MOLECULAR ECOLOGY

Molecular Ecology (2016) doi: 10.1111/mec.13922

Population structure and local selection yield high genomic variation in *Mimulus guttatus*

JOSHUA R. PUZEY,*† JOHN H. WILLIS† and JOHN K. KELLY‡

*Department of Biology, College of William and Mary, Williamsburg, VA 23187, USA, †Department of Biology, Duke University, Durham, NC 27708, USA, ‡Department of Ecology and Evolution, University of Kansas, Lawrence, KS 27708, USA

Abstract

Across western North America, *Mimulus guttatus* exists as many local populations adapted to site-specific environmental challenges. Gene flow between locally adapted populations will affect genetic diversity both within demes and across the larger metapopulation. Here, we analyse 34 whole-genome sequences from the intensively studied Iron Mountain population (IM) in conjunction with sequences from 22 *Mimulus* individuals sampled from across western North America. Three striking features of these data address hypotheses about migration and selection in a locally adapted population. First, we find very high levels of intrapopulation polymorphism (synonymous $\pi = 0.033$). Variation outside of genes is likely even higher but difficult to estimate because excessive divergence reduces the efficiency of read mapping. Second, IM exhibits a significantly positive genomewide average for Tajima's *D*. This indicates allele frequencies are typically more intermediate than expected from neutrality, opposite the pattern observed in many other species. Third, IM exhibits a distinctive haplotype structure with a genomewide excess of positive associations between rarer alleles at linked loci. This suggests an important effect of gene flow from other *Mimulus* populations, although a residual effect of population founding might also contribute. The combination of multiple analyses, including a novel tree-based analytic method, illustrates how the balance of local selection, limited dispersal and metapopulation dynamics manifests across the genome. The overall genomic pattern of sequence diversity suggests successful gene flow of divergent immigrant genotypes into IM. However, many loci show patterns indicative of local adaptation, particularly at SNPs associated with chromosomal inversions.

Keywords: evolution, inversions, local selection, migration, *Mimulus*, population genomics

Received 12 November 2015; revision received 30 September 2016; accepted 7 November 2016

Introduction

A fundamental question in evolutionary biology is how genetic variation is maintained in the face of selection. Here, we consider this question as it relates to a species with many locally adapted, but subdivided demes (Wright 1932; Slatkin 1987; Wade 2016). Considering a species as a metapopulation, as opposed to a homogenous unit, is necessary when local populations are genetically distinct. Local differentiation is more the rule than the exception in plant species. Govindaraju (1988) summarized studies of allozyme variation within and among populations of 102 plant species: On average, 26% of allelic variation [measured as F_{ST} or G_{st}; (Nei 1973)] is distributed among populations and this percentage is highly variable among species [see also (Hamrick & Godt 1996; Loveless & Hamrick 1984)]. High mean F_{ST} has been corroborated by more recent surveys (Nybom 2004; Duminil *et al.* 2009) expanded to include DNA-based markers and many additional plant species. The nature and extent of differentiation has critical implications not only for the process of evolution, but how we study that process. In molecular

Correspondence: John K. Kelly, Fax: 785-867-5309; E-mail: jkk@ku.edu

population genetics for example, evolutionary inferences are usually based on summary statistics calculated from samples of gene sequences (Watterson 1975; Tajima 1989; Charlesworth *et al.* 1997). In a metapopulation, the interpretation of these statistics depends *entirely* on the scale of sampling. Sampling of individuals at both local and geographical scale is essential to address questions about balance of evolutionary forces in a metapopulation (Ross-Ibarra *et al.* 2008; Lack *et al.* 2015; Roesti *et al.* 2015).

In this study, we combine whole-genome sequencing with a multilevel sampling approach to investigate the maintenance of variation in the wildflower *Mimulus guttatus*. We examine a single focal population in conjunction with genomic data from individuals across the entire species complex. We use these data in a series of analyses to test basic predictions of the evolutionary theory of migration–selection balance (Wright 1931; Charlesworth *et al.* 1997; Yeaman & Whitlock 2011). We first confirm that previous marker-based and single-gene studies of *M. guttatus*, which indicated both high differentiation among populations and high intrapopulation (local) variation, are fully supported by genome-wide data. Second, we test the prediction that genomic regions subject to local selection will be less permeable to incoming haplotypes. Loci under local selection should have lower intrapopulation nucleotide diversity (π), but also higher F_{ST} and elevated absolute divergence between populations (Dxy), if gene flow occurs in other regions of the genome (Lewontin & Krakauer 1973; Beaumont & Nichols 1996; Charlesworth *et al.* 1997; Cruickshank & Hahn 2014). Third, we test the prediction that loci under local selection should exhibit distinct patterns of linkage disequilibria from the rest of the genome (Strobeck 1983; Charlesworth 2006; Storz & Kelly 2008; Jacobs *et al.* 2016). Finally, theory predicts that population genetic signatures of selection should be most pronounced when recombination is reduced (Kaplan *et al.* 1989; Begun & Aquadro 1992; Charlesworth *et al.* 1993). We test this prediction by comparing patterns of polymorphism, interpopulation divergence and linkage disequilibrium (LD) between genomic regions that harbour recombination-suppressing chromosomal inversion with the remainder of the genome.

An interesting and underappreciated prediction of migration–selection balance concerns the genomewide pattern of LD. Local selection and limited dispersal will allow populations to become differentiated in allele frequencies. As a consequence, when successful gene flow does (occasionally) occur, it can introduce divergent haplotypes into a population. Alleles that are rare in the focal population (or previously absent) are introduced in combinations. In this way, migration can generate positive associations between (locally) rare alleles at linked single nucleotide polymorphisms (SNPs). To evaluate this prediction, we calculate 'polarized' linkage disequilibrium (D) measuring the association of the minor alleles (the less common base at each contrasted SNP pair). Positive D indicates that minor alleles are positively associated (Langley & Crow 1974). To measure this signal, we compare the observed distribution of polarized D estimates between nearby SNPs with the predicted distribution under mutation–recombination–drift balance.

Our focal population for these studies, Iron Mountain (IM), has been the subject of intense evolutionary and ecological research for the past 30 years (Willis 1993, 1996; Scoville *et al.* 2011; Flagel *et al.* 2014; Fishman & Kelly 2015). In IM, plant lifespan is strictly limited by water availability. During the short window between the spring snow melt and summer drought (routinely 6–10 weeks), seedlings must grow, flower, mate and set seed. These abiotic pressures impose strong selection (Willis 1996; Mojica & Kelly 2010), and the population exhibits adaption to local conditions (Hall & Willis 2006). Despite this, IM retains high internal variability in both molecular and quantitative genetic traits (Kelly & Willis 1998; Kelly & Arathi 2003). Chromosomal inversions segregate both within IM (Scoville *et al.* 2009; Lee *et al.* 2016) and also between IM and other populations (Lowry & Willis 2010; Holeski *et al.* 2014; Twyford & Friedman 2015).

Iron Mountain is annual population within the *M. guttatus* species complex, an enormous collection of localized populations. Some of these populations are recognized as distinct taxa (e.g. *Mimulus nasutus*) (Fenster & Ritland 1994; Fishman *et al.* 2002) or ecotypes (e.g. annual and perennial ecotypes of *M. guttatus*) (Lowry & Willis 2010). The *M. guttatus* complex occurs across western North America and has adapted to a wide range of habitats including serpentine barrens, heavy-metal-rich mine tailings, huge elevation ranges and oceanic salt spray (Kelly 2003; Hall & Willis 2006; Lowry *et al.* 2009; Mojica *et al.* 2012). Populations within the complex are often interfertile to varying degrees, and gene flow occurs between populations, including those named as distinct species (Brandvain *et al.* 2014). However, potentially, strong selection against immigrant genotypes allows substantial genetic differentiation (Hall & Willis 2006; Hall *et al.* 2010; Friedman *et al.* 2015; Kooyers *et al.* 2015). The level of genetic differentiation increases with distance among populations in *M. guttatus*. Lowry & Willis (2010) estimated $F_{ST} = 0.48$ across a set of 30 populations spanning a latitudinal range from 35° to 45°. Similarly, Twyford & Friedman (2015) obtained an F_{ST} of 0.46 for populations sampled across a large extent of *M. guttatus*' native range (latitudinal range: 31.2–53.8). F_{ST} is slightly lower (0.43) in a

more geographically limited sampling of *M. guttatus* populations across Oregon, including IM (V. Koelling and J. K. Kelly, unpublished results from whole-genome sequencing study). If IM is compared to populations at much smaller distances (3 and 6 km, respectively), F_{ST} declines to 0.07 and 0.13, respectively (Monnahan *et al.* 2015). The geographical genetic data indicate limits on gene flow. Local adaptation should elevate differentiation relative to the level predicted by migration/drift balance.

In this study, we analyse genome sequences from 34 IM individuals in conjunction with 22 individuals from other populations across the species complex. We use a combination of techniques in two stages of analysis to test the migration–selection balance predictions outlined above. First, we characterize patterns of polymorphism and linkage disequilibria within IM. These analyses yield several striking results including a remarkably high level of synonymous site diversity, a tilt of the site frequency spectrum towards intermediate allele frequencies, and a distinctive haplotype structure in which minor alleles are positively associated at linked sites. In the second phase of the analysis, we compare IM sequences to individuals from allopatric populations within the species complex, estimating divergence relative to polymorphism. By reconstructing distance-based trees for thousands of intervals across the entire genome, we evaluate patterns of relatedness and the possibility of gene flow into IM. The varying structure of polymorphism within IM relative to divergence from other *M. guttatus* populations indicates an overall pattern of successful gene flow of divergent immigrant genotypes into IM. However, many loci show patterns indicative of local adaptation, particularly at SNPs associated with chromosomal inversions.

Materials and methods

Focal population and plant samples

All sequences in this study (IM and allopatric) are based on samples from natural populations. IM is located in the cascade mountains of central Oregon (44.402217N, −122.153317W) at an elevation of approximately 1400 m. The population is predominantly outcrossing (Willis 1993), but plants are self-compatible. IM exhibits minimal internal spatial structure: The genetic relatedness of neighbouring plants, typically separated by <1 cm, is essentially zero at microsatellite loci (Sweigart *et al.* 1999). We propagated approximately 1200 independent lines of *Mimulus guttatus* by single-seed descent (self-fertilization) for 5–13 generations. Each line was founded from the seed set of a separate field-collected plant sampled from the IM (Willis 1999a). As expected, the inbred lines are almost completely homozygous at microsatellite loci with different lines fixed for different alleles (Kelly 2003). Some novel mutations may have been introduced over the course of line formation, but the number of such mutations should be miniscule relative to standing variation (see 'Results'). It is likely that recessive alleles causing lethality or sterility (under greenhouse conditions) were lost during line formation (Willis 1999b). DNA from 39 of these lines was newly extracted and sequenced.

After eliminating some lines as redundant (see below), we combined these data with nine previously sequenced IM lines (Flagel *et al.* 2014) and 21 allopatric samples (individuals from other populations or species in the complex (Table S1, Supporting information); reads downloaded from the JGI Short Read Archive). For all samples (both IM and allopatric individuals), average read depth after filtering (as determined from VCF file using vcftools –depth) ranged from 2.6 to 24.4 (mean = 6.7; calculated for genotyped bases on chromosome 1, no indels included, Table S1, Supporting information). We also newly sequenced an additional allopatric individual, hereafter called Iron Mtn. Perennial (IMP). This perennial population occurs in close proximity to the annual IM population. IMP might be taxonomically classified as *Mimulus decorus* on the basis of its distinctive morphology, including numerous long and thin underground stems (stolons), although the results of this study suggest substantial genetic similarity to IM. Like the IM samples, the sequenced allopatric individuals are inbred lines; descendants of wild plants propagated through multiple generations of self-fertilization in the greenhouse.

DNA extraction, library preparation and sequencing

We collected and froze leaf tissue and extracted DNA using the Epicentre Leaf MasterPure kit (Epicentre, USA). Libraries for Illumina sequencing were made using the Illumina Nextera DNA kit (Illumina, USA). Individual barcodes were added during library preparation to facilitate multiplexing. Libraries were pooled in equal molar amounts based on concentrations measured using the Qubit high-sensitivity DNA assay and insert size distributions obtained from a Agilent bioanalyser (HS-DNA chip, Agilent Technologies, USA). Up to 24 libraries were pooled in a single Illumina HiSeq 2500 Rapid-Run sequencing run generating 150-bp paired-end reads.

Alignment, genotype calling and residual heterozygosity

After sequencing, we demultiplexed reads into individual samples and mapped them independently. Reads

were aligned to the unmasked *M. guttatus* v2.0 reference genome (http://www.phytozome.net/) (assembled length of 321 Mb) using BOWTIE2 (Langmead & Salzberg 2012). Next, we converted SAM alignment files to binary format using SAMTOOLS (Li *et al.* 2009) and then processed alignments with PICARD TOOLS (http://broadinstitute.github.io/picard; Commands: FixMates, MarkDuplicates, and AddReadGroups). The Picard processing validated read pairing, removed duplicate reads and added read groups for analysis in the Genome Analysis Toolkit (GATK) (DePristo *et al.* 2011). We called genotypes using GATK UnifiedGenotyper (details in Supporting information). Genotype VCF files were converted to tabbed format using VCFTOOLS (vcf-to-tab) (Danecek *et al.* 2011).

After the initial genotyping, we masked putative SNPs that were excessively heterozygous (in more than 25% of lines) as these are likely due to mismapped reads. Mismapping is likely in regions of the reference genome where paralogs are incorrectly collapsed into a single gene. We further suppressed entire genomic intervals where, across lines, the mean heterozygosity divided by the average expected heterozygosity (given by the Hardy–Weinberg proportions) exceeds 0.5. Finally, for each individual, we calculated the ratio of observed to expected heterozygosity within 500 SNP windows across the genome. Within each line, we called a region heterozygous if the average was elevated across 10 successive windows. We identified a total of 429 residually heterozygous regions across all lines/chromosomes. This corresponds to 1.29% of the sequence in total. The Mendelian prediction for residual heterozygosity with single-seed descent is 1.56% after six generations and 0.78% after seven generations. The size distribution of putative residual heterozygous regions is consistent with the number of generations of selfing, the size of the genome 450–500 mB and map length of about 125 cM per chromosome (Fig. S1, Supporting information).

Identification of related lines

After the genotype filtering described above, we constructed a similarity matrix for all IM lines using the Emboss fdnadist program with the Jukes–Cantor substitution matrix. A total of 4.1 million SNPs were called in 43 or more IM lines. Based on this approach, we identified lines that were excessively similar (Fig. S2, Supporting information). A distribution of pairwise similarity between IM lines shows clear outliers (Fig. S2A, Supporting information). For instance, IM777 is 0.997 similar to IM323. We thus determined these lines to be relatives and eliminated the IM323 from subsequent analyses. For each pair of IM lines with >0.98 similarity, one line was removed from analysis. After filtering relatives, the following 34 lines were included for all subsequent tests: 62, 106, 109, 115, 116, 138, 170, 179, 238, 239, 266, 275, 359, 412, 479, 502, 549, 624, 657, 667, 693, 709, 742, 767, 777, 785, 835, 886, 909, 922, 1054, 1145, 1152, 1192 (Fig. S3, Supporting information).

Relationship of missing data and divergence

We delineated windows containing 500 SNPs and, within each window of each line, calculated (i) the number of called and uncalled sites and (ii) the number of SNPs called for the reference allele as opposed to the alternative allele. The fraction of missing data was calculated from (i), and the window divergence (fraction of calls to alternate) from (ii). We performed a logistic regression in R with fraction missing as the response and divergence as the predictor: glm(formula = logit1*frac.missing*~*logit*1divergence, family = binomial). This revealed a strong relationship between missing data and divergence (fraction of called SNPs that differ from the reference genome; Figure S4A, Supporting information). Given this relationship, we opted to focus our analyses where data were most complete using two complementary approaches. First, based on the fact that the fraction of missing data is much lower in coding regions (Fig. S4B, Supporting information), we conducted a series of gene-based analyses (e.g. synonymous vs. nonsynonymous diversity). The mean fraction of called bases in coding regions, calculated for each line, ranged from 0.63 to 0.86 (Table S2, Supporting information). Second, we identified genomic windows (genic and intergenic DNA) each consisting of 10 000 genotyped bases (monomorphic and polymorphic sites both count as genotyped bases). To qualify as a genotyped base, a site had to be scored in at least 30 of the 34 unrelated IM lines. The resulting windows ranged from 10 000 to 3 427 432 bases (of the reference genome) with a mean and median of 39 044 and 18 478 bases, respectively. Allowing 1000 genotyped base overlapping steps between windows, a total of 74 445 windows span the 14 chromosomes. For these windows, we calculated population genetic and tree-based statistics.

Nucleotide diversity within genes (synonymous and nonsynonymous π)

We converted filtered genotype files to fasta format for the entire genome for each separate line. When re-creating line-specific fasta files, missing data were not imputed, and indels and heterozygous sites were suppressed. We extracted coding sequences using GFFREAD (Langmead & Salzberg 2012). Each gene was individually extracted from the line-specific coding sequences

libraries and combined into a single fasta file containing 34 individual coding sequences for each gene. We calculated synonymous and nonsynonymous diversity through pairwise comparisons of all lines using the KaKs_Calculator (Nei and Gojobori model) (Zhang *et al.* 2006). Only diversity measurements derived from genes with alignment lengths >1000 bases were included (n = 29 421). A $\pi_{\text{non-syn}}$ and π_{syn} value was computed for each gene and was used to calculate genomewide mean Ka and Ks.

Window analyses

We calculated statistics of polymorphism, divergence and genealogy within windows of 10 000 genotyped bases. We then created a phylogenetic tree for every window using EMBOSS fdnadist (Rice *et al.* 2000) to calculate a nucleotide distance matrix (Jukes–Cantor substitution model). Trees were inferred from the distance matrix using EMBOSS fneighbor (Rice *et al.* 2000) and rooted using *Mimulus dentilobius*. Of the 74 445 windows, fneighbor failed to parse the distance matrix for only 28 windows, which we excluded from subsequent analysis. Next, for each tree, we determined whether IM formed a monophyletic clade or was polyphyletic using a custom perl script (Vos 2015) dependent on the Bio::Phylo toolkit (Talevich *et al.* 2012). In cases that IM was polyphyletic, we determined how many allopatric samples had to be removed to restore IM monophyly using the perl monophyletic output (Vos 2015) and custom perl scripts.

We calculated S (the number of polymorphisms), π (nucleotide diversity), Tajima's D (Tajima 1989) and LD statistics in each window using custom python scripts. For the linkage disequilibrium (D), we estimated the association of the minor alleles (less common base) at each contrasted SNP pair. Positive D indicates that minor alleles are positively associated (Langley & Crow 1974). We also standardize D as the correlation coefficient, $r = D/\sqrt{p(1-p)q(1-q)}$ and from that, calculate r^2 [the Z_{nS} test for selection is r^2 conditioned on S (Kelly 1997)]. We calculated r and r^2 for SNP pairs across each chromosome to estimate the long-range pattern of LD. For comparison with observed LD, we performed neutral simulations using calibrated, empirical estimates for 4Nu (from nucleotide diversity) and 4Nr (from LDhelmet as described below) by updating the programs used in Storz *et al.* (2012). Absolute nucleotide divergence, Dxy, between IM annuals and all allopatric individuals were calculated using a perl script (LaMariposa) dependent on BioPerl::PopGen modules (Dxy is equivalent to π_{XY} (Nei & Li 1979)]. Dxy was calculated on a single base increment for all sites that had at least one IM and one allopatric individual genotyped. Next, using these values, an average Dxy value was calculated for the same 10 000 genotyped base windows used for other population genetic statistics.

Recombination rates within IM

We used LDhelmet (Chan *et al.* 2012) to estimate fine-scale recombination rates with recalled genomes in fasta format as inputs. First, using the 'find_confs' command, 50 SNP windows were used to scan the genome and create a haplotype configuration file. Next, a likelihood lookup table and Pade coefficients were generated using a population-scaled mutation rate of 0.015 (this was based on a preliminary estimate for genomewide π within IM). In the final step, the 'rjmcmc' command was run using the previously generated haplotype configuration, likelihood table, and Pade coefficients and a Jukes–Cantor mutation matrix to estimate recombination rates. Exon-specific recombination rates were calculated. Only pairs of SNPs contained within exons were used. The genomewide mean estimate for 4Nr was used to calibrate the neutral simulator described above.

Specieswide statistics

As a contrast to the population genetic estimates within IM, we constructed a sample with a single IM line (IM767) and 21 of the allopatric individuals (excluding *M. dentilobius* as it is outside the *M. guttatus* complex). For this specieswide sample, we calculated nucleotide diversity within genes and other statistics (S, Tajima's D, LD) genomewide, requiring that SNP be called in at least 15 of 21 lines for inclusion.

Results

Nucleotide diversity within IM and divergence from allopatric populations

Within genes, mean synonymous nucleotide diversity (π_{syn} = 0.033) is fivefold greater than mean nonsynonymous diversity ($\pi_{\text{non-syn}}$ = 0.006) within IM (Fig. S5, Supporting information). Nucleotide diversity (genic and nongenic) varied across windows with a mean value of π_{Genome} = 0.014. The variance in pairwise π among the 561 contrasts between 34 IM lines within each genomic window also exhibits many localized peaks across the genome (Fig. S8, Supporting information). The mean of Var[π] is 0.0000831, and this statistic is positively correlated with nucleotide diversity in the window and with LD measured as r or r^2 (Figs S6 and S8, S9, Supporting information).

Diversity within IM was lower than divergence of IM sequences from allopatric individuals (Dxy): mean

Dxy = 0.038, range 0.002–0.166 (Fig. S6E, Supporting information) with several clear peaks of high Dxy (Fig. 2). Dxy is the IM vs. allopatric sequences analog of π_{Genome} (IM vs. IM), and thus, the estimated ratio of divergence to polymorphism is about 2.7.

We delineated each of the three inversions mapped in the IMxPR RIL population (Holeski *et al.* 2014) by locating genetic markers used in the cross to locations in the v2 genome build (bars in Fig. 2). This is a cross between annual (IM) and perennial (PR) genotypes. We did not analyse variation within the chromosome 6 inversion that segregates within IM because it has been largely purged from the sequenced lines (Lee *et al.* 2016). Sequence divergence (Dxy) is significantly elevated within the inversion regions on chromosomes 5, 8 and 10 relative to genomewide averages: $Dxy_{Genome} = 0.037$, $Dxy_{inversion(8)} = 0.044$ ($P < 0.0001$), $Dxy_{inversion(5)} = 0.068$ ($P < 0.0001$) and $Dxy_{inversion(10)} = 0.042$ ($P < 0.0001$). The chromosome 8 and chromosome 10 inversions show significantly lower overall nucleotide diversity within IM while the chromosome 5 is not statistically different from genomewide levels: $\pi_{Genome} = 0.014$, $\pi_{inversion(10)} = 0.010$ ($P < 0.0001$), $\pi_{inversion(8)} = 0.012$ ($P < 0.0001$) and $\pi_{inversion(5)} = 0.014$ ($P = 0.66$). Interestingly, Var[π] is statistically elevated in the chromosome 5 inversion but statistically lower in the regions on chromosomes 8 and 10: $Var[\pi]_{Genome} = 0.000085$, $Var[\pi]_{inversion(5)} = 0.000109$ ($P = 0.0002$), $Var[\pi]_{inversion(10)} = 0.000060$ ($P < 0.0001$), and $Var[\pi]_{inversion(8)} = 0.000053$ ($P < 0.0001$).

Linkage disequilibrium and recombination rate

Chromosomal segments that have recently entered the IM population by migration from other differentiated demes are likely to consist of SNP alleles that are rare within IM. Such migration should generate patterns of positive association between rare alleles at linked SNPs. To investigate this possibility, we estimated LD using a consistent coding for minor alleles at each SNP. This 'polarized' measure of LD indicates a striking excess of positive associations (Fig. 1) relative to the level predicted with neutrality. If an IM line harbours the less frequent base at a SNP, it is much more likely to have the less frequent base at neighbouring SNPs. The neutral distribution of Fig. 1 was obtained using the average, genomewide ρ (4Nr) of 0.0042 obtained from LDhelmet (Chan *et al.* 2012). When measured as r^2, LD is high at short distances (~100 bp) and shows a rapid decay with sequence distance (Fig. S10, Supporting information). It should be noted that the pattern of long-range LD differs among chromosomes (Fig. S10, Supporting information). The average genic ρ is 0.0052 and recombination hotspots are clearly evident (Fig. S11, Supporting information).

Distribution of monophyletic IM clusters across the genome

To compare the genomic patterns of diversity in IM with divergence among populations, we constructed a phylogenetic tree for every window of 10 000 genotyped bases in the genome including both IM and allopatric samples. The monophyly of IM samples was evaluated individually for each tree. A minority (10 504) phylogenetic trees were monophyletic for IM while the majority (64 913) showed IM as polyphyletic; in those genomic regions, some IM lines were more similar to allopatric sequences than to other IM lines (Table S4, Supporting information). Genomewide, monophyletic windows were found both in gene-dense and gene-sparse regions. For each polyphyletic IM window, we calculated the number of allopatric samples that would have to be removed from the tree for IM to be monophyletic. For 6958 of the polyphyletic

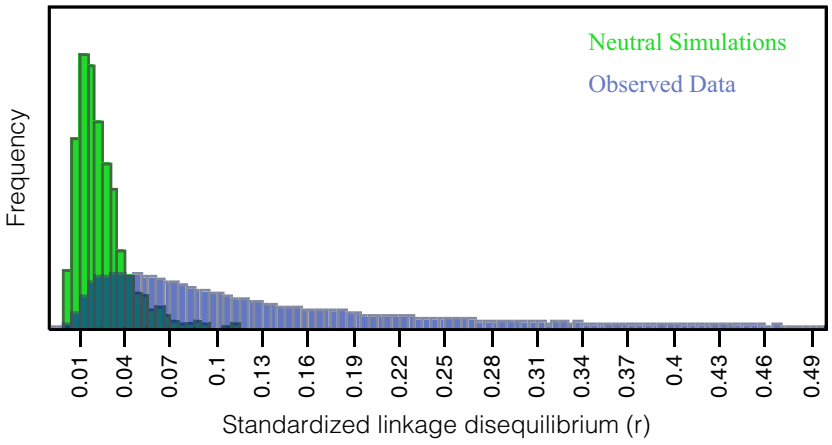

Fig. 1 Genomewide distribution of *r* in IM annuals shows an excess of positive associations between rare alleles relative to neutral simulations. IM, Iron Mountain. [Colour figure can be viewed at wileyonlinelibrary.com]

trees (11%), only one allopatric sequence would have to be removed to restore IM monophyly (Table S5, Supporting information). For 48% of these 6958 trees, the geographically proximate Iron Mtn. Perennial (IMP) was responsible for IM polyphyly (Table S5, Supporting information). This pattern is evident across all linkage groups, but the incidence is highest for chromosome 8 where 65% of polyphyletic trees where IM monophyly is broken by a single individual are due to IMP (Table S5, Supporting information). Table 1 summarizes relationship between population genetic statistics, IM mono/polyphyly and the location of structural polymorphisms. Nucleotide diversity (π), Var[π], polarized LD (r), the number of segregating sites and Tajima's D were all significantly elevated in polyphyletic windows when compared with monophyletic windows (Table 1). Dxy was significantly lower in polyphyletic windows relative to monophyletic windows (Table 1).

Relationship of Var[π] and tree topology

Combining Var[π] and tree topology allows for identification of genomic regions that have experienced a selective sweep or introgression event (Fig. 3 and Fig. S13, Supporting information). On average, we expect monophyletic regions of the genome to have lower Var[π], a trend we observe (Table 1). First, we selected the lowest Var[π] window and the topology suggests a selective sweep – short branches within a monophyletic IM clade (IM individuals within this block possess nearly the exact same haplotype; Fig. 3A). Next, we picked the highest Var[π] region for all monophyletic windows (Fig. 3B). Here, the IM population contains several highly diverged sequences. Introgression may well have been the original source of the divergent lineages in Fig. 3B, although locally they may be maintained by selective processes within IM. It is also possible to identify the genomic effects of introgression from individual allopatric sequences. To illustrate this, we extracted all regions of the genome where IMP is solely responsible for breaking IM monophyly and extracted high Var[π] regions. Tree and π distributions from these windows show evidence of introgression and segregation of very divergent haplotypes (long branches within IM and multimodal π distribution; Fig. 3C, Fig. S13, Supporting information). Two key features of the tree-inference analysis are indicated by a specific (but typical) interval on chromosome 5 (Fig. 4, Supporting information). Most windows are paraphyletic, but the number and identity of allopatric individuals that break monophyly change along a chromosome. Second, genomic windows where IM sequences are monophyletic are distinctly clustered, such as around the 1.875 mb and 2.175 mb locations on chromosome 5 (Fig. 4, Supporting information).

Statistics in the specieswide sample

Across the species complex, nucleotide diversity within genes is very high: $\pi_{syn} = 0.063$ (SE = 0.0001), $\pi_{non\text{-}syn} = 0.0099$ (SE = 0.00005). Across genotyped bases (genic and nongenic), there are approximately twice as many SNPs in the specieswide sample than in the IM sample (S = 11 715 727 vs. 5 676 399). However, differences in the missing data pattern between these samples impede comparisons based on the windows used for IM only. Thus, we calculated Tajima's D and LD for each in consecutive nonoverlapping 50 SNP windows. For the 113 519 windows of IM sample, the mean for Tajima's D was 0.159 (SE = 0.003; SD = 1.082) and the mean for r was 0.245 (SE = 0.001, SD = 0.194). For the 234 308 windows of the specieswide sample, the mean Tajima's D was −0.906 (SE = 0.001, SD = 0.472) and the mean for r was 0.036 (SE = 0.0001, SD = 0.043).

Table 1 Contrast of population genetic statistics between monophyletic and polyphyletic windows. With these groups, we classified windows as within one of the mapped inversions or elsewhere in genome. n = number of windows. The difference between monophyletic and polyphyletic windows is highly significant for all statistics

Tree status	Genomic location	N	π	Tajima's D	Var[π]	r	Dxy
IM monophyletic	Inversion 5	35	0.0139	−0.3277	3.59E-05	0.0782	0.0907
	Inversion 8	1455	0.0117	0.1091	3.62E-05	0.0553	0.0465
	Inversion 10	700	0.0091	0.2108	3.33E-05	0.1082	0.0458
	Rest of genome	8314	0.0119	0.0793	4.81E-05	0.0871	0.0499
IM polyphyletic	Inversion 5	42	0.0136	−0.2747	17.06E-05	0.2769	0.0482
	Inversion 8	925	0.0135	0.2818	7.99E-05	0.1213	0.0399
	Inversion 10	630	0.0110	0.2820	8.93E-05	0.2474	0.0370
	Rest of genome	62 316	0.0143	0.1498	8.94E-05	0.1339	0.0356

IM, Iron Mountain.

Discussion

Traditionally, sequencing efforts in evolutionary genomics have focused on sampling a single individual from each of multiple populations distributed across the full range of a species. Only recently have evolutionary biologists begun generating whole-genome sequence data sets specific to demes, populations of individuals connected by mating in recent time, for example (Mackay *et al.* 2012; Burri *et al.* 2015; Kubota *et al.* 2015). This study uses a combination of intensive within- and across-population whole-genome sequencing to test hypotheses about the balance between evolutionary forces acting in a metapopulation. Within the IM population, we observe very high sequence diversity and atypically intermediate allele frequencies. Our data support the prediction that genomic regions subject to local selection are less permeable to introgressing haplotypes introduced by immigrant pollen or seed. Estimates for polarized linkage disequilibria (LD) provide evidence that migration does introduce alleles into IM, at least within genomic regions where gene flow is not impeded by local selection. Finally, the data provide further support to the growing body of examples that chromosomal inversions are an important component of local adaptation (Krimbas & Powell 1992; Gilburn & Day 1999; Coluzzi *et al.* 2002; Balanyà *et al.* 2003; Feder *et al.* 2003; Hoffmann & Rieseberg 2008; Lowry & Willis 2010; Cheng *et al.* 2012; Fang *et al.* 2012; Jones *et al.* 2012). Owing to recombination suppression, these inversions can have pronounced effects on gene sequence evolution (Fig. 2).

Diversity within IM

Variation is extremely high for a single population ($\pi_{syn} = 0.033$, $\pi_{non\text{-}syn} = 0.006$, $\pi_{Genome} = 0.014$), consistent with a previous study of several autosomal genes within *M. guttatus* [$\pi_{syn} = 0.061$ in Puzey & Vallejo-Marín (2014)]. The estimate for nonsynonymous diversity is much lower, reiterating the usual pattern of purifying selection on amino acid changes that is observed in most species. Our genomic estimate ($\pi_{Genome} = 0.014$), which includes both genic and noncoding sequence, is almost certainly an underestimate. There is a strong tendency for missing data to increase with divergence from the reference genome (Fig. S4, Supporting information). The consequence is a downward bias in π due to ascertainment; we are less likely to map (and thus analyse) sequences that are most divergent. This is not surprising but, to our knowledge, has not been demonstrated previously. We expect this to be a general phenomenon extending across most studies of this kind. High levels of insertion/deletion variation (Flagel *et al.* 2014) is likely contributing to incomplete read mapping and subsequent underestimation of nucleotide diversity.

The high diversity within IM is notable for a single local population. Leffler *et al.* (2012) recently summarized nucleotide diversity across a wide range of species, classifying estimates by sampling strategy (one population or multiple populations), site type (synonymous, nonsynonymous, etc.) and chromosome type (autosome or sex). Considering single population ($n = 9$, species from five phyla: Arthropoda, Chlorophyta, Chordata, Mollusca and Pinophyta), the mean and median of autosomal π_{syn} were 0.014 and 0.011, respectively (range 0.001–0.033). Across multiple populations ($n = 50$), the mean and median were 0.010 and 0.006, respectively. *Mimulus guttatus* thus represents one of the most variable species yet described, excepting the hyperdiverse nematode recently reported by Dey *et al.* (2013).

The specieswide F_{ST} of ~0.5 for *M. guttatus* (see 'Introduction') implies that only half of allelic variation resides within demes, on average. This high F_{ST} implies that migration, if successful, should introduce novel haplotypes into IM. This may help to explain the pattern of LD observed in IM, where rarer alleles at proximate SNP pairs are positively associated across the genome (Fig. 1). Most sequencing studies do not specify associations between SNPs in terms of features of alternative alleles, and as a consequence, the direction of LD estimates is meaningless. Indeed, direction is lost when calculating r^2, which is commonly used to measure the strength of association between SNPs (e.g. Fig. S10, Supporting information). Here, we polarize bases (minor vs. major) based on allele frequency so that the absolute value of LD can inform questions about evolutionary process. Langley and Crow (1974) developed an epistatic selection model to explain the negative LD observed in allozyme data. In IM, LD is highly variable in both direction and magnitude. However, we suggest that the positive mean for LD is due, at least in part, to migration. Immigrants from divergent populations tend to generate positive LD by introducing novel alleles in combinations.

The third striking feature of intra-IM variation, the positive average value for Tajima's *D* (Table 1), also requires a careful consideration of both genomic and spatial scale. At the scale of individual loci, this statistic is routinely used as a test for selection. Negative Tajima's *D* (extreme allele frequencies) can result from strong purifying selection or a recent selective sweep, while positive values (intermediate allele frequencies) suggest balancing selection (Tajima 1989). Different genomic windows of our survey illustrate each of these outcomes. The window on chromosome 8 at position

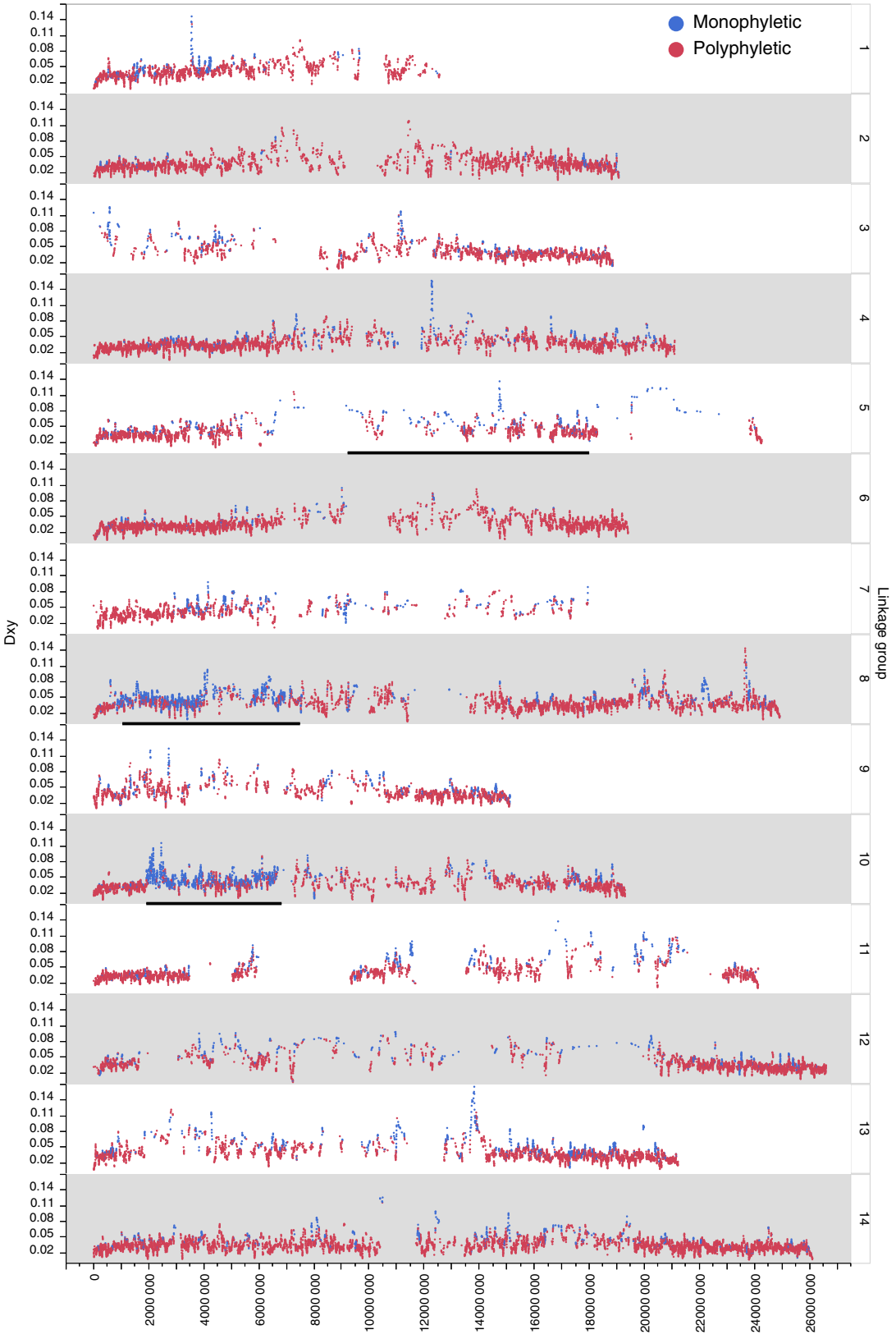

Fig. 2 Genomic distribution of absolute divergence (Dxy). Blue dots = monophyletic windows; red dots = polyphyletic windows. Bars on chromosomes 5, 8 and 10 denote approximate locations of chromosomal inversions. [Colour figure can be viewed at wileyonlinelibrary.com]

8.25 Mb exhibits the topology predicted by a recent selective sweep within IM (Fig. S14, Supporting information) and Tajima's $D = -2.31$ across 191 SNPs. In contrast, the meiotic drive locus on chromosome 11 exhibits positive Tajima's D over a large genomic interval: Mean value = 0.46 across 561 genomic windows from position 9.5 to 11.7 Mb. Previous study of this region indicates a balanced polymorphism owing to meiotic drive. The drive allele maintained at a population frequency of approximately 35% owing to balancing benefits and costs (Fishman & Saunders 2008; Fishman & Kelly 2015).

Demographic arguments are typically invoked if the genomic average of Tajima's D is significantly nonzero: population expansion to explain negative values, and population contraction for positive values (Tajima 1989). However, these interpretations depend on the spatial scale of sampling. A species that is expanding through many local populations, each founded from a limited propagule, can exhibit positive Tajima's D within demes (Ross-Ibarra *et al.* 2008), even if the statistic is negative in a specieswide sample. To illustrate, consider a population founded by a single seed that rapidly expands to large population size. At the beginning, all SNPs in this population will be at loci that were heterozygous in the founder. Such SNPs will have initial population frequencies of 0.5 and thus produce the highest possible values for Tajima's D. The statistic will be reduced by subsequent drift, pushing allele frequencies away from 0.5, and novel mutations that necessarily start at low frequency ($1/2N$). However, simulations indicate a substantial time persistence of this founding effect (Ross-Ibarra *et al.* 2008) and local demes of many species are likely to be quite young on the 'coalescent' time scale.

The comparison of genetic diversity statistics between our IM sample and our specieswide sample further indicates the need to interpret population genetic statistics in a metapopulation context. Consistent with the high specieswide F_{ST}, the amount variation in our complexwide sample is about double that within IM in terms of the number of polymorphic sites. However, the nature variation with respect to allele frequencies is quite different. While the mean Tajima's D is positive within IM, it is very significantly negative (mean = −0.90) in the complexwide sample. This is similar to what has been observed in several *Drosophila* species (Fabian *et al.* 2012; Mackay *et al.* 2012; Nolte *et al.* 2013), as well as in Arabidopsis (Schmid *et al.* 2005). It is not surprising that lineages within the species complex have acquired lineage-specific mutations (adaptive, neutral or deleterious), and by definition, such mutation will be rare in the complexwide sample. In the next section, we describe results from our second phase of analysis where we compare IM sequences to 22 genomes from 'allopatric individuals' sampled from across the *M. guttatus* species complex.

Genomewide effects of migration and localized signatures of selection

Most genomic windows exhibit polyphyly of the IM samples, that is some IM lines are more similar to other populations than to other IM lines. The high frequency of polyphyly is not caused by a few divergent IM lines repeatedly breaking monophyly. Instead, most IM lines exhibit similarity to allopatric sequences within portions of their genomes, the identity of lines that break monophyly changing across the genome. This is expected given previous evidence that IM is an internally well-mixed, outbred population (Sweigart *et al.* 1999). IM polyphyly could be due to either ancestral polymorphism that is continuing to segregate or introgression from neighbouring populations. These are difficult to distinguish and both are likely important. However, several observations suggest low, but nontrivial, immigration of genotypes into IM. First, the previously noted positive LD pattern (Fig. 1) is significantly elevated in polyphyletic windows (Table 1). Second, the geographically proximate IMP is the most frequent cause of IM polyphyly (when a single allopatric is the cause; Table S5, Supporting information) consistent with migration from this source. Third, the number and identity of allopatric sequences breaking IM monophyly turn over rapidly as one moves along a chromosome (Fig. 4) suggesting a diverse set of contributors to IM.

If there is gene flow into IM, we expect its molecular signal to vary across the genome. In particular, introgression should be reduced at loci subject to local adaptation (Lewontin & Krakauer 1973; Beaumont & Nichols 1996). Despite the general tendency towards polyphyly, many genomic windows exhibit a distinct pattern suggesting local adaptation. The clearest signature of selection is a genomically localized reduction in nucleotide diversity coupled with increased divergence of IM from other populations of *M. guttatus* (Fig. 2) (Nosil *et al.* 2009). One of the most striking observations from Table 1 is the elevated Dxy, depressed π and Var[π] in monophyletic windows. Reduced π is a one signal of directional selection while elevated Dxy is indicative of locally beneficial variants.

Selection effects should be most pronounced when recombination is suppressed. Chromosomal inversions suppress recombination within heterozygotes over broad genomic scales, and a number of inversions have been identified in *M. guttatus* (Lowry & Willis 2010; Holeski *et al.* 2014; Twyford & Friedman 2015). Three of these, on chromosomes 5, 8 and 10, respectively, are

located approximately to the genome sequence in Fig. 2. Consistent with a migration–selection balance model for these loci (Guerrero *et al.* 2012), we find that sequence divergence (Dxy) is significantly elevated within all three inversion regions relative to genomewide observations (Fig. 2). The inversion regions on chromosomes 8 and 10 show significantly lower overall nucleotide diversity within IM, while the chromosome 5 inversion is on not statistically different from genomewide levels.

Direct evidence for local selection on the chromosome 8 inversion comes from experiments mapping QTL important for life-history traits and flowering time to this locus (Hall *et al.* 2006, 2010). Reciprocal transplant experiments have directly shown local selection on inversion 8 (Lowry & Willis 2010). In fact, these previous experiments predict the current findings of increased IM monophyly (61% of genomic windows are monophyletic in the inversion 8 region relative to 14% for the genome as a whole 14%), and increased Dxy within this region. This region also shows slightly lower nucleotide diversity (0.012) than the genome average (0.14), consistent with increased monophyly (Table 1). Also, while the general pattern appears to be resistance to introgression, there is evidence for gene exchange via recombination or gene conversion between IM and IMP within inversion 8. IMP is the sole cause of IM polyphyly far more frequently within inversion 8 than genomewide. IMP is responsible for 409 of 925 (44%) polyphyletic windows, while outside the inversion, IMP is solely responsible for only 2923 of 62 988 (4.6%) polyphyletic windows. This collection of estimates is interesting given that inversion 8 is the genomic region most closely association life-history variation the species range (Lowry & Willis 2010).

The phenotypic and fitness effects of the inversions on chromosomes 5 and 10 remain to be investigated, although the recent genome scan by Twyford & Friedman (2015) suggests that inversion 5 may be associated with life-history differences among populations. Like inversion 8, inversion 10 exhibits reduced intra-IM variation but increased divergence. In contrast, intra-IM π within inversion 5 is comparable to the genomewide average. The very elevated Dxy, high Var[π], but moderate π within the inversion 5 could be explained by reduced gene flow with allopatric populations and balancing local selection. The proportion of monophyletic windows within the inversion was substantially elevated: 35 of 77 (45%) of windows are monophyletic. The proportion of monophyletic windows within inversion 10 is even higher (700 of 1330, 52%). Interestingly, IMP is not solely responsible for breaking IM monophyly in any of the 42 polyphyletic chromosome 5 windows while IMP is the solely responsible for breaking IM monophyly in 241 of 630 chromosome 10 windows. The variable patterns of ancestry across inverted regions are perhaps not too surprising. Many different population/species of the *M. guttatus* complex are potential contributors to IM, and they may differ in the whether they have the IM orientation for a particular inversion.

Candidate genes for local adaptation

Haasl & Payseur (2016) recently reviewed the literature on genome scans for natural selection and concluded that 'the ability to detect individual instances of selection can decrease as the fraction of genome affected by linked selection grows'. This can be true for a number of reasons, but our study certainly one illustrates one them. Perhaps our most compelling evidence of selection is the pattern of sequence data in regions with suppressed recombination owing to inversions. This suppression produces the broad signal that we detect (many sites affected), but it also hinders discrimination of the specific genetic changes that are affecting fitness. However, we do see more localized patterns of polymorphism and divergence that are potential signatures of selection. We report these as tentative candidates, worthy of further study.

As a first step to identify specific genes potentially involved in local adaptation, we calculated outlier residuals from a Dxy vs. π contrast from all monophyletic windows. Appendix S2 reports the 2.5% most negative windows for IM π (controlling for Dxy). A total of 882 genes were located in these outlier windows, and they have significantly lower π_{syn} ($\pi_{syn\ outlier} = 0.014$, $\pi_{syn} = 0.035$, $P < 0.0001$; for genes with alignment length $\geq$200 and P-value ≤ 0.05, see 'Materials and methods') and lower nonsynonymous diversity ($\pi_{non\text{-}syn\ outlier} = 0.002$, $\pi_{non\text{-}syn} = 0.006$, $P < 0.0001$). Genes involved in the flowering pathway, germination timing, stress responses and trichome development are present in this outlier class. Several very intriguing candidates are immediately worth follow-up functional work. *DELAY OF GERMINATION1* (*DOG*1), a gene involved in timing of germination and flowering in *Arabidopsis thaliana* (Bentsink *et al.* 2006; Chiang *et al.* 2013) and several genes involved in flowering time in *A. thaliana,* including *Short Vegetative Phase* (*SVP*) (Lee *et al.* 2013) and ATMBD9 (Peng *et al.* 2006), are contained within the outlier windows. To complement the Dxy vs. π residual analysis, we selected monophyletic windows whose Var[π] and π values are below the 10% minimum for all monophyletic windows (Var[π] ≤ 0.00001 and $\pi \leq 0.00591$). For the low Var[π] and π windows, a total of 690 windows were identified (Appendix S3). Genes within this outlier group include a possible ortholog of *A. thaliana* gene AtMBD9 (Arabidopsis METHYL-CpG

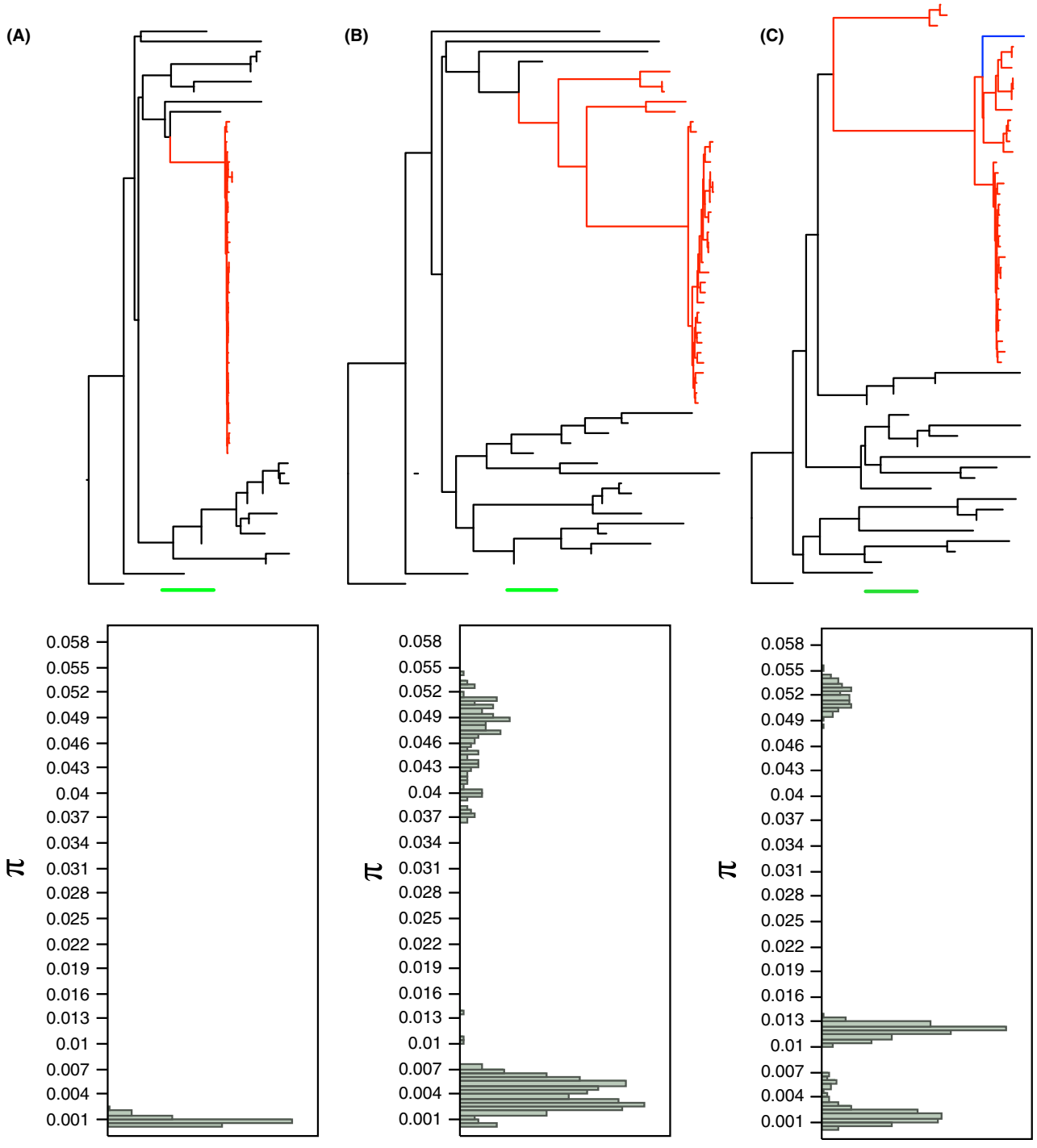

Fig. 3 Relationship of tree topology and distribution of pairwise π values within IM annuals. Highlighted red branches are IM annuals. Black bars are allopatric individuals. Green scale bar equals 0.01 for panels A, B and C. Pairwise π values between all 34 IM annuals were calculated (total 561 comparisons). Var[π] was calculated using these values. (A) Lowest Var[π] region for all monophyletic windows shows evidence of selective sweep. Top: tree shows that all IM annuals are very similar. Bottom: Distribution of raw π values for within IM comparisons shows that all IM annuals possess almost the exact same haplotype. (B) Highest Var[π] region for all monophyletic windows shows evidence of multiple distinct haplotypes within IM. (C) Second highest Var[π] region of all trees where IMP alone ruins IM monophyly shows evidence of introgression event including IMP and multiple distinct segregating haplotypes. Blue branch = IMP. IM, Iron Mountain. [Colour figure can be viewed at wileyonlinelibrary.com]

BINDING DOMAIN 9), which is an important regulator of flowering time through interactions with FLC (Peng *et al*. 2006), as well as a possible ortholog of Incurvata2 (ICU2) whose *Arabidopsis* mutant exhibits early flowering (Barrero *et al*. 2007). The fact that genes in both outlier groups regulate phenological transitions in *A. thaliana* and that phenology is critical for IM fitness suggests that research following up on these candidate loci may move us a step closer to a gene-level understanding of local adaptation.

Gene genealogies in a meta-population

A component of our analysis is the construction of distance-based trees relating IM and allopatric sequences for thousands of intervals across the entire genome. These trees estimate the genealogy of sequences subject to the important caveat that historical recombination within a locus will make our estimate a compromise among multiple true genealogies. Recognizing this, we do not use the trees for formal hypothesis testing, for example to provide a *P*-value on the null hypothesis that a genomic region is selectively neutral. Instead, the trees provide a compelling visual illustration of the relationship between molecular summary statistics, such as Tajima's *D* or r^2, and hypothesized evolutionary events, such as selective sweeps or introgression (Fig. 3, Figs S13 and S14, Supporting information). Second, the trees can provide classifiers, for example Is IM monophyletic or polyphyletic? This classification is useful for the analysis and interpretation of population genetic statistics that more formally characterize intrapopulation polymorphism, interpopulation divergence and linkage disequilibria (Table 1). The relative occurrence of these classes (monophyletic vs. polyphyletic) is clearly affected by major evolutionary events such as chromosomal inversion (Fig. 2).

The trees also provide an avenue for thinking about the genealogical process in a structured population with recombination. The observed phylogenetic relationship between sequences changes as one moves along a chromosome of *M. guttatus* (Fig. 4), which is to be expected

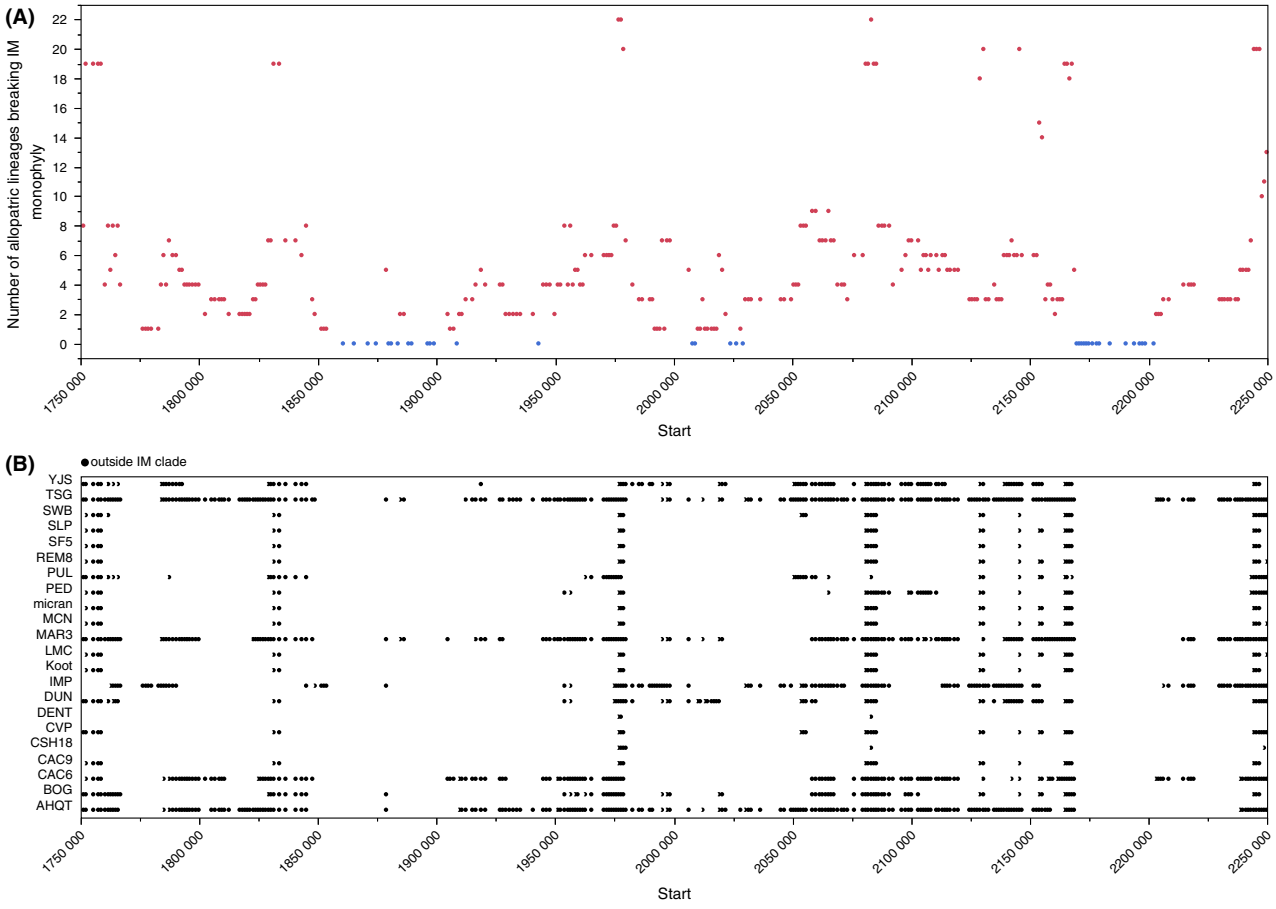

Fig. 4 (A) The number of allopatric individuals that break IM monophyly is depicted as a function of window location along chromosome 5. Monophyletic windows are blue dots. (B) The identity of allopatric individuals breaking IM monophyly is indicated for the same genomic region. IM, Iron Mountain. [Colour figure can be viewed at wileyonlinelibrary.com]

in an outbred species with recombination. Beyond that, several features of the data suggest that unlike selection, lineage coalescence is perhaps rarely a local phenomenon. First, the number of reproductive adults in IM (far exceeding 100 000 in most years) is at least an order of magnitude greater than the number of generations since the population was founded (likely <10 000 and perhaps much less). Second, the high frequency of polyphyly is predicted if lineages are coalescing outside of IM within the larger metapopulation. Third, even genomic windows where IM sequences are monophyletic do not imply a most recent common ancestor (MRCA) within IM. Sequences that coalesce within IM should differ only at sites that experienced mutation within the clade. We would not expect the same nucleotide positions to be polymorphic in neighbouring *M. guttatus* populations (except due to occasional independent, parallel mutation). In fact, pooled population sequencing data indicate that IM polymorphisms are *usually* shared between populations with the same alternative bases segregating (Monnahan *et al.* 2015). Perhaps most important is the simple fact that any two IM sequences are likely to exhibit a large number of nucleotide differences (high π) in a given genomic window. This suggests a substantial time since their MRCA. If the neutral mutation rate per base pair is 10^{-8} or 10^{-9}, the high π observed even in monophyletic windows (Table 1) implies numbers of generations to the MRCA that are far greater than the age of IM (Watterson 1975; Hudson 1990). In aggregate, these observations suggest that the genealogy of sequences within IM is determined more by population founding, migration and natural selection (hitch-hiking effects associated with linkage), than by the standard coalescent process of genetic drift *within* IM.

An important open question is the extent to which population-level processes, such as selection generated by local environmental conditions, influence the amount and distribution of genetic variation specieswide. Local adaptation not only influences the specific loci that are targets of selection, but also the 'effective migration rate' via its effect on the relative fitness of immigrant individuals (or gametes). As described in the 'Introduction', there is a long history of spatial population genetics in plant biology. Species vary enormously in the proportion of genetic variation that exists within relative to among demes. Genome sequencing can now provide a much more thorough characterization, quantifying absolute levels of variation within and among populations, and providing novel information from haplotype structure (LD). This study of *M. guttatus* illustrates that high differentiation among populations does not imply low intrapopulation variation. Hierarchical studies of other species, for example (Long *et al.* 2013), are required to determine the generality of this pattern.

Acknowledgements

We would like to thank Stephen Wright, Patrick Monnahan and Jenn Coughlan for advice on the content and/or comments on this manuscript. This work was supported by grants from the National Institutes of Health to J.K. and J.W. (R01 GM073990) and the National Science Foundation to J.R.P. (NPGI-IOS-1202778).

References

Balanyà J, Serra L, Gilchrist GW, Huey RB (2003) Evolutionary pace of chromosomal polymorphism in colonizing populations of *Drosophila subobscura*: an evolutionary time series. *Evolution*, **57**, 1837–1845.

Barrero JM, González-Bayón R, del Pozo JC, Ponce MR, Micol JL (2007) INCURVATA2 encodes the catalytic subunit of DNA polymerase α and interacts with genes involved in chromatin-mediated cellular memory in *Arabidopsis thaliana*. *The Plant Cell*, **19**, 2822–2838.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **263**, 1619–1626.

Begun DJ, Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*, **356**, 519–520.

Bentsink L, Jowett J, Hanhart CJ, Koornneef M (2006) Cloning of DOG1, a quantitative trait locus controlling seed dormancy in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 17042–17047.

Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL (2014) Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. *PLoS Genetics*, **10**, e1004410.

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.

Chan AH, Jenkins PA, Song YS (2012) Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genetics*, **8**, e1003090.

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, **2**, e64.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.

Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research*, **70**, 155–174.

Cheng C, White BJ, Kamdem C *et al.* (2012) Ecological genomics of *Anopheles gambiae* along a latitudinal cline: a population-resequencing approach. *Genetics*, **190**, 1417–1432.

Chiang GC, Barua D, Dittmar E *et al.* (2013) Pleiotropy in the wild: the dormancy gene DOG1 exerts cascading control on life cycles. *Evolution*, **67**, 883–893.

Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V (2002) A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science*, **298**, 1415–1418.

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.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Dey A, Chan CKW, Thomas CG, Cutter AD (2013) Molecular hyperdiversity defines populations of the nematode *Caenorhabditis brenneri*. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 11056–11060.

Duminil J, Hardy OJ, Petit RJ (2009) Plant traits correlated with generation time directly affect inbreeding depression and mating system and indirectly genetic structure. *BMC Evolutionary Biology*, **9**, 1–14.

Fabian DK, Kapun M, Nolte V *et al.* (2012) Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. *Molecular Ecology*, **21**, 4748–4769.

Fang Z, Pyhäjärvi T, Weber AL *et al.* (2012) Megabase-scale inversion polymorphism in the wild ancestor of maize. *Genetics*, **191**, 883–894.

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.

Fenster CB, Ritland K (1994) Quantitative genetics of mating system divergence in the yellow monkeyflower species complex. *Heredity*, **73**, 422–435.

Fishman L, Kelly JK (2015) Centromere-associated meiotic drive and female fitness variation in *Mimulus*. *Evolution*, **69**, 1208–1218.

Fishman L, Saunders A (2008) Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. *Science*, **322**, 1559–1562.

Fishman L, Kelly AJ, Willis JH (2002) Minor quantitative trait loci underlie floral traits associated with mating system divergence in *Mimulus*. *Evolution*, **56**, 2138–2155.

Flagel LE, Willis JH, Vision TJ (2014) The standing pool of genomic structural variation in a natural population of *Mimulus guttatus*. *Genome Biology and Evolution*, **6**, 53–64.

Friedman J, Twyford AD, Willis JH, Blackman BK (2015) The extent and genetic basis of phenotypic divergence in life history traits in *Mimulus guttatus*. *Molecular Ecology*, **24**, 111–122.

Gilburn AS, Day TH (1999) Female mating behaviour, sexual selection and chromosome I inversion karyotype in the seaweed fly, *Coelopa frigida*. *Heredity*, **82**, 276–281.

Govindaraju DR (1988) Mating systems and the opportunity for group selection in plants. *Evolutionary Trends in Plants*, **2**, 99–106.

Guerrero RF, Rousset F, Kirkpatrick M (2012) Coalescent patterns for chromosomal inversions in divergent populations. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **367**, 430–438.

Haasl RJ, Payseur BA (2016) Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, **25**, 5–23.

Hall MC, Willis JH (2006) Divergent selection on flowering time contributes to local adaptation in *Mimulus guttatus* populations. *Evolution*, **60**, 2466–2477.

Hall MC, Basten CJ, Willis JH (2006) Pleiotropic quantitative trait loci contribute to population divergence in traits associated with life-history variation in *Mimulus guttatus*. *Genetics*, **172**, 1829–1844.

Hall MC, Lowry DB, Willis JH (2010) Is local adaptation in *Mimulus guttatus* caused by trade-offs at individual loci? *Molecular Ecology*, **19**, 2739–2753.

Hamrick JL, Godt MJW (1996) Effects of life history traits on genetic diversity in plant species. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **351**, 1291–1298.

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.

Holeski L, Monnahan P, Koseva B *et al.* (2014) A high-resolution genetic map of yellow monkeyflower identifies chemical defense QTLs and recombination rate variation. *G3: Genes|Genomes|Genetics*, **4**, 813–821.

Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, **7**, 1–44.

Jacobs GS, Sluckin TJ, Kivisild T (2016) Refining the use of linkage disequilibrium as a robust signature of selective sweeps. *Genetics*, **203**, 1807–1825.

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

Kaplan NL, Hudson RR, Langley CH (1989) The "hitchhiking effect" revisited. *Genetics*, **123**, 887–899.

Kelly JK (1997) A test of neutrality based on interlocus associations. *Genetics*, **146**, 1197–1206.

Kelly JK (2003) Deleterious mutations and the genetic variance of male fitness components in *Mimulus guttatus*. *Genetics*, **164**, 1071–1085.

Kelly JK, Arathi HS (2003) Inbreeding and the genetic variance of floral traits in *Mimulus guttatus*. *Heredity*, **90**, 77–83.

Kelly AJ, Willis JH (1998) Polymorphic microsatellite loci in *Mimulus guttatus* and related species. *Molecular Ecology*, **7**, 769–774.

Kooyers NJ, Greenlee AB, Colicchio JM, Oh M, Blackman BK (2015) Replicate altitudinal clines reveal that evolutionary flexibility underlies adaptation to drought stress in annual *Mimulus guttatus*. *New Phytologist*, **206**, 152–165.

Krimbas CB, Powell JR (1992) *Drosophila Inversion Polymorphism*. CRC Press, Boca Raton, Florida.

Kubota S, Iwasaki T, Hanada K *et al.* (2015) A genome scan for genes underlying microgeographic-scale local adaptation in a wild *Arabidopsis* species. *PLoS Genetics*, **11**, e1005361.

Lack JB, Cardeno CM, Crepeau MW *et al.* (2015) The *Drosophila* genome nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. *Genetics*, **199**, 1229–1241.

LaMariposa GitHub: popgen_scripts, https://github.com/LaMariposa/popgen_scripts.

Langley CH, Crow JF (1974) The direction of linkage disequilibrium. *Genetics*, **78**, 937–941.

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

Lee JH, Ryu H-S, Chung KS *et al.* (2013) Regulation of temperature-responsive flowering by MADS-box transcription factor repressors. *Science*, **342**, 628–632.

Lee YW, Fishman L, Kelly JK, Willis JH (2016) A segregating inversion generates fitness variation in yellow monkeyflower (*Mimulus guttatus*). *Genetics*, **202**, 1473–1484.

Leffler EM, Bullaughey K, Matute DR *et al.* (2012) Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biology*, **10**, e1001388.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Long Q, Rabanal FA, Meng D *et al.* (2013) Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nature Genetics*, **45**, 884–890.

Loveless MD, Hamrick JL (1984) Ecological determinants of genetic structure in plant populations. *Annual Review of Ecology and Systematics*, **15**, 65–95.

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.

Lowry DB, Hall MC, Salt DE, Willis JH (2009) Genetic and physiological basis of adaptive salt tolerance divergence between coastal and inland *Mimulus guttatus*. *New Phytologist*, **183**, 776–788.

Mackay TF, Richards S, Stone EA *et al.* (2012) The *Drosophila melanogaster* genetic reference panel. *Nature*, **482**, 173–178.

Mojica JP, Kelly JK (2010) Viability selection prior to trait expression is an essential component of natural selection. *Proceedings of the Royal Society of London B: Biological Sciences*, **277**, 2945–2950.

Mojica JP, Lee YW, Willis JH, Kelly JK (2012) Spatially and temporally varying selection on intrapopulation quantitative trait loci for a life history trade-off in *Mimulus guttatus*. *Molecular Ecology*, **21**, 3718–3728.

Monnahan PJ, Colicchio J, Kelly JK (2015) A genomic selection component analysis characterizes migration-selection balance. *Evolution*, **69**, 1713–1727.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, **70**, 3321–3323.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, **76**, 5269–5273.

Nolte V, Pandey RV, Kofler R, Schlötterer C (2013) Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in *Drosophila mauritiana*. *Genome Research*, **23**, 99–110.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

Nybom H (2004) Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Molecular Ecology*, **13**, 1143–1155.

Peng M, Cui Y, Bi Y-M, Rothstein SJ (2006) AtMBD9: a protein with a methyl-CpG-binding domain regulates flowering time and shoot branching in *Arabidopsis*. *The Plant Journal*, **46**, 282–296.

Puzey J, Vallejo-Marín M (2014) Genomics of invasion: diversity and selection in introduced populations of monkeyflowers (*Mimulus guttatus*). *Molecular Ecology*, **23**, 4472–4485.

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. *Trends in Genetics*, **16**, 276–277.

Roesti M, Kueng B, Moser D, Berner D (2015) The genomics of ecological vicariance in threespine stickleback fish. *Nature Communications*, **6**, 8767.

Ross-Ibarra J, Wright SI, Foxe JP *et al.* (2008) Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS One*, **3**, e2411.

Schmid KJ, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T (2005) A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics*, **169**, 1601–1615.

Scoville A, Lee YW, Willis JH, Kelly JK (2009) Contribution of chromosomal polymorphisms to the G-matrix of *Mimulus guttatus*. *New Phytologist*, **183**, 803–815.

Scoville AG, Lee YW, Willis JH, Kelly JK (2011) Explaining the heritability of an ecologically significant trait in terms of individual QTLs. *Biology Letters*, **7**, 896–898.

Slatkin M (1987) Gene flow and the geographic structure of natural populations. *Science*, **236**, 787–792.

Storz JF, Kelly JK (2008) Effects of spatially varying selection on nucleotide diversity and linkage disequilibrium: insights from deer mouse globin genes. *Genetics*, **180**, 367–379.

Storz JF, Natarajan C, Cheviron ZA, Hoffmann FG, Kelly JK (2012) Altitudinal variation at duplicated β-globin genes in deer mice: effects of selection, recombination, and gene conversion. *Genetics*, **190**, 203–216.

Strobeck C (1983) Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. *Genetics*, **103**, 545–555.

Sweigart A, Karoly K, Jones A, Willis JH (1999) The distribution of individual inbreeding coefficients and pairwise relatedness in a population of *Mimulus guttatus*. *Heredity*, **83**, 625–632.

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

Talevich E, Invergo BM, Cock PJ, Chapman BA (2012) Bio. Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, **13**, 209.

Twyford A, Friedman J (2015) Adaptive divergence in the monkey flower *Mimulus guttatus* is maintained by a chromosomal inversion. *Evolution*, **69**, 1476–1486.

Vos R (2015) Monophylizer, https://github.com/naturalis/monophylizer/tree/v1.0.1.

Wade MJ (2016) *Adaptation in Metapopulations: How Interaction Changes Evolution*. University of Chicago Press, Chicago, Illinois.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Willis JH (1993) Partial self fertilization and inbreeding depression in two populations of *Mimulus guttatus*. *Heredity*, **71**, 145–154.

Willis JH (1996) Measures of phenotypic selection are biased by partial inbreeding. *Evolution*, **50**, 1501–1511.
Willis JH (1999a) Inbreeding load, average dominance, and the mutation rate for mildly deleterious alleles in *Mimulus guttatus*. *Genetics*, **153**, 1885–1898.
Willis JH (1999b) The role of genes of large effect on inbreeding depression in *Mimulus guttatus*. *Evolution*, **53**, 1678–1691.
Wright S (1931) Evolution in Mendelian population. *Genetics*, **16**, 97–159.
Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: *Proceedings of the VI International Congress of Genetics*, pp. 356–366.
Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration–selection balance. *Evolution*, **65**, 1897–1911.
Zhang Z, Li J, Zhao X-Q *et al.* (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics, Proteomics & Bioinformatics*, **4**, 259–263.

J.P., J.W. and J.K. designed this experiment. J.P. made the libraries and directed sequencing. J.P. and J.K. performed all genomic analyses and wrote the manuscript.

Data availability

All sequence data generated here are available on the Short Read Archive. SRA numbers: SAMN05852485-SAMN05852522.

Supporting information

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

Table S1 Summary of samples and sequence data used in this study.

Table S2 Fraction of bases called in coding regions (CDS).

Table S3 Fraction of bases called by linkage group (averaged across all lines).

Table S4 Fraction of windows per linkage group where IM is monophyletic.

Table S5 Fraction of polyphyletic trees ruined by a single outgroup sample.

Fig. S1 Size distribution of residual heterozygosity blocks in filtered lines.

Fig. S2 Similarity of all IM lines before filtering.

Fig. S3 Average pairwise divergence after filtering.

Fig. S4 Lower amount of missing data in less diverged regions.

Fig. S5 Gene specific nucleotide diversity values.

Fig. S6 Genome-wide distributions of core population genetic statistics.

Fig. S7 Spatial distribution of pi across the entire genome.

Fig. S8 Spatial distribution of variance(π) across the entire genome.

Fig. S9 Relationship of variance(π) to nucleotide diversity and LD.

Fig. S10 Linkage Disequilibrium measured as r^2 within IM.

Fig. S11 Genome-wide patterns of recombination shows clear evidence of recombination hotspots.

Fig. S12 Fraction of monophyletic windows was calculated in 250 000 bp non-overlapping windows.

Fig. S13 Relationship of variance(π) to tree topology.

Fig. S14 This window on chromosome 8 at position 8.25 Mb exhibits the topology predicted by a recent selective sweep within IM.

Fig. S15 Geographic distribution of samples used in this study.

Appendix S1. Alignment and genotype calling

Appendix S2 The genomic locations of windows with low Var [π] and low π.

Appendix S3 Interval list of outlier windows.