# Whole genome sequencing of 56 *Mimulus* individuals illustrates population structure and local selection

Joshua R. Puzey[1,2], John H. Willis[2] and John K. Kelly[3,*]

[1] Department of Biology, College of William and Mary, Williamsburg, Virginia, 23187

[2] Department of Biology, Duke University, Durham, North Carolina, 27708

[3] Department of Ecology and Evolution, University of Kansas, Lawrence, Kansas, 27708

*Author for correspondence, Email: jkk@ku.edu

## ABSTRACT

Across western North America, *Mimulus guttatus* exists as many local populations adapted to site-specific challenges including salt spray, temperature, water availability, and soil chemistry. Gene flow between locally adapted populations will effect genetic diversity in both local demes and across the larger meta-population. A single population of annual *M. guttatus* from Iron Mountain, Oregon (IM) has been extensively studied and we here building off this research by analyzing whole genome sequences from 34 inbred lines from IM in conjunction with sequences from 22 *Mimulus* individuals from across the geographic range. Three striking features of these data address hypotheses about migration and selection in a locally adapted population. First, we find very high intra-population polymorphism (synonymous $\pi$ = 0.033). Variation outside genes may be even higher, but is difficult to estimate because excessive divergence affects read mapping. Second, IM exhibits a significantly positive genome-wide average for Tajima's D. This indicates allele frequencies are typically more intermediate than expected from neutrality, opposite the pattern observed in other species. Third, IM exhibits a distinctive haplotype structure. There is a genome-wide excess of positive associations between minor alleles; consistent with an important effect of gene flow from nearby *Mimulus* populations. The combination of multiple data types, including a novel, tree-based analytic method and estimates for structural polymorphism (inversions) from previous genetic mapping studies, illustrates how the balance of strong local selection, limited dispersal, and meta-population dynamics manifests across the genome.


Keywords: population genomics, evolution, genomics, migration, selection, *Mimulus*, inversion

Running title: Population genomics of monkeyflower

## INTRODUCTION

The *Mimulus guttatus* species complex is an enormous collection of localized populations, some of which are recognized as distinct taxa (e.g. *Mimulus nasutus*) [1, 2] or ecotypes (e.g. annual and perennial) [3]. Over the past two decades, *M. guttatus* has emerged as a powerful model system for addressing evolutionary and ecological questions related to local adaptation, inbreeding, and maintenance of variation [4, 5]. Endemic to western North America, *M. guttatus* has adapted to a wide range of habitats including serpentine barrens, metal-rich mine tailings, huge elevation ranges, and oceanic salt spray [6-9]. These populations are often inter-fertile to varying degrees. Gene flow occurs between populations [10], but potentially strong selection against immigrant genotypes [6] allows substantial genetic differentiation.

Previous studies have clearly shown the level of genetic differentiation increases with distance among populations in *M. guttatus*. Lowry and Willis [11] estimated $F_{ST}$ = 0.48 across a set of 30 populations spanning a latitudinal range from 35-45 degrees. Similarly, Twyford and Friedman [12] 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* across Oregon (V. Koelling and J. K. Kelly, unpublished results). This survey included the Iron Mountain population (hereafter IM). 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 [13]. Taken in total, these data indicate geographic limits on gene flow, suggesting that *M. guttatus* populations should be able to adapt to local environmental conditions. Local adaptation has been demonstrated in several reciprocal transplant experiments [6, 8]. When gene flow does occur, it is likely to introduce divergent alleles into a locally adapted population. Genomic migration-selection balance predicts that local populations experiencing high levels of gene flow will result in high genome-wide levels of nucleotide diversity [14, 15]. Genomic regions subject to local selection are expected to be less permeable to incoming haplotypes and should

have lower nucleotide diversity, higher levels of absolute divergence between populations (Dxy), and distinct patterns of linkage disequilibria [14, 16-18].

To explore these predictions, we examine the population genomics of a single focal population in conjunction with data from the larger meta-population. Our focal population, IM, has been the subject of intense evolutionary and ecological research for the past 30 years. IM is an annual population where 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 on IM [19, 20] and the population exhibits adaption to local conditions [6]. Despite this, IM retains high internal variability in both molecular and quantitative genetic traits [21, 22]. Recent studies have also revealed extensive structural (chromosomal) variation. Major inversions segregate both within IM (on Linkage Groups 6 and 11 [23, 24]) and also between IM and other populations (on Linkage Groups 5, 8, and 10 [3, 12, 25]). Inversions are predicted to directly influence migration–selection balance, and may have extensive effects on molecular variation owing to recombination suppression.

We use a combination of analytical techniques to characterize patterns of polymorphism, divergence and haplotype structure. 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. Two primary conclusions emerge. First, despite the fact that IM is a localized population composed of only a few hundred thousand individuals, synonymous site diversity is among the highest reported for any organism. Second, the varying structure of polymorphism within IM relative to divergence from other *M. guttatus* populations indicates an overall genomic 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

***Plant samples*:** Approximately 1200 independent lines of *M. guttatus,* each founded from the seed set of a separate field-collected plant sampled from the IM [26]. Each line was subsequently maintained in a greenhouse by single-seed descent (self-fertilization) for 5-13 generations. As expected, these lines are almost completely homozygous at microsatellite loci with different lines fixed for different alleles [7]. DNA from 39 of these lines was newly generated and subsequently combined with data from 9 previously sequenced IM lines ([27]; reads downloaded from the JGI Short Read Archive). Sequence data from 18 "outgroups" (other populations or species in the complex; Supplemental Table 1) data was downloaded from SRA [27]. For all samples (both IM and outgroups), average read depth after filtering (as determined from VCF file using vcftools --depth) ranged from 2.6-24.27 (mean=6.68; calculated for genotyped bases on LG1, no indels included, Supplemental Table 1). We also sequenced the perennial species, *Mimulus decorus,* hereafter called Iron Mtn. Perennial (IMP). This species occurs in close proximity to the annual IM population.

***DNA extraction, library preparation, and sequencing*:** We collected and froze leaf tissue for DNA extraction. We extracted DNA from leaf tissue using the Epicentre Leaf MasterPure kit (Epicentre, USA). Libraries for Illumina sequencing were made using the Illumina Nextera DNA kit (Illumina, USA), which utilizes a transposon-based system to integrate within and tag DNA. 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 bioanalyzer (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 *Mimulus guttatus* v2.0 reference genome (http://www.phytozome.net/) using bowtie2 [28]. Next, we converted SAM alignment files to binary format using samtools [29] and then processed alignments with Picardtools (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) [30]. GATK UnifiedGenotyper was used to call genotypes (details in supplement). Genotype VCF files were converted to tabbed format using vcftools (vcf-to-tab) [31]. Details can be found in supplementary methods.

After the initial genotyping, we masked putative SNPs that were excessively heterozygous (in more than 25% of lines) as these are likely due to mis-mapped reads. This is likely, for example, 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 dividing 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 6 generations and 0.78% after 7 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 (about 125 cM per chromosome) (Supplemental Figure 1).

***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 (Supplemental Figure 2). For instance, IM777 is 0.997 similar to IM323. We thus determined these lines to be relatives and eliminated the IM323 from subsequent analyses. In addition, we also calculated the proportion of divergent sites through pairwise comparisons and used these values to identify related individuals. However, a low level of variability is consistent with a random sample from a large well-mixed population. IM109 is more divergent, but when considering the whole genome, is not a genuine outlier (Supplemental Figure 3). 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.

***Relationship of missing data and divergence*:** We delineated windows containing 500 SNPs, and within each window of each line, calculated (1) the number of called and uncalled sites and (2) the number of SNPs called for the reference allele as opposed to the alternative allele. The fraction missing data was calculated from (1), and the window divergence (fraction of calls to alternate) from (2). We performed a logistic regression in R with fraction missing as the response and divergence as the predictor: glm(formula = logit1$frac.missing ~ logit1$divergence, family = binomial). This revealed a strong relationship between data missingness and divergence (fraction of called SNPs that differ from the reference genome; Supplemental Figure 4A). Given this relationship, we opted to focus our analyses where data was most complete using two complimentary approaches. First, based on the fact that the fraction of missing data is considerably lower in coding regions (Supplemental Figure 4B) we conducted a series of gene-based analyses (e.g. synonymous versus non-synonymous diversity). The mean fraction of called bases in coding regions, calculated for each line, ranged from 0.63 to 0.86 (Supplemental Table 2). Second, we identified genomic windows (genic and inter-genic 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 30 of the 34 unrelated IM lines. The resulting windows ranged from 10,000-3,427,432 bases with a mean and median of 39,044 and 18,478 bases, respectively.

Allowing 1,000 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 non-synonymous π)*:** We converted filtered genotype files to fasta format for the entire genome for each separate line. When recreating line specific fasta files, missing data was not imputed, and indels and heterozygous sites were suppressed. Gffread [28] was used to extract coding sequences from individual fasta file. 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 non-synonymous diversity through pairwise comparisons of all lines using the KaKs_Calculator (Nei and Gojobori model) described in Zhang et al [32]. Only diversity measurements derived from genes with alignment lengths greater than 1000 bases were included (N=29,421). A Ka and Ks value was computed for each gene and was used to calculate genome-wide mean Ka and Ks.

***Window analyses*:** We calculated statistics of polymorphism, divergence, and genealogy within windows of 10,000 genotyped bases. We created a phylogenetic tree for every window using EMBOSS fdnadist [33] to calculate a nucleotide distance matrix (Jukes-Cantor substitution model). Trees were inferred from the distance matrix using EMBOSS fneighbor [33] and rooted using *Mimulus dentilobius*. Of the 74,445 windows, fneighbor failed to parse the distance matrix for 28 windows. These windows were excluded from population genetic statistics results. Next, for each tree we determined whether IM was monophyletic or polyphyletic using a custom perl script [34] dependent on the Bio::Phylo toolkit [35]. This perl script searches a newick file and asks whether a specified group of individuals form a monophyletic clade. In the cases that IM was polyphyletic, we further explored the data by asking how many outgroup samples had to be removed to restore IM monophyly using the perl monophyletic output [34] and custom perl scripts.

188 We calculated S (the number of polymorphisms), π (nucleotide diversity), Tajima's D [36], haplotype
189 homozygosity, and LD statistics in each window using custom python scripts and VariScan [37]. For the
190 linkage disequilibrium (D), we estimated the association of the minor alleles (less common base) at each
191 contrasted SNP pair. Positive D indicates that minor alleles are positively associated [38]. We
192 standardize D as the correlation coefficient, r = D/ Sqrt[p(1-p)q(1-q)], and from that, calculate $r^2$ (the $Z_{nS}$
193 test for selection is $r^2$ conditioned on S [39]). We also calculated r and $r^2$ for SNP pairs across each
194 chromosome to estimate the long-range pattern of LD. For comparison to observed LD, we performed
195 neutral simulations using calibrated, empirical estimates for 4Nu (from nucleotide diversity) and 4Nr
196 (from LDhelmet as described below) by updating the programs used in Storz et al [40]. Absolute
197 nucleotide divergence, Dxy, between IM annuals and all outgroups was calculated using a perl script [41]
198 dependent on BioPerl::PopGen modules (Dxy is equivalent to $\pi_{XY}$ [42]). Dxy was calculated on a single
199 base increment for all sites that had at least one IM and one outgroup individual genotyped. Next, using
200 these values, an average Dxy value was calculated for the same 10,000 genotyped base windows used for
201 other population genetic statistics.
202
203 ***Recombination rates within IM*:** We used LDhelmet [43] to estimate fine-scale recombination
204 rates with recalled genomes in fasta format as inputs. First, using the "find_confs" command, 50 SNP
205 windows were used to scan the genome and create a haplotype configuration file. Next, a likelihood
206 lookup table and Pade coefficients were generated using a population scaled mutation rate of 0.015 (this
207 was based on a preliminary estimate for genome-wide π within IM). In the final step, the "rjmcmc"
208 command was run using the previously generated haplotype configuration, likelihood table, and Pade
209 coefficients and a Jukes-Cantor mutation matrix to estimate recombination rates. Exon specific
210 recombination rates were calculated. Using bedtools [44] intersect command, coordinates of exons
211 extracted from the *M. guttatus* gff3 gene annotation file (phytozome.net) were combined with
212 recombination rates calculated by LDhelmet [43]. Only pairs of SNPs contained within exons were used.
213 Using this information, we are able to look at gene and exon specific recombination rates.

## RESULTS

***Nucleotide diversity within IM*:** Within genes, synonymous nucleotide diversity is very high $\pi_{syn}$ = 0.033 (Supplement Figure 5). Mean non-synonymous diversity is $\pi_{non\text{-}syn}$ = 0.006 (Supplemental Figure 5). Nucleotide diversity was significantly lower in interior exons (Supplemental Figure 6). Exon numbers one through five had mean π values of 0.0122, 0.0111, 0.0102, 0.0098, and 0.0092, respectively. Exon specific GC content was correlated with nucleotide diversity; π and GC content were both highest in the first exon (Supplemental Figure 6). When comparing the first exon to numbers 2-5, interior exons have lower GC ($\%GC_{exon\#1}$=48.0, $\%GC_{exon\#2\text{-}5}$=44.4, $p<0.0001$, *t-test*). To explore this relationship further, exons were placed in bins based on position (1st, 2nd, 3rd exon, etc.) and further binned by GC content (40-45, 45-50, 50-55, and 55-60% GC content). Interestingly, in the first exon, π exhibits a negative relationship with GC content, while in exons 2-5, π is positively related with GC content (Supplemental Figure 7). Within the 10,000 genotyped base windows of IM lines (genic and inter-genic DNA), the average nucleotide diversity was 0.014. Across the genome, π varied from 0.00-0.03 (Supplemental Figure 8A and 9).

***Linkage disequilibrium and recombination rate:*** In most genomic windows there is a striking excess of positive LD (Figure 1). A large proportion of genomic windows exhibit stronger association of minor alleles than predicted with neutrality. If an IM line harbors the less frequent base at a SNP, it is much more likely to have the less frequent base at neighboring SNPs. The neutral distribution of Figure 1 was obtained using the average, genome-wide ρ (4Nr) of 0.0042 obtained from LDhelmet [43]. When measured as $r^2$, linkage disequilibrium is high at short distances (~100bp) and shows a rapid decay with sequence distance (Supplemental Figure 12). It should be noted that the pattern of long-range LD differs among chromosomes (Supplemental Figure 12). The average genic ρ was 0.0052 and recombination hotspots were clearly evident (Supplemental Figure 13). On an exon specific level, recombination rates are highest in the first exons and decrease in interior exons (Supplemental Figure 6) ($r_{exon\#1}$=0.0073,

SE=0.0002; $r_{exon\#2\text{-}5}$=0.0037, SE=0.0002; p<0.0001, *t-test*). Interestingly, in the first exon, π exhibits a negative relationship with GC content, while in exons 2-5, π is positively related with GC content (Supplemental Figure 7).

***Divergence of IM lines from other M. guttatus populations:*** Diversity within IM was lower than divergence of IM sequences from outgroups (Dxy): Mean Dxy = 0.038, range 0.002-0.166 (Supplemental Figure 8E) with several clear peaks of high Dxy (Figure 2). The variance in pairwise divergence (among the 561 contrasts between 34 lines within each genomic window) also exhibits many localized peaks across the genome (Supplemental Figure 10). 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 $r^2$ (Supplemental Figures 8F and 10, 11).

We located each of the three inversions mapped in the IMxPR RIL population [25] by locating the markers to locations in the v2 genome build (bars in Figure 2). This is a cross between annual (IM) and perennial (PR) genotypes. Absolute sequence divergence (Dxy) is significantly elevated within all three inversion regions relative to genome-wide averages: $Dxy_{Genome}$ = 0.037, $Dxy_{inversion(Lg8)}$ = 0.044 (p<0.0001), $Dxy_{inversion(LG5)}$ = 0.068 (p<0.0001), and $Dxy_{inversion(LG10)}$ = 0.042 (p<0.0001). The LG10 and LG8 inversions shows significantly lower overall nucleotide diversity while the LG5 is on not statistically different from genome-wide levels: $\pi_{Genome}$=0.014, $\pi_{inversion(LG10)}$ = 0.010 (p<0.0001), $\pi_{inversion(LG8)}$ = 0.012 (p<0.0001), and $\pi_{inversion(LG5)}$ = 0.014 (p=0.6603). Interestingly, Var[π] is statistically elevated in the LG5 inversion but statistically lower in the LG10 and LG8 inversions: $Var[\pi]_{Genome}$= 0.000085, $Var[\pi]_{inversion(LG5)}$ =0.000109 (p=0.0002), $Var[\pi]_{inversion(LG10)}$ =0.000060 (p<0.0001), and $Var[\pi]_{inversion(LG8)}$ = 0.000053 (p<0.0001).

***Distribution of monophyletic IM clusters across the genome:*** We constructed a phylogenetic tree for every window in the genome including both IM and outgroup samples. The monophyly of IM samples

was evaluated individually for each tree. In 10,504 phylogenetic trees all IM samples were monophyletic (Supplemental Table 4). The majority (64,913) of windows showed IM as polyphyletic – some IM lines more similar to outgroup sequences than other IM lines (Supplemental Table 4). Genome-wide, $IM_{monphyletic}$ windows were found both in gene dense and gene sparse regions. To further explore this data, we calculated the number of outgroup samples that would have to be removed from the tree for IM to be monphyletic. For 6,958 of the trees, only one outgroup sequence would have to be removed to restore IM monophyly (Supplemental Table 5). For 48% of these 6,958 trees, the geographically proximate Iron Mtn. Perennial (IMP) was responsible for IM paraphyly (Supplemental Table 5). This pattern is evident across all linkage groups but the incidence is highest for LG8 where 65% of polyphyletic trees where IM monophyly is ruined by a single individual are due to IMP (Supplemental Table 5).

***Relationship of Var[π] and tree topology:*** Combining within population Var[π] and tree topology allows for identification of genomic regions that have experienced a selective sweep or introgression event (Figure 3 and Supplemental Figure 15). On average, we expect monophyletic regions of the genome to have lower Var[π]. Introgression of divergent haplotypes should elevate Var[π], a trend we observe (Table 1). To test the effectiveness of the Var[π] statistic as a technique for identifying selective sweeps, we selected the lowest Var[π] tree from all monophyletic windows and looked at its topology and distribution of pairwise π. The tree created from the lowest Var[π] window shows a topology indicative of a selective sweep – short branches within the monophyletic IM clade. For this same window, within IM pairwise-π values are very small supporting the finding that all IM individuals within this block possess nearly the exact same haplotype (Figure 3A).

Next, we picked the highest Var[π] region for all monophyletic windows and created a tree and π distribution for this region (Figure 3B). These results indicate that for this genomic window the IM population contains several highly diverged sequences. Introgression may well have been the original source of the divergent lineages in Figure 3B, although local they may be maintained by selective

processes within IM. It is also possible to use this dataset to identify the genomic effects of introgression from individual outgroup 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 hapolotypes (long branches within IM and multimodal π distribution) (Figure 3C, Supplemental Figure 15).

***Contrast of population genetic statistics between monophyletic and polyphyletic regions:*** To determine the relationship between gene trees and population genetic statistics, we compared the distribution of each statistic between IM monophyletic windows and IM polyphyletic windows. 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 to monophyletic windows (Table 1). Dxy was significantly lower in polyphyletic windows relative to monophyletic windows (Table 1).

## 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 datasets specific to demes, populations of individuals connected by mating in recent time, e.g. [45]. This study uses a combination of intensive within population and across population whole-genome sequencing to explore genomic patterns indicative of selection and introgression. In this discussion, we summarize genome-wide trends, discuss methodological approaches (tree based analysis combined with Var[π]) to explore within population sequencing data, and take a first step from genomic observations to underlying genes.

Nucleotide diversity within Iron Mountain (IM) is extremely high for a single population ($\pi_{Genome}$ = 0.014, $\pi_{syn}$ = 0.033). The genomic estimate of 0.014 is likely underestimated given the strong tendency for

318 missing data to increases with divergence from the reference genome (Supplemental Figure 4).
319 Downward bias of $\pi$ outside of genic regions results from ascertainment; we are less likely to map (and
320 thus analyze) sequences that are most divergent. This is not surprising, but to our knowledge, has not
321 been clearly demonstrated previously. We expect this to be a general phenomenon extending across most
322 studies of this kind. Here, we suspect that the high level of within population insertion/deletion variation
323 reported for IM annuals contributes to incomplete mapping and subsequent underestimation of nucleotide
324 diversity [27].
325
326 Leffler [46] has recently summarized nucleotide diversity across a wide-range of species, classifying
327 estimates by sampling strategy (one population or multiple populations), site type (synonymous, non-
328 synonymous, etc.), and chromosome type (autosome or sex). From their dataset, we extracted 37 one-
329 population autosomal nucleotide diversity estimates spanning eight Phyla (Arthropoda, Chloropyta,
330 Chordata, Echinodermata, Magnoliophyta, Mollusca, Pinophyta, and Porifera). Considering single
331 population autosomal $\pi_{syn}$ (N=9, species from five phylum represented: Arthropoda, Chlorophyta,
332 Chordata, Mollusca, and Pinophyta), Leffler [46] observed a mean and median of 0.014 and 0.011,
333 respectively (range 0.001-0.033). Across multiple populations (N=50) for which autosomal $\pi_{syn}$ has been
334 reported, values ranged from 0.000-0.035 with a mean and median of 0.010 and 0.006, respectively. It has
335 previously been reported that *Mimulus guttatus* has multiple-population synonymous diversity values of
336 $\pi_{syn}$=0.061 [47]. Thus, for both single population and across multiple-populations, *Mimulus guttatus* ties
337 and exceeds, respectively, the highest levels of autosomal synonymous nucleotide diversity reported. It is
338 worth noting of all autosomal nucleotide diversity ($\pi$) values reported by Leffler (N=207) (not filtered
339 based on site-type), four-fold degenerate site diversity ($\pi$=0.080) of the sea squirt (Ciona roulei) was the
340 only higher $\pi$ value than *M. guttatus* multiple population $\pi_{syn}$=0.061 [46]. These results indicate that
341 *Mimulus guttatus* has one of the highest within population and within species level of nucleotide diversity
342 known for any organism.
343

344 Nucleotide diversity at a locus is determined by the number of SNPs and the frequencies of alternative
345 alleles at each SNP. Relating to the latter, a second notable feature of IM is the substantially positive
346 mean value for Tajima's D (Table 1). Genomic surveys in several Drosophila species [45, 48, 49], as
347 well as in Arabidopsis [50], reveal negative mean values for Tajima's D. In other words, the allele
348 frequency spectrum is skewed towards rare alleles in most genomic windows. In contrast, we find a
349 tendency for allele frequencies to be more intermediate than expected from the equilibrium neutral model
350 (predicted mean zero for Tajima's D). The meiotic drive locus on chromosome 11 illustrates this pattern:
351 Mean Tajima's D = 0.46 across 561 genomic windows from position 9.5 Mb to 11.7 Mb. Previous study
352 of this region has revealed a balanced polymorphism [24, 51]. An allele that causes meiotic drive has
353 recently increased to a population frequency of approximately 35%. The genomic signature of this sort of
354 recent event is subtle, unlike ancient balanced polymorphisms where alternative alleles may have highly
355 divergent sequences [52]. In essence, the Drive haplotype has been 'sampled' from the SNPs resident in
356 the population, and its increase owing to selection is associated with incremental shifts in the allele
357 frequencies across a large genomic region. It is remarkable that the sequence-level pattern of variation
358 within the Drive locus – a particular haplotype is present in a non-trivial minority of lines (20-30%) – is
359 observed at many loci in the genome.
360
361 *Genome-wide effects of migration and localized signatures of selection:* Most genomic windows exhibit
362 polyphyly and a haplotype structure suggesting low but significant migration of immigrant genotypes into
363 IM. The former is simply the observation, that in most of the genome, some IM lines are more similar to
364 lines from other populations than to other IM lines. The high frequency of polyphyly is not caused by a
365 few divergent IM lines repeatedly breaking monophyly. Instead, most IM lines exhibit similarity to
366 outgroup sequences within portions of their genomes; the identity of lines that break monophyly changing
367 across the genome. This is expected given previous evidence that IM is an internally well-mixed, outbred
368 population [53].
369

370 IM polyphyly could be due to either ancestral polymorphism that is continuing to segregate or
371 introgression from neighboring populations. Both are likely relevant, but two observations suggest an
372 important contribution of migration. First, we observed a specific pattern of LD in which the rarer
373 alleles at pairs of SNPs are positively associated on average across the genome (Figure 1). This
374 tendency is significantly elevated in polyphyletic windows (Table 1). Most sequencing studies do not
375 specify associations between SNPs in terms of features of alternative alleles, and as a consequence, the
376 direction of LD is meaningless. Indeed, direction is lost when calculating $r^2$, which is commonly used to
377 measure the strength of association between SNPs (e.g. Supplemental Figure 12). Here, in order to
378 evaluate the potential contribution of introgression, we polarize alleles based on allele frequency in IM.
379 Measured this way, the absolute value of LD can inform questions about evolutionary process. For
380 example, Langley and Crow [38] developed an epistatic selection model to explain the negative LD
381 observed in allozyme data. In IM, LD is highly variable in both direction and magnitude, but we suggest
382 that the positive average is due, at least in part, to migration. Immigrants from divergent populations tend
383 to generate positive LD by introducing novel alleles in combinations. The second observation is that the
384 geographically proximate outgroup IMP is the most frequent cause of IM polyphyly (when a single
385 outgroup is the cause; Supplemental Table 5). This suggests that migration is contributing to elevated $\pi$
386 and Var[$\pi$] and that geographically proximate species are contributing divergent haplotypes.
387
388 Despite these trends, many genomic windows exhibit a distinct pattern suggesting local adaptation. The
389 clearest signature of selection is a genomically localized reduction in nucleotide diversity coupled with
390 increased divergence of IM from other populations of *M. guttatus* (Figure 2) [54]. One of the most
391 striking observations from Table 1 is the elevated Dxy, depressed $\pi$ and Var[$\pi$] in monophyletic windows.
392 Reduced $\pi$ is a one signal of directional selection while elevated Dxy is indicative of locally beneficial
393 variants. Biologically, these statistics indicate the presence of locally favored variants present at high
394 frequency with low levels of within population variation – all pointing to a local selective advantage.
395

396 The variance in pairwise nucleotide differences, Var[π], is an informative statistic, particularly when
397 combined with gene-tree reconstructions (Table 1, Figure 3). Filtering genomic windows by Var[π]
398 demonstrated that low Var[π] regions may have topologies indicative of selective sweeps (extremely short
399 branch lengths), while high Var[π] regions exhibit multiple segregating haplotypes separated by longer
400 branch lengths. These patterns suggest the genomic regions are subject to distinct evolutionary forces.
401 For instance, balancing selection or soft sweeps may cause two haplotypes to be present at intermediate
402 frequency within a population and would result in a local increase the variance of pairwise π. Another
403 possible cause of elevated Var[π] would be sampling of a recently immigrated haplotype for which there
404 has been insufficient time for recombination to breakdown. Patterns of LD may be used to distinguish
405 between these two possibilities as large extended swaths of elevated Var[π] may be indicative of a recent
406 introgression events, while sharp localized Var[π] peaks may be due to extended periods of balancing
407 selection.
408
409 *Chromosomal Inversions*: We expect that selection effects will be most pronounced when recombination
410 is suppressed. Recent studies suggest reduced gene flow between annual and perennial 'ecotypes' of *M.*
411 *guttatus* within a chromosomal inversion on LG8 [3, 12]. Interestingly, QTL studies demonstrate that
412 genes contained within this inversion are important for life history type and flowering time [3]. Holeski
413 et al [25] recently mapped two additional putative inversions in a cross between IM and a perennial *M.*
414 *guttatus* genotype from Point Reyes, CA. These are located on linkage groups 5 and 10 respectively
415 (Figure 2). QTLs for traits related to reproductive isolation map to the LG8 inversion [3]; while the
416 phenotypic effects of the other two loci remain to be investigated. The present dataset reveals that
417 sequence divergence (Dxy) is significantly elevated within all three inversion regions relative to genome-
418 wide observations (Figure 2). The LG10 and LG8 inversions shows significantly lower overall nucleotide
419 diversity within IM, while the LG5 inversion is on not statistically different from genome-wide levels.
420

These results provide an interesting contrast to the literature on "genomic islands of speciation [55]." Genome scans in a number of systems have identified loci with elevated divergence among populations (populations sometimes described as species or nascent species) and these regions may harbor changes that reduce gene flow among populations. Cruickshank and Hahn recently reviewed the data from five systems (rabbits, mice, Heliconius butterflies, mosquitos, and flycatchers) and found that while relative divergence (measured by Fst) is higher in genomic islands, absolute divergence (measured by Dxy) is not (see Table 1 in [14]). They argue that this pattern, in conjunction with other observations, support the hypothesis that genomic islands are formed post-speciation. In other words, the relevant loci may not have actually reduced gene flow prior to speciation. The inversions in *M. guttatus* provide an interesting contrast to these examples for two reasons. First, they exhibit elevated divergence in both relative and absolute terms. Second, most of the outgroups included this survey (Supplemental Table 1) are described as divergent populations of one species (*M. guttatus*) rather than distinct/nascent species. Several self-fertilizing lineages (*M. nasutus, M. micranthus, M. platycalyx*) were included, but even for these, there is evidence of ongoing gene flow or introgression in the recent past [10, 56]. While gene flow seems to persist within this species complex, the populations that we contrast to IM are actually much more divergent (in terms of Dxy) than among taxa in four of the five systems reviewed by Cruickshank and Hahn (mice, butterflies, mosquitos, and flycatchers).

Patterns of genetic variation across the LG8 inversion within the IM population are surprising. First, the reduced intra-IM $\pi$ for the LG8 inversion is noteworthy given that other studies have reported higher inter-population annual $\pi$ than perennial $\pi$ [57]. Second, it is very clear that the fraction of IM monophyletic trees is elevated within the LG8 inversion. The ratio of monophyletic to polyphyletic windows genome-wide (outside inversion) is 0.14 (9,049 monophyletic to 62,988 polyphyletic windows) while within the LG8 inversion this ratio is 1.58 (1455 monophyletic windows to 925 polyphyletic windows). Third, Var[$\pi$] is reduced and Dxy is elevated relative to genome-wide patterns. Fourth, of IMP is the sole cause of IM polyphyly far more frequently within the LG8 inversion than genome-wide.

447 Within the LG8 inversion, the fraction of IMP broken IM monophyly is considerably higher than
448 genome-wide values. IMP is responsible for 409 of 925 (44%) polyphyletic windows, while outside the
449 inversion, IMP is solely responsible for 2923 of 62,988 (4.6%) polyphyletic windows.
450
451 These data suggest that alleles contained with the LG8 inversion are important in local adaptation of the
452 IM annuals. This is supported by the finding of reduced within IM $\pi$, reduced Var[$\pi$], dramatically
453 increased frequency of IM monophyletic windows, and increased Dxy within this region. These data
454 strongly suggest that and these regions are resistant to introgression due to selective pressures. Second,
455 these data provide evidence for gene exchange via recombination or gene conversion between IM and
456 IMP within LG8 region. This is supported by the fact that IMP is the sole cause of IM polyphyly far more
457 frequently within the LG8 inversion that genome-wide. Put simply, IM and IMP share a unique set of
458 sequences, not shared with outgroups, within the lg8 region more frequently than elsewhere in the
459 genome. This will be of potential interest given the observation that the LG8 inversion is the genomic
460 marker most closely association life-history of *M. guttatus* across its entire range.
461
462 Like the LG8 inversion, the LG10 inversions exhibits reduced intra-IM variation but increased
463 divergence. In contrast, intra-IM $\pi$ within the LG5 inversion is comparable to the genome-wide average.
464 The very elevated Dxy, high Var[$\pi$], but moderate $\pi$ within the LG5 inversion could be explained by
465 reduced gene flow with outgroup populations and balancing local selection. The proportion of
466 monophyletic windows within the inversion was substantially elevated when compared to outside LG5
467 inversion ratios. Within the LG5 inversion, 35 of 77 (45%) of windows were monophyletic. The
468 proportion of monophyletic windows within the LG10 inversion is even more elevated (700 of 1330
469 (52%) windows monophyletic). Interestingly, IMP is not solely responsible for breaking IM monophyly
470 in any of the 42 polyphyletic LG5 windows while IMP is the solely responsible for breaking IM
471 monophyly in 241 of 630 LG10 windows.

472 The variable patterns of ancestry across inverted regions is perhaps not too surprising. Many different
473 population/species of the M. guttatus complex are potential contributors to IM, and they may differ in the
474 whether they have the IM orientation for a particular inversion.
475
476 *Candidate genes for future study: Mimulus guttutus* annuals have evolved a "live-fast-die-young"
477 life history. To reproduce, they must take advantage of a narrow window of time between the spring
478 snowmelt and summer drought. Abiotic tolerances and timing of life processes (germination, flowering,
479 seed set, etc.) are particularly important in the success of IM annuals. Using the data assembled here, we
480 sought to provide the foundation to move from overall genomic signals to identifying genes potentially
481 important in these processes and possibly involved in local adaptation of the IM annual population. To
482 begin to identify genes that are potential candidates of strong selection within the IM population, we
483 calculated outlier residuals from a Dxy vs. $\pi$ contrast (for all monophyletic window) and selected the
484 2.5% most extreme outlier windows where IM $\pi$ is low for absolute divergence Dxy (Supplemental File
485 1; interval list of outlier windows). A total of 882 genes were located in these outlier windows, and they
486 have significantly lower Ks ($Ks_{outlier}$=0.014, $Ks_{Genome}$=0.035, $p<0.0001$; for genes with alignment length
487 >= 200 and p-value <=0.05, see methods) and lower non-synonymous diversity ($Ka_{outlier}$=0.002,
488 $Ka_{Genome}$=0.006, $p<0.0001$). Genes involved in the flowering pathway, germination timing, stress
489 responses, and trichome development are present in this outlier classes. While it is not possible to
490 mention all interesting outlier genes, several very intriguing candidates are worth mentioning for follow-
491 up functional work. *DELAY OF GERMINATION1* (*DOG*1), a gene involved in timing of germination and
492 flowering in *Arabidopsis thaliana* [58, 59], as well as several genes involved in flowering time in *A.*
493 *thaliana,* including *Short Vegetative Phase* (*SVP*) [60] and ATMBD9 [61], were contained within the
494 outlier windows. The fact that these genes regulate phenological transitions in *A. thaliana* and that
495 phenology is critical for IM fitness suggests that research following up on these candidate loci may move
496 us a step closer to a mechanistic understanding of local adaptation.
497

498

499 **Data Availability:** All sequence data generated here will be available on the Short Read Archive
500 following acceptance of the manuscript. SRA numbers will be listed here following acceptance.

501 **Acknowledgements:** We would like to thank Patrick Monnahan and Jenn Coughlan for providing
502 extensive comments on this manuscript. This work was supported by grants from the National Institutes
503 of Health to J.K. and J.W. (R01 GM073990) and the National Science Foundation to J.R.P. (NPGI-IOS-
504 1202778).

505 **Author Contributions:** JP, JW, and JK designed this experiment. JP made the libraries and directed
506 sequencing. JP and JK performed all genomic analyses and wrote the paper.

507

508 Cited


509 1. Fishman, L., A.J. Kelly, and J.H. Willis, *Minor quantitiative trait loci underlie floral*
510 *traits associated with mating system divergence in Mimulus.* Evolution, 2002. **56**(11):
511 p. 2138-2155.
512 2. Fenster, C.B. and K. Ritland, *Quantitative Genetics Of Mating System Divergence In the*
513 *Yellow Monkeyflower Species Complex.* Heredity, 1994. **73**(Pt4): p. 422-435.
514 3. Lowry, D.B. and J.H. Willis, *A Widespread Chromosomal Inversion Polymorphism*
515 *Contributes to a Major Life-History Transition, Local Adaptation, and Reproductive*
516 *Isolation.* PLoS Biol, 2010. **8**(9): p. e1000500.
517 4. Wu, C., et al., *Mimulus is an emerging model system for the integration of ecological*
518 *and genomic studies.* Heredity, 2008. **100**(2): p. 220-230.
519 5. Twyford, A.D., et al., *Genomic studies on the nature of species: adaptation and*
520 *speciation in Mimulus.* Molecular ecology, 2015. **24**(11): p. 2601-2609.
521 6. Hall, M.C. and J.H. Willis, *Divergent selection on flowering time contributes to local*
522 *adaptation in Mimulus guttatus populations.* Evolution, 2006. **60**(12): p. 2466-2477.
523 7. Kelly, J.K., *Deleterious mutations and the genetic variance of male fitness components*
524 *in Mimulus guttatus.* Genetics, 2003. **164**(3): p. 1071-1085.
525 8. Lowry, D.B., et al., *Genetic and physiological basis of adaptive salt tolerance*
526 *divergence between coastal and inland Mimulus guttatus.* New Phytologist, 2009.
527 **183**(3): p. 776-788.
528 9. Mojica, J.P., et al., *Spatially and temporally varying selection on intrapopulation*
529 *quantitative trait loci for a life history trade-off in Mimulus guttatus.* Molecular
530 ecology, 2012. **21**(15): p. 3718-3728.
531 10. Brandvain, Y., et al., *Speciation and Introgression between <italic>Mimulus*
532 *nasutus</italic> and <italic>Mimulus guttatus</italic>.* PLoS Genet, 2014. **10**(6): p.
533 e1004410.
534 11. Lowry, D.B., R.C. Rockwood, and J.H. Willis, *Ecological reproductive isolation of coast*
535 *and inland races of Mimulus guttatus.* Evolution, 2008. **62**(9): p. 2196-2214.
536 12. Twyford, A.D. and J. Friedman, *Adaptive divergence in the monkey flower Mimulus*
537 *guttatus is maintained by a chromosomal inversion.* Evolution, 2015. **69**(6): p. 1476-
538 1486.
539 13. Monnahan, P.J., J. Colicchio, and J.K. Kelly, *A genomic selection component analysis*
540 *characterizes migration-selection balance.* Evolution, 2015. **69**(7): p. 1713-1727.
541 14. Cruickshank, T.E. and M.W. Hahn, *Reanalysis suggests that genomic islands of*
542 *speciation are due to reduced diversity, not reduced gene flow.* Molecular Ecology,
543 2014. **23**(13): p. 3133-3157.
544 15. Charlesworth, B., M. Nordborg, and D. Charlesworth, *The effects of local selection,*
545 *balanced polymorphism and background selection on equilibrium patterns of genetic*
546 *diversity in subdivided populations.* Genetical research, 1997. **70**(02): p. 155-174.
547 16. Lewontin, R.C. and J. Krakauer, *Distribution of gene frequency as a test of the theory of*
548 *the selective neutrality of polymorphisms.* Genetics, 1973. **74**: p. 175-95.
549 17. Beaumont, M.A. and R.A. Nichols, *Evaluating loci for use in the genetic analysis of*
550 *population structure.* Proceedings Of the Royal Society Of London Series B-Biological
551 Sciences, 1996. **263**(1377): p. 1619-1626.

18. Storz, J.F. and J.K. Kelly, *Effects of Spatially Varying Selection on Nucleotide Diversity and Linkage Disequilibrium: Insights From Deer Mouse Globin Genes.* Genetics, 2008. **180**(1): p. 367-379.
19. Mojica, J.P. and J.K. Kelly, *Viability selection prior to trait expression is an essential component of natural selection.* Proceedings of the Royal Society B-Biological Sciences, 2010. **277**(1696): p. 2945-2950.
20. Willis, J.H., *Measures of phenotypic selection are biased by partial inbreeding.* Evolution, 1996. **50**: p. 1501-1511.
21. Kelly, A.J. and J.H. Willis, *Polymorphic microsatellite loci in Mimulus guttatus and related species.* Mol. Ecol., 1998. **7**: p. 769-774.
22. Kelly, J.K. and H.S. Arathi, *Inbreeding and the genetic variance of floral traits in Mimulus guttatus.* Heredity, 2003. **90**: p. 77-83.
23. Scoville, A., et al., *Contribution of chromosomal polymorphisms to the G-matrix of Mimulus guttatus.* New Phytologist, 2009. **183** p. 803–815.
24. Fishman, L. and A. Saunders, *Centromere-Associated Female Meiotic Drive Entails Male Fitness Costs in Monkeyflowers.* Science, 2008. **322**(5907): p. 1559-1562.
25. Holeski, L., et al., *A High-Resolution Genetic Map of Yellow Monkeyflower Identifies Chemical Defense QTLs and Recombination Rate Variation.* G3: Genes|Genomes|Genetics, 2014. **4**(5): p. 813-821.
26. Willis, J.H., *Inbreeding load, average dominance, and the mutation rate for mildly deleterious alleles in Mimulus guttatus.* Genetics, 1999. **153**: p. 1885-1898.
27. Flagel, L.E., J.H. Willis, and T.J. Vision, *The standing pool of genomic structural variation in a natural population of Mimulus guttatus.* Genome biology and evolution, 2014. **6**(1): p. 53-64.
28. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357-359.
29. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.
30. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nature genetics, 2011. **43**(5): p. 491-498.
31. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-2158.
32. Zhang, Z., et al., *KaKs_Calculator: calculating Ka and Ks through model selection and model averaging.* Genomics, proteomics & bioinformatics, 2006. **4**(4): p. 259-263.
33. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: the European molecular biology open software suite.* Trends in genetics, 2000. **16**(6): p. 276-277.
34. Vos, R., *Monophylizer, https://github.com/naturalis/monophylizer/tree/v1.0.1*. 2015.
35. Talevich, E., et al., *Bio. Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython.* BMC bioinformatics, 2012. **13**(1): p. 209.
36. Tajima, F., *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.* Genetics, 1989. **123**(3): p. 585-595.
37. Vilella, A.J., et al., *VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data.* Bioinformatics, 2005. **21**(11): p. 2791-2793.
38. Langley, C.H. and J.F. Crow, *The direction of linkage disequilibrium.* Genetics, 1974. **78**(3): p. 937-941.

39. Kelly, J.K., *A test of neutrality based on interlocus associations.* Genetics, 1997. **146**(3): p. 1197-1206.
40. Storz, J.F., et al., *Altitudinal variation at duplicated β-globin genes in deer mice: effects of selection, recombination, and gene conversion.* Genetics, 2012. **190**(1): p. 203-216.
41. LaMariposa, *GitHub: popgen_scripts, https://github.com/LaMariposa/popgen_scripts.*
42. Nei, M. and W.H. Li, *Mathematical model for studying genetic variation in terms of restriction endonucleases.* Proceedings of the National Academy of Sciences, 1979. **76**(10): p. 5269-5273.
43. Chan, A.H., P.A. Jenkins, and Y.S. Song, *Genome-wide fine-scale recombination rate variation in Drosophila melanogaster.* PLoS genetics, 2012. **8**(12): p. e1003090.
44. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-842.
45. Mackay, T.F., et al., *The Drosophila melanogaster genetic reference panel.* Nature, 2012. **482**(7384): p. 173-178.
46. Leffler, E.M., et al., *Revisiting an old riddle: what determines genetic diversity levels within species?* PLoS biology, 2012. **10**(9): p. e1001388.
47. Puzey, J. and M. Vallejo-Marín, *Genomics of invasion: diversity and selection in introduced populations of monkeyflowers (Mimulus guttatus).* Molecular ecology, 2014. **23**(18): p. 4472-4485.
48. Nolte, V., et al., *Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in Drosophila mauritiana.* Genome research, 2013. **23**(1): p. 99-110.
49. Fabian, D.K., et al., *Genome-wide patterns of latitudinal differentiation among populations of Drosophila melanogaster from North America.* Molecular ecology, 2012. **21**(19): p. 4748-4769.
50. Schmid, K.J., et al., *A multilocus sequence survey in Arabidopsis thaliana reveals a genome-wide departure from a neutral model of DNA sequence polymorphism.* Genetics, 2005. **169**(3): p. 1601-1615.
51. Fishman, L. and J.K. Kelly, *Centromere-associated meiotic drive and female fitness variation in Mimulus.* Evolution, 2015. **69**(5): p. 1208-1218.
52. Delph, L.F. and J.K. Kelly, *On the importance of balancing selection in plants.* New Phytologist, 2014. **201**(1): p. 45-56.
53. Sweigart, A., et al., *The distribution of individual inbreeding coefficients and pairwise relatedness in a population of Mimulus guttatus.* Heredity, 1999. **83**(5): p. 625-632.
54. Nosil, P., D.J. Funk, and D. ORTIZ-BARRIENTOS, *Divergent selection and heterogeneous genomic divergence.* Molecular ecology, 2009. **18**(3): p. 375-402.
55. Turner, T., M. Hahn, and S. Nuzhdin, *Genomic islands of speciation in Anopheles gambiae* PLoS Biology, 2005. **3**: p. e285.
56. Sweigart, A.L. and J.H. Willis, *Patterns of nucleotide diversity are affected by mating system and asymmetric introgression in two species of Mimulus.* Evolution, 2003. **57**(11): p. 2490-2506.
57. Oneal, E., et al., *Divergent population structure and climate associations of a chromosomal inversion polymorphism across the Mimulus guttatus species complex.* Molecular Ecology, 2014. **23**(11): p. 2844-2860.
58. Chiang, G.C., et al., *Pleiotropy in the wild: the dormancy gene DOG1 exerts cascading control on life cycles.* Evolution, 2013. **67**(3): p. 883-893.

643 59. Bentsink, L., et al., *Cloning of DOG1, a quantitative trait locus controlling seed*
644 *dormancy in Arabidopsis.* Proceedings of the National Academy of Sciences, 2006.
645 **103**(45): p. 17042-17047.
646 60. Lee, J.H., et al., *Regulation of temperature-responsive flowering by MADS-box*
647 *transcription factor repressors.* Science, 2013. **342**(6158): p. 628-632.
648 61. Peng, M., et al., *AtMBD9: a protein with a methyl-CpG-binding domain regulates*
649 *flowering time and shoot branching in Arabidopsis.* The Plant Journal, 2006. **46**(2): p.
650 282-296.
651

652

653 **Table 1.** Contrast of population genetic statistics between monophyletic and polyphyletic windows.

| | Monophyletic | Polyphyletic | p-value |
|---|---|---|---|
| $\pi$ | 0.012 | 0.014 | <0.0001 |
| Var[$\pi$] | 0.000045 | 0.000089 | <0.0001 |
| r | 0.084 | 0.135 | <0.0001 |
| S | 453 | 541 | <0.0001 |
| Tajima's D | 0.135 | 0.206 | <0.0001 |
| Dxy | 0.049 | 0.036 | <0.0001 |

654

655

656

**Figure 1.** Genome-wide distribution of r in IM annuals shows an excess of positive associations between rare alleles relative to neutral simulations.

**Figure 2.** Genomic distribution of absolute divergence (Dxy). Blue dots = monophyletic windows; red dots = polyphyletic windows. Bars on LG5, LG8, and LG10 denote approximate locations of chromosomal inversions.

**Figure 3.** Relationship of tree topology and distribution of pairwise π values within IM annuals. Highlighted red branches are IM annuals. Black bars are outgroups. 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.

Figure 1

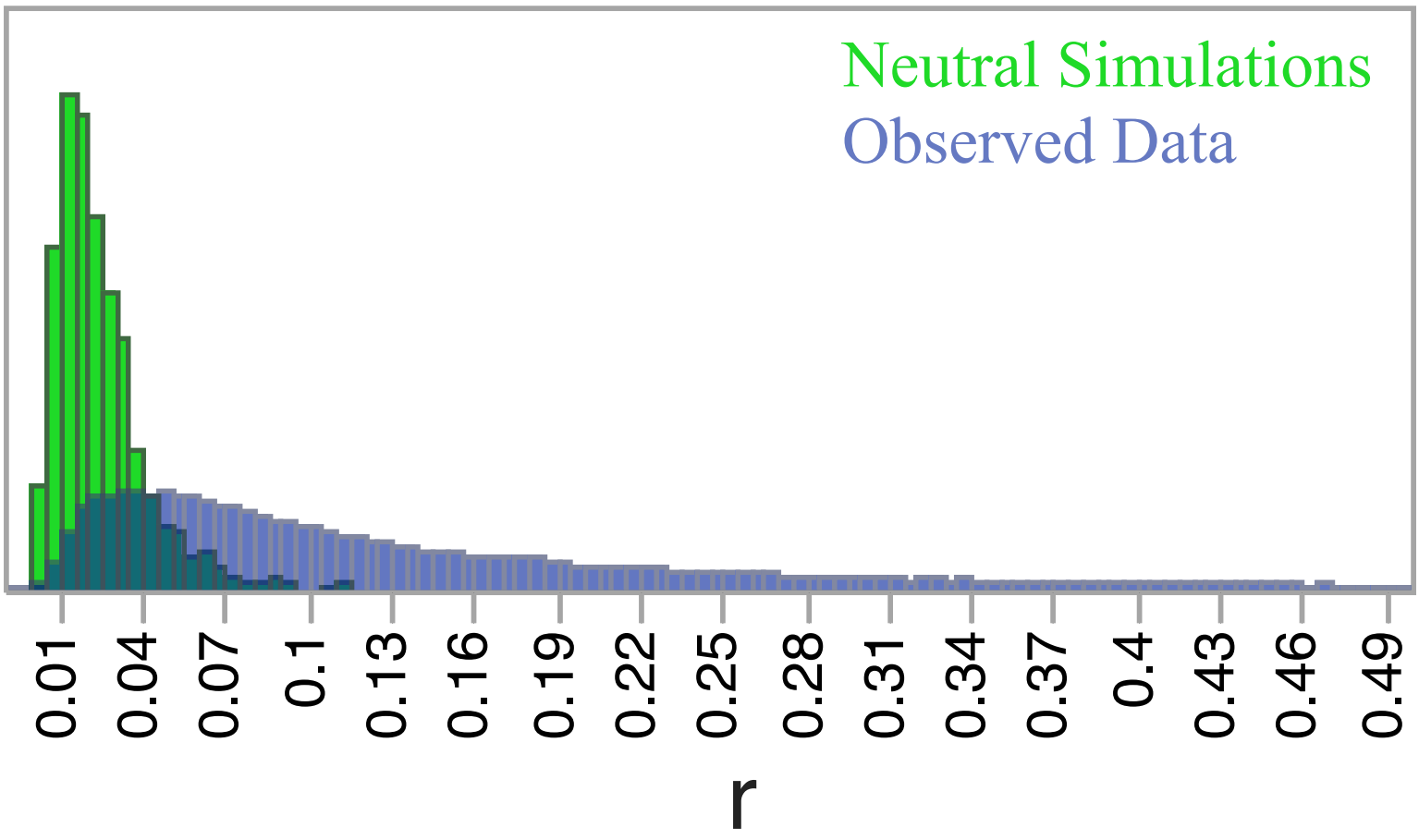

Figure 2

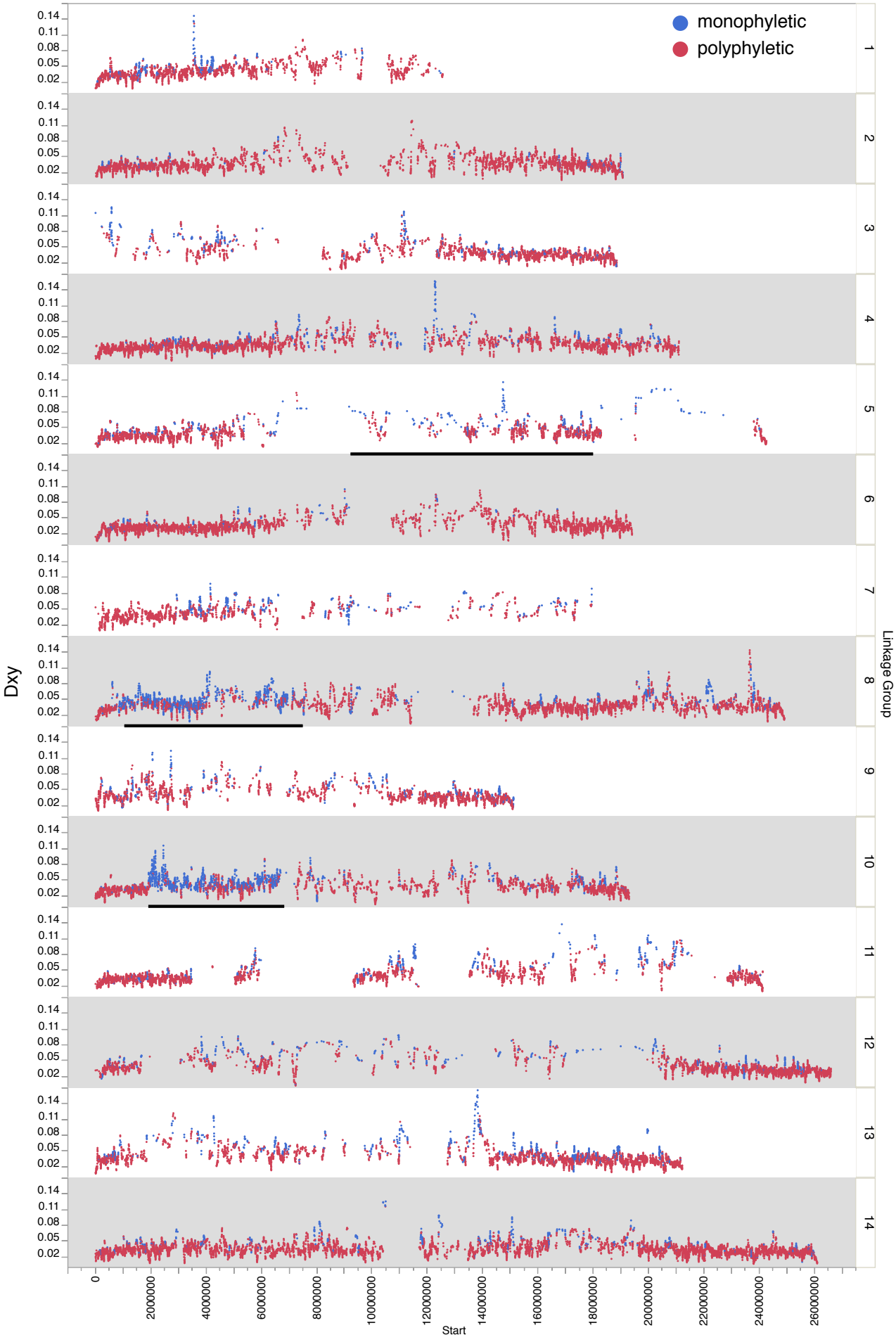

Figure 3