypotheses, our analyses provide strong support that selection is responsible for the divergence of these markers despite gene flow.

The estimates of cline centre and cline slope provide the most compelling evidence that these loci are associated with floral trait divergence. In a previous study, we showed sharp, coincident clines across the transect for six divergent floral traits (Stankowski *et al.* 2015). The shape of these clines contrasts with the shallow gradient in genome-wide differentiation, suggesting that the divergent floral traits have been maintained by a common selective agent despite ongoing gene flow, rather than reflecting recent secondary contact (Streisfeld & Kohn 2005; Stankowski *et al.* 2015). Thus, we predicted that loci associated with pollinator isolation should show cline shapes that recapitulate the spatial transition in floral traits. Indeed, we observed 130 RAD markers that have sharp clines and coincide with the narrow phenotypic transition zone between the ecotypes. In ecological models of cline formation and maintenance, the cline centre represents the geographic position where the direction of selection switches to favour the alternative form of a trait (i.e. Haldane 1948; Endler 1977; Kruuk *et al.* 1999). Thus, these data are consistent with divergence of these loci due to pollinators, which generally require divergence in multiple traits to maximize attraction and successful pollen transfer (Fenster *et al.* 2004). Moreover, the clines with the steepest slopes were almost exclusively positioned in this region. Indeed, 56 markers show cline slopes that are steeper than the *MaMyb2*-M3 marker. While some may be physically linked to *MaMyb2,* the vast majority of these markers show weak LD with the *MaMyb2*-M3 marker in the hybrid zone, suggesting that they are located in different genomic regions. Thus, even though allele frequency differences are generally modest, the relationship between cline centre and slope suggests that many of these loci are associated with the primary barrier to gene flow between these ecotypes (Sobel & Streisfeld 2015).

Although the remarkable geographic coincidence of the trait and SNP clines is consistent with them becoming diverged via pollinator-mediated selection, there are other explanations for this pattern that must be considered. First, some of these markers could be differentiated due to selective gradients that are unrelated to pollinators but positioned in the same geographic location as the floral traits (Barton & Hewitt 1985). However, there is currently little evidence to support this conclusion. While other nonfloral traits differ between the ecotypes (Hare 2002; J. M. Sobel in prep), they show linear rather than sigmoid shaped clines across the study area (J. M. Sobel in prep). Thus, we would expect

loci associated with these differences in abiotic factors to match the more gradual transitions in these traits. Future studies may reveal that some of the outlier loci that show shallow slopes but large differences in allele frequency across the transect may be associated with nonfloral adaptations.

Another alternative explanation for the association of the marker and floral trait clines is that multiple independent isolating barriers have become coupled together (Barton & De Cara 2009; Bierne *et al.* 2011; Abbott *et al.* 2013). For example, in systems where there are intrinsic incompatibilities between ecologically divergent taxa, endogenous clines are likely to be attracted to and become stabilized by ecological barriers to gene flow (Bierne *et al.* 2011). Thus, clines that coincide with an ecological boundary need not reside within the genomic regions that contribute to ecological divergence (Bierne *et al.* 2011). However, endogenous barriers to gene flow are effectively absent between the red and yellow ecotypes (Sobel & Streisfeld 2015), suggesting that intrinsic barriers to gene flow are not affecting our ability to identify the loci associated with ecological divergence. Thus, while additional studies will be necessary to determine whether non-pollinator-mediated selection is maintaining the divergence of these markers, current evidence suggests that those with steep coincident clines likely reside in the genomic regions that underlie the divergent floral traits.

In addition to identifying highly differentiated loci, we also gained insight into how these outliers are distributed throughout the genome. While some studies have shown that outlier loci are scattered across the genome (Egan *et al.* 2015; Gompert *et al.* 2014; Roesti *et al.* 2015), others have found that they are concentrated in one or a few narrow genomic regions that often have diverse phenotypic effects (Lowry & Willis 2010; Fishman *et al.* 2013; Poelstra *et al.* 2014). Indeed, the co-localization of adaptive loci appears to have facilitated rapid and robust adaptation in several examples of divergence with gene flow by limiting breakdown of linkage disequilibrium in hybrids (Jones *et al.* 2012; Joron *et al.* 2011; Lowry & Willis 2010; Twyford & Friedman 2015). At this point, our genome assembly consists of relatively short scaffolds, which limits our ability to establish the physical relationships among all of the outlier loci. However, our analysis in the hybrid zone suggests that the outlier loci do not co-localize to a small number of genomic regions. Specifically, we observed extensive variation in admixture scores in the hybrid zone, indicating that the divergent alleles are inherited largely independently of one another. In addition, we detected significant reductions in linkage disequilibrium (LD) between pairs of markers in the hybrid zone. Although this could result from the breakup of a group of tightly linked loci due to extensive fine-scale recombination in hybrids, our analysis suggests that this is not the case. Indeed, pairs of loci on the same genome scaffold (maximum of 74 kb apart) showed higher mean LD both inside and outside the hybrid zone compared with loci on different scaffolds. Thus, our results suggest that many of the outlier loci reside either on different chromosomes or in loosely linked colinear regions of the same chromosome. This result is in agreement with the substantial breakup of divergent floral traits in these same hybrid populations, and in an experimental F_2 population where selection was relaxed (Stankowski *et al.* 2015). Future studies that take advantage of chromosome-length genomic scaffolds will help to resolve the full extent of LD across the genome to determine the potential for structural variants that might limit recombination among loci.

Geographic cline analysis as a tool for studying genome-wide divergence

In addition to characterizing the genomics of floral divergence between the red and yellow ecotypes of *M. aurantiacus*, an additional goal of our study was to explore the utility of geographic cline analysis as a tool for studying genome-wide patterns of variation. Geographic cline analysis has long been used for studying barriers to gene flow between closely related taxa. However, despite a call for better integration of cline theory into genomic studies of adaptive divergence and speciation (Bierne *et al.* 2011), geographic cline analysis has not been applied to large data sets with the explicit purpose of studying patterns of genome-wide variation. This seems to reflect the history of development and use of cline analysis and its associated theory, and the relative difficulty of applying cline analysis to large marker data sets.

While geographic clines and hybrid zones have long been recognized as excellent systems for developing and testing ideas about speciation (Huxley 1939; Haldane 1948; Bazykin 1969; Clarke 1966; Endler 1977), most theory was developed in the 1980s and 1990s to provide a framework for making inferences about the nature, strength and symmetry of reproductive isolation between hybridizing taxa (Barton 1983; Barton & Hewitt 1985; Szymura & Barton 1986; Mallet & Barton 1989; Barton & Gale 1993). Although these methods require just a handful of differentially fixed, unlinked loci, cline analysis can be applied to linked loci and loci with considerable allele sharing if the goal is to study patterns of genome-wide variation. For example, consider a scenario where local adaptation across a sharp ecological gradient arises from selection on standing genetic variation. In this case, considerable allele sharing between divergent populations is expected at sites linked to the

causal variants (Hermisson & Pennings 2005; Pritchard *et al.* 2010). Although allele frequency differences at linked sites may be relatively small, any difference in allele frequency between populations should manifest itself as a sharp cline due to the indirect effects of selection on linked variants. Indeed, a model of ecological cline maintenance for a selected locus and a tightly linked neutral locus predicts the formation of clines with similar shapes, although the level of allele sharing is higher at the neutral marker (Durrett *et al.* 2000). Similarly, in polygenic models of adaptation, where traits are controlled by many loci each of small effect, smaller differences in allele frequency are expected among diverging populations even at causal loci (Pritchard *et al.* 2010).

In our study, the loci with cline shapes that recapitulated the spatial patterns of floral trait variation were associated with modest allele frequency differences. In contrast, only 106 of the 427 outlier loci showed an allele frequency difference across the transect >0.8 ($\Delta P > 0.8$). Even more striking, these markers tended to have cline shapes that were neither coincident nor concordant with the clines in floral traits. Rather, they tended to show very broad clines that were often positioned large distances from the transition in the floral traits. While these loci may be associated with other forms of local adaptation or reflect a potentially complex history of divergence, our analysis suggests that they do not make a major contribution to floral trait divergence. Thus, if we had limited our cline analysis only to these loci, as is usually done, we would draw very different conclusions about the pattern of genome-wide divergence in this system.

Another reason that cline analysis has not been applied to genome-wide data is that it is more complex and computationally intensive compared with other widely used methods. For example, F_{ST} and other similar measures of differentiation are easily calculated, even for thousands to millions of loci. Genomic cline analyses, which fit functions to allele frequency data plotted against a hybrid index instead of geographic distance, are fully automated (i.e. Gompert & Buerkle 2012). However, under certain situations, geographic cline analysis has several advantages over these methods. For example, traditional F_{ST} and genomic cline analyses cannot incorporate spatial, phenotypic, or environmental data to assist in the identification of ecologically important loci (but see Coop *et al.* 2010; Günther & Coop 2013; Gautier 2015). Moreover, our analysis revealed that key cline parameters, including the cline centre and slope, showed only weak correlations with F_{CT}. These results suggest that cline analysis provides a more informative view of ecological divergence in this system, that can be related directly to the patterns of divergence in ecologically important traits.

As a preliminary study of clinal variation in this system, we only fitted clines to the top 1% of the F_{CT} distribution. Given that many of the loci that show sharp geographic clines also show relatively modest differences in allele frequency between the ecotypes, other loci of interest may lie further down the F_{CT} distribution. Thus, rather than conducting cline analysis only on 'highly differentiated' loci, full genome-wide cline analyses of all variable sites may be conducted independent of other genome scans. Efficient software is now available to allow the automated fitting of cline models to large data sets (Derryberry *et al.* 2014). Ideally, the fit of a set of alternative cline models of varying complexity would be compared to one another, and to the fit of a null model ($m = 0$) to identify loci that show clinal variation. Estimated cline parameters for each locus may be mapped across chromosomes to reveal the consequences of selection and reproductive isolation across the entire genome. While the small scaffolds in our current genome assembly currently preclude this, our analysis of SNPs from the same genomic scaffold indicates that cline shape parameters are correlated across small chromosomal regions, demonstrating the potential of this method in our system. The interpretation of genome-wide patterns of clinal variation will be aided by genomic models of cline formation and maintenance across a range of divergence histories. Integration of these genome-wide patterns of divergence with studies of quantitative trait locus (QTL) mapping of trait variation, historical demographic modelling and haplotype-based analyses will enhance our understanding of ecological divergence in this system, and other examples of divergence with gene flow.

Acknowledgements

We would like to thank Susie Bassham for advice on sequencing library preparation, Josh Burkhart for assistance with the assembly and Julian Catchen for modifying the STACKS pipeline. Madeline Chase, Thomas Nelson and William Cresko provided fruitful discussion. Lorne Curran provided computer support. We would also like to thank Sean Rogers, Philipp M. Schlüter and Shuqing Xu for organizing and editing this special issue. The project was supported by National Science Foundation grant DEB-1258199.

References

Abbott R, Albach D, Ansell S *et al.* (2013) Hybridization and speciation. *Journal of Evolutionary Biology*, **26**, 229–246.

Baldassarre DT, White TA, Karubian J, Webster MS (2014) Genomic and morphological analysis of a semipermeable avian hybrid zone suggests asymmetrical introgression of a sexual signal. *Evolution*, **68**, 2644–2657.

Barrett RD, Schluter D (2008) Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, **23**, 38–44.

Barton NH (1983) Multilocus clines. *Evolution*, **37**, 454–471.
Barton NH, Baird SJE (1995) Analyse: an application for analysing hybrid zones. Freeware, Edinburgh isolation. *Evolution*, **63**, 1171–1190.
Barton NH, De Cara MAR (2009) The evolution of strong reproductive isolation. *Evolution*, **63**, 1171–1190.
Barton NH, Gale KS (1993) Genetic analysis of hybrid zones. In: *Hybrid Zones and the Evolutionary Process* (ed Harrison R. G.), pp. 13–45. Oxford University Press, Oxford, UK.
Barton NH, Hewitt GM (1985) Analysis of hybrid zones. *Annual Review Ecology and Systematics*, **16**, 113–148.
Bazykin AD (1969) Hypothetical mechanism of speciaton. *Evolution*, **23**, 685–687.
Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.
Bridle JR, Baird SJ, Butlin RK (2001) Spatial structure and habitat variation in a grasshopper hybrid zone. *Evolution*, **55**, 1832–1843.
Burri R, Nater A, Kawakami T *et al.* (2015) Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula flycatchers*. *Genome Research*, **25**, 1656–1665.
Butlin RK, Galindo J, Grahame JW (2008) Sympatric, parapatric or allopatric: the most important way to classify speciation? *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**, 2997–3007.
Butlin RK, Debelle A, Kerth C *et al.* (2012) What do we need to know about speciation? *Trends in Ecology & Evolution*, **27**, 27–39.
Butlin RK, Saura M, Charrier G *et al.* (2014) Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. *Evolution*, **68**, 935–949.
Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171–182.
Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.
Clarke B (1966) The evolution of morph-ratio clines. *American Naturalist*, **100**, 389–402.
Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.
Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.
Derryberry EP, Derryberry GE, Maley JM, Brumfield RT (2014) HZAR: hybrid zone analysis using an R software package. *Molecular Ecology Resources*, **14**, 652–663.
Durrett R, Buttel L, Harrison R (2000) Spatial models for hybrid zones. *Heredity*, **84**, 9–19.
Egan SP, Ragland GJ, Assour L *et al.* (2015) Experimental evidence of genome wide impact of ecological selection during early stages of speciation with gene flow. *Ecology Letters*, **18**, 817–825.
Endler JA (1977) *Geographic Variation, Speciation, and Clines*. Princeton University Press, Princeton, New Jersey.
Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (2011) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In: *Molecular Methods for Evolutionary Genetics* (eds Orgogozo V, Rockman MV), vol. **772**, pp. 157–178. Humana Press, New York City.
Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**, 47–50.
Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342–350.
Fenster CB, Armbruster WS, Wilson P, Dudash MR, Thomson JD (2004) Pollination syndromes and floral specialization. *Annual Review of Ecology, Evolution, & Systematics*, **35**, 375–403.
Fishman L, Stathos A, Beardsley PM, Williams CF, Hill JP (2013) Chromosomal rearrangements and the genetics of reproductive barriers in *Mimulus* (monkey flowers). *Evolution*, **67**, 2547–2560.
Gautier M (2015) Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, **201**, 1555–1579.
Gay L, Crochet PA, Bell DA, Lenormand T (2008) Comparing clines on molecular and phenotypic traits in hybrid zones: a window on tension zone models. *Evolution*, **62**, 2789–2806.
Gompert Z, Buerkle CA (2012) bgc: Software for Bayesian estimation of genomic clines. *Molecular Ecology Resources*, **12**, 1168–1176.
Gompert Z, Comeault AA, Farkas TE *et al.* (2014) Experimental evidence for ecological selection on genome variation in the wild. *Ecology Letters*, **17**, 369–379.
Grant V (1981) *Plant Speciation*. Columbia University Press, New York, NY.
Grant V (1993a) Effect of hybridization and selection on floral isolation. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 990–993.
Grant V (1993b) Origin of floral isolation between ornithophilous and sphingophilous plant species. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 7729–7733.
Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205–220.
Haldane JBS (1948) The theory of a cline. *Journal of Genetics*, **48**, 277–284.
Handelman C, Kohn JR (2014) Hummingbird color preference within a natural hybrid population of *Mimulus aurantiacus* (Phrymaceae). *Plant Species Biology*, **29**, 65–72.
Hare JD (2002) Geographic and genetic variation in the leaf surface resin components of Mimulus aurantiacus from southern California. *Biochemical Systematics and Ecology*, **30**, 281–296.
Harr B (2006) Genomic islands of differentiation between house mouse subspecies. *Genome Research*, **16**, 730–737.
Hebert FO, Renaut S, Bernatchez L (2013) Targeted sequence capture and resequencing implies a predominant role of regulatory regions in the divergence of a sympatric lake whitefish species pair (*Coregonus clupeaformis*). *Molecular Ecology*, **22**, 4896–4914.
Hermisson J, Pennings PS (2005) Soft sweeps molecular population genetics of adaptation from standing genetic variation. *Genetics*, **169**, 2335–2352.
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.

Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **367**, 395–408.

Huxley JS (1939) Clines: an auxiliary method in taxonomy. *Bijdragen tot de Dierkunde*, **27**, 491–520.

Jiggins CD, Mallet J (2000) Bimodal hybrid zones and speciation. *Trends in Ecology & Evolution*, **15**, 250–255.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Joron M, Frezal L, Jones RT *et al.* (2011) Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, **477**, 203–206.

Kruuk LEB, Baird SJE, Gale KS, Barton NH (1999) A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. *Genetics*, **153**, 1959–1971.

Lafontaine G, Prunier J, Gérardi S, Bousquet J (2015) Tracking the progression of speciation: variable patterns of introgression across the genome provide insights on the species delimitation between progenitor–derivative spruces (*Picea mariana*× *P. rubens*). *Molecular Ecology*, **24**, 5229–5247.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Larson EL, White TA, Ross CL, Harrison RG (2014) Gene flow and the maintenance of species boundaries. *Molecular Ecology*, **23**, 1668–1678.

Lowry DB, Willis JH (2010) A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biology*, **8**, e1000500.

Mallet J (2008) Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **363**, 2971–2986.

Mallet J, Barton N (1989) Inference from clines stabilized by frequency-dependent selection. *Genetics*, **122**, 967–976.

Martin SH, Dasmahapatra KK, Nadeau NJ *et al.* (2013) Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research*, **23**, 1817–1828.

Murovec J, Bohanek B (2013) Haploid induction in *Mimulus aurantiacus* Curtis obtained by pollination with gamma irradiated pollen. *Scientia Horticulturae*, **162**, 218–225.

Nosil P (2012) *Ecological Speciation*. Oxford University Press, Oxford.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Poelstra JW, Vijay N, Bossu CM *et al.* (2014) The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, **344**, 1410–1414.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, **20**, 208–215.

Ralph P, Coop G (2010) Parallel adaptation: one or many waves of advance of an advantageous allele? *Genetics*, **186**, 647–668.

Renaut S, Grassa CJ, Yeaman S *et al.* (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, **4**, 1827.

Roesti M, Kueng B, Moser D, Berner D (2015) The genomics of ecological vicariance in threespine stickleback fish. *Nature Communications*, **6**, 1–13.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters*, **8**, 336–352.

Schluter D (2009) Evidence for ecological speciation and its alternative. *Science*, **323**, 737–741.

Seehausen O, Butlin RK, Keller I *et al.* (2014) Genomics and the origin of species. *Nature Reviews Genetics*, **15**, 176–192.

Shin JH, Blay S, McNeney B, Graham J (2006) LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *Journal of Statistical Software*, **16**, 1–10.

Sobel JM, Streisfeld MA (2015) Strong premating isolation exclusively drives insipient speciation in *Mimulus aurantiacus*. *Evolution*, **69**, 447–461.

Sobel JM, Chen GF, Watt LR, Schemske DW (2010) The biology of speciation. *Evolution*, **64**, 295–315.

Soria-Carrasco V, Gompert Z, Comeault AA *et al.* (2014) Stick insect genomes reveal natural selection's role in parallel speciation. *Science*, **344**, 738–742.

Stankowski S (2013) Ecological speciation in an island snail: evidence for the parallel evidence of a novel ecotype and maintenance by ecologically dependent postzygotic isolation. *Molecular Ecology*, **22**, 2726–2741.

Stankowski S, Streisfeld MA (2015) Introgressive hybridization facilitates adaptive divergence in a recent radiation of monkeyflowers. *Proceedings of the Royal Society B: Biological Sciences*, **282**, 20151666.

Stankowski S, Sobel JM, Streisfeld MA (2015) The geography of divergence with gene flow facilitates multitrait adaptation and the evolution of pollinator isolation in Mimulus aurantiacus. *Evolution*, **69**, 3054–3068.

Streisfeld MA, Kohn JR (2005) Contrasting patterns of floral and molecular variation across a cline in *Mimulus aurantiacus*. *Evolution*, **59**, 2548–2559.

Streisfeld MA, Kohn JR (2007) Environment and pollinator-mediated selection on parapatric floral races of Mimulus aurantiacus. *Journal of Evolutionary Biology*, **20**, 122–132.

Streisfeld MA, Young WN, Sobel JM (2013) Divergent selection drives genetic differentiation in an R2R3-Myb transcription factor that contributes to incipient speciation in *Mimulus aurantiacus*. *PLoS Genetics*, **9**, e1003385.

Szymura JM, Barton NH (1986) Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegata*, near Cracow in southern Poland. *Evolution*, **40**, 141–1159.

Szymura JM, Barton NH (1991) The genetic structure of the hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata*: comparisons between transects and between loci. *Evolution*, **45**, 237–261.

Tulig M (2000). Morphological variation in Mimulus section *Diplacus* (Scrophulariaceae). Doctoral dissertation, California State Polytechnic University.

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, **3**, 572.

Twyford AD, Friedman J (2015) Adaptive divergence in the monkeyflower *Mimulus guttatus* is maintained by a chromosomal inversion. *Evolution*, **000**, 000–000.

Waayers GM (1996) Hybridization, introgression, and selection in *Mimulus aurantiacus* ssp. *australis* and *Mimulus puniceus*. Doctoral dissertation, San Diego State University.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

S.S., J.M.S. and M.A.S. conceived and designed the study and collected the data. S.S and M.A.S. analyzed the data. S.S. prepared the manuscript and figures with input from M.A.S. and J.M.S.

Data accessibility

Sequence data for the RAD-seq samples are archived at the Short Read Archive (SRA) under bioproject IDs PRJNA299226 and PRJNA317601. Sequence data used for the draft assembly are archived at the SRA under bioproject ID PRJNA317499. The draft assembly has been archived at DRYAD: http://dx.doi.org/10.5061/dryad.7j3rq.

Supporting information

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

Data S1. Supplementary methods.

Fig. S1. Pairs of SNP markers associated with the same RAD cut site show very similar levels of differentiation.

Fig. S2. Markers with different levels of allele sharing show contrasting cline shapes.

Fig. S3. Pairs of SNP markers on the same genomic scaffold have similar cline shapes.

Table S1. Geographic coordinates, *MaMyb2*-M3 allele frequencies, and RADseq sample sizes for the sample sites used in this study.