**Classification:** Biological Sciences: Evolution

# The tempo of linked selection: rapid emergence of a heterogeneous genomic landscape during a radiation of monkeyflowers

Sean Stankowski[1,2,§], Madeline A. Chase[1,3,§], Allison M. Fuiten[1], Peter L. Ralph[1], and Matthew A. Streisfeld[1*].

[1]Institute of Ecology and Evolution, 335 Pacific Hall 5289, University of Oregon, Eugene, OR 97405, USA
[2]Present address: Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK
[3]Present address: Department of Evolutionary Biology, Evolutionary Biology Centre (EBC), Uppsala University, Uppsala, Sweden
[§]These authors contributed equally to this work
*Corresponding author: mstreis@uoregon.edu

## ORCID ids:

SS: https://orcid.org/0000-0003-0472-9299; MAC: https://orcid.org/0000-0002-7916-3560; AMF: https://orcid.org/0000-0003-3926-5455 PLR: https://orcid.org/0000-0002-9459-6866; MAS: https://orcid.org/0000-0002-2660-8642

**Keywords:** genome assembly | lineage sorting | *Mimulus* | differentiation landscape

**Abstract—**What are the processes that shape patterns of genome-wide variation between emerging species? This question is central to our understanding of the origins of biodiversity and the fundamental principles governing molecular evolution. It is becoming clear that indirect selection on linked neutral variation (hereafter 'linked selection') plays a pervasive role in shaping heterogeneous patterns of genome-wide diversity and differentiation within and between species, but we do not know how these signatures of linked selection evolve over time. To fill this critical knowledge gap, we construct the first chromosome-level genome assembly for the bush monkeyflower, and use it to show that linked selection has been a primary architect of heterogeneous patterns of lineage sorting, differentiation, and nucleotide diversity across a recent radiation. By taking advantage of the range of divergence times between the different pairs of monkeyflower taxa, we also show how the signatures of linked selection evolve as populations diverge: linked selection occurring within lineages acts to conserve an ancestral pattern of diversity after a population split, while its joint action in separate lineages causes a common differentiation landscape to rapidly emerge between them. Together, our study demonstrates how pervasive linked selection shapes patterns of genome-wide variation within and between taxa, and provides critical insight into how its singiature evolves during the first 1.5 million years of divergence.

## Significance

What are the processes that shape patterns of genome-wide variation between emerging species? Because nucleotides are linked together on chromosomes, even neutral variants are impacted by selection on mutations that arise at neighboring sites. We show that this phenomenon, referred to as linked selection, was important in causing common patterns of differentiation to evolve between taxa during a radiation of monkeyflowers. This signature begins to emerge shortly after divergence begins, but it takes 1.5 million years to become pronounced. This result fills a critical gap in our knowledge about how genomes evolve, and it shows how linked selection shapes patterns of differentiation soon after a population split, which is critical to our understanding of divergence and speciation.

## Introduction

Since the first discoveries of abundant genetic variation in nature, evolutionary geneticists have sought to understand the processes that shape patterns of polymorphism and divergence within and between species (1-6). The neutral theory explained how mutation and drift could shape genetic variation (3, 7). Despite work suggesting the importance of non-neutral forces (5, 8-10), it has remained the default assumption of most molecular genetic analyses, partly because of a lack of concrete, alternative models. However, genome-wide studies have revealed heterogeneous patterns of genetic variation that are inconsistent with purely neutral forces (11-14). These genomic 'landscapes' can be important confounders for work in other fields, such as speciation research (15-17), and they provide intriguing clues in their own right into the ongoing evolutionary forces shaping our own genomes.

Heterogeneous genomic landscapes are increasingly understood to be formed due to the indirect effects of selection on linked neutral variation (hereafter, linked selection) (13). For example, variable patterns of genetic diversity ($\pi$) have now been observed across the genomes of a diverse range of plants and animals, and appear to have been shaped by variation in the intensity of linked selection across the genome (14, 18). This occurs because natural selection reduces the amount of genetic variation available to future generations at linked sites, similar to a reduction in local effective population size ($N_e$) (19-23). It is important to note that all forms of selection, whether acting on deleterious or beneficial mutations or on epistatic interactions, have linked effects. In theoretical models of recurrent linked selection, its local intensity is determined by the density of targets of selection relative to the recombination rate, such that larger reductions in diversity occur in genomic regions with more frequent selection and less recombination (19-22).

Compared to patterns of within-species diversity, we know relatively little about how linked selection shapes patterns of genome-wide differentiation between emerging species (15) Although rates of differentiation and lineage sorting should be accelerated in genomic regions that have experienced long-term reductions in diversity (24-26), we do not know how long it takes for linked selection to generate heterogeneous patterns of between-species variation (17). Unlike patterns of diversity ($\pi$), which are inherited from an ancestral population and maintained in diverging taxa by ongoing linked selection, a heterogeneous pattern of differentiation ($F_{ST}$) should emerge gradually owing to the accumulating effects of lineage-specific linked selection following a population split.

However, the temporal dynamics of genomic landscape evolution, which have been outlined in a verbal model by Burri (16), have never been explicitly tested. Therefore, empirical studies are needed to fill this critical gap in our knowledge of the processes that shape patterns of genome-wide variation between emerging species.

In this paper, we study the temporal signatures of linked selection using taxa from the bush monkeyflower radiation (Figure 1). This recent radiation of perennial shrubs is distributed throughout California and consists of seven subspecies of *Mimulus aurantiacus*, one with two ecotypes (27). Together with their sister species *M. clevelandii,* they span a variety of divergence times, ranging from locally adapted ecotypes to species separated by ~1.5 million years (28). The plants inhabit multiple environments, including temperate coastal regions, mountain ranges, semi-arid habitats, and offshore islands (29). Most of the taxa are geographically isolated from one another, though some have parapatric distributions and hybridize in narrow regions where their distributions overlap (30-36). Recent phylogenetic studies have confirmed the monophyly of the radiation and revealed the basic relationships among its taxa (27, 33), making it an excellent system to study how the signatures of linked selection evolve over time.

Using whole-genome sequencing and the first chromosome-level reference assembly for the bush monkeyflower, we reveal heterogeneous patterns of lineage sorting, diversity, and differentiation that have been shaped by variation in the intensity of linked selection acting across the genome. By using the different taxon pairs as points along a divergence continuum, we then show how these signatures of linked selection evolve over the first 1.5 million years of divergence. These results have important implications for our understanding of the origins of biodiversity, speciation, and the basic principles governing molecular evolution.

## Results and Discussion

*A chromosome-level genome assembly, map, and annotation for the bush monkeyflower*

To facilitate the analysis of genome-wide variation in this group, we constructed the first chromosome-level reference genome for the bush monkeyflower using a combination of long-read Single Molecule Real Time (SMRT) sequencing reads (PacBio), overlapping and mate-pair short-reads (Illumina), and a high-density genetic map (7,589 segregating markers across 10 linkage groups; Fig. S1; Table S1). Contig building and scaffolding yielded 1,547 scaffolds, with an N50 size of 1,578 kbp, and a total length of 207 Mbp. The high-density map allowed us to anchor and orient 94% of the assembled genome onto 10 linkage groups, which is the number of chromosomes inferred from karyotypic analyses in all subspecies of *M. aurantiacus* and *M. clevelandii* (37). Analysis of assembly completeness based on conserved gene space (38) revealed that 93% of 1440 universal single copy orthologous genes were completely assembled (Table S2). Subsequent annotation yielded 23,018 predicted genes.

*Variation in the extent of lineage sorting across the genome*

As a first step toward understanding the processes that have shaped patterns of genome-wide variation during this radiation, we inferred phylogenetic relationships among its taxa. Rapid diversification is a hallmark of radiations and can result in extensive phylogenetic discordance between genomic regions due to incomplete lineage sorting (ILS) (39-42). To do this, we sequenced 37 whole genomes from the seven

subspecies and two ecotypes of *Mimulus aurantiacus* ($n$ = 4-5 per taxon) and its sister species *M. clevelandii* ($n$ = 3) (Fig. S2; Table S3). Close sequence similarity allowed us to align reads from all samples to the reference assembly with high confidence (average 91.7% reads aligned; Table S3). After mapping, we identified 13.2 million variable sites that were used in subsequent analyses (average sequencing depth of 21x per individual, Table S3). Relationships were inferred among the nine taxa using maximum-likelihood (ML) phylogenetic analysis (43) based on three different datasets: whole-genome concatenation and 500 kb and 100 kb non-overlapping genomic windows.

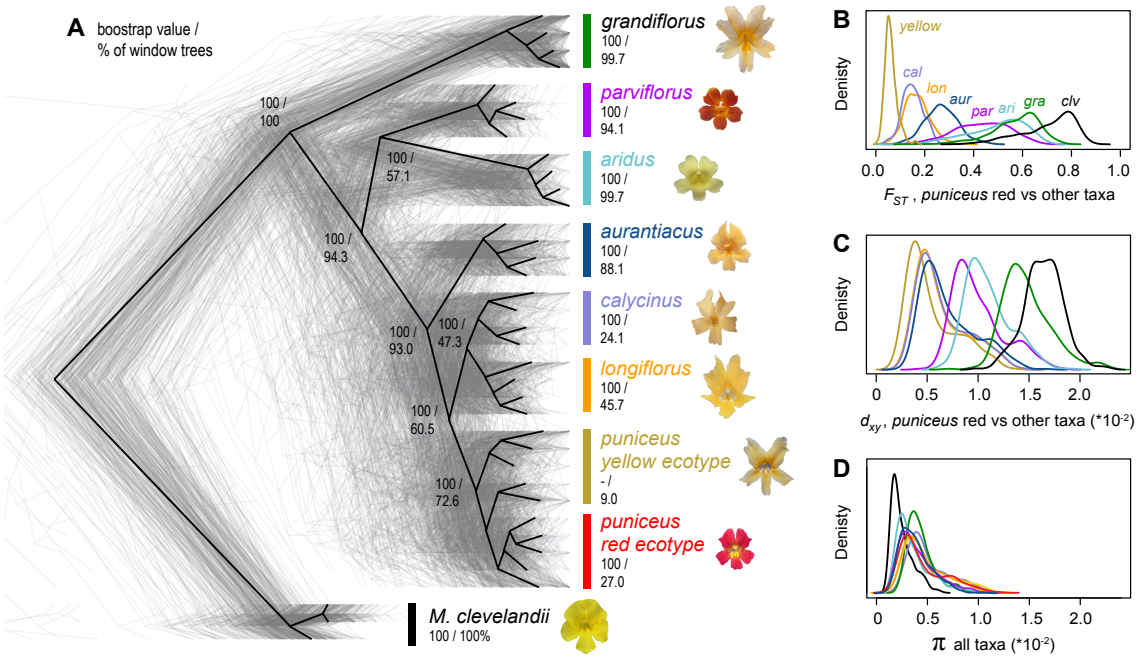



**Figure 1. Evolutionary relationships and patterns of genome-wide variation across the radiation**. A) The black tree was constructed from a concatenated alignment of genome-wide SNPs and is rooted using *M. clevelandii*. The 387 gray trees were constructed from 500 kb genomic windows. The first number associated with each node or taxon is the bootstrap support for that clade in the whole genome tree, and the second number is the percentage of window-based trees in which that clade is present. B) Levels of differentiation ($F_{ST}$), C) divergence ($d_{xy}$), and D) diversity ($\pi$) within and among taxa based on the same 500 kb windows. For simplicity, $F_{ST}$ and $d_{xy}$ are shown only for comparisons with the red ecotype of subspecies *puniceus*.

The tree topology obtained from the whole genome (concatenated) dataset (Fig. 1) confirmed the same phylogenetic relationships as previous analyses based on reduced-representation sequencing and five methods of phylogenetic reconstruction (27, 33), and were supported by patterns of clustering from principal components analysis (Fig. S3). All seven subspecies formed monophyletic groups with 100% bootstrap support. Relationships within subspecies *puniceus* were more complex, as the red ecotype formed a monophyletic sub-clade within the paraphyletic yellow ecotype, reflecting the recent origin of red flowers from a yellow-flowered ancestor (33).

164 Although the whole genome phylogeny provides a well-supported summary of the
165 relationships among the taxa, window-based analyses revealed extensive phylogenetic
166 discordance at a finer genomic scale (Fig 1A). Despite four of the subspecies forming
167 separate monophyletic groups in nearly all of the 387 window-based trees (500 kb scale;
168 *grandiflorus* 99.7%, *aridus* 99.7%, *parviflorus* 94.1%, and *aurantiacus* 88.1%), only 22
169 (6%) trees showed the same taxon branching order as the whole-genome tree. While
170 some of this discordance could be generated by gene flow after divergence (33, 42, 44,
171 45), our data indicate that the majority is due to incomplete lineage sorting (ILS).
172 Specifically, higher levels of discordance were observed at nodes that were separated by
173 shorter internode lengths ($r^2$ = 0.94, $p$ < 0.001; Fig. S4). For example, even though
174 subspecies *puniceus* was monophyletic in the majority of trees (72.6%), individuals from
175 the red and yellow ecotypes only formed monophyletic groups in 27% and 9% of the
176 trees, respectively. Similarly, the closely related subspecies *longiflorus* and *calycinus*
177 were monophyletic in fewer trees than the other subspecies (45.7% and 24.1%,
178 respectively). Thus, it appears that ILS is the primary source of phylogenetic conflict,
179 especially near rapid divergence events, as predicted by theory (41, 46) and shown in
180 other diverse radiations (42, 47).
181 Next, we examined how the pattern of lineage sorting varied across the bush
182 monkeyflower genome. Tree discordance resulting from the stochastic effects of neutral
183 demography should be distributed uniformly across the genome (39). To test this
184 prediction, we computed the correlation between the distance matrix generated from each
185 window-based tree and the whole-genome tree, with a stronger relationship indicating
186 that they are more similar (i.e., less ILS). Plotting this tree concordance score across the
187 10 linkage groups revealed a striking pattern (Fig. 2A; Fig. S5 for plots along each
188 chromosome and S6 for results for 100 kb windows). Rather than being randomly distr-
189 ibuted, trees with low concordance scores tended to cluster together in relatively narrow
190 regions of all 10 chromosomes (Fig. 2A; autocorrelation analysis permutation tests $p$ =
191 0.023 - 0.001; Fig. S7). This non-random pattern suggests that local rates of lineage
192 sorting are determined by differences in the nature and/or strength of selection acting
193 across the bush monkeyflower genome.
194
195 *Patterns of genome-wide variation and lineage sorting have been shaped by recurrent*
196 *linked selection*
197 To gain deeper insight into the evolutionary processes that have shaped patterns
198 of genome-wide variation, we used summary statistics in 500 kb windows to quantify
199 patterns of differentiation ($F_{ST}$), divergence ($d_{xy}$), and diversity ($\pi$) among and within
200 these taxa. The variation in $F_{ST}$ among all 36 pairs of taxa highlights the continuous
201 nature of differentiation across the group (Fig. 1B; Fig. S8), with mean window-based
202 estimates ranging from 0.06 (red vs. yellow ecotypes of *puniceus*) to more than 0.70.
203 Distributions of absolute divergence ($d_{xy}$) show a similar pattern (Fig. 1C), with mean
204 values ranging from 0.54% (red vs. yellow ecotypes) to 1.6% (yellow ecotype vs. *M.*
205 *clevelandii*). More strikingly, the broad distributions of window-based estimates revealed
206 heterogeneity in levels of differentiation and divergence among genomic regions.
207 Window-based estimates of nucleotide diversity also vary markedly ($\pi$; Fig. 1D), ranging
208 from 0.09% to 1.26%, even though mean estimates were very similar among the ingroup
209 taxa (0.37% to 0.53%) and were only slightly lower in *M. clevelandii* (0.26%).

As with tree concordance, these summary statistics showed non-random patterns of variation across broad regions of the genome ($p < 0.005$; Fig. 2; Fig. S5 Fig. S6; Fig. S7). To account for the large magnitude of variation in these statistics across all nine taxa (for $\pi$) or among the 36 pairs of taxa (for $d_{xy}$ and $F_{ST}$), we normalized the window-based estimates using *z*-transformation and plotted them across the genome (Fig. S5). After noting that the genome-wide patterns for each statistic were qualitatively similar among all comparisons, we used principal components analysis to quantify their similarity and extract a single variable (PC1) that summarized this common pattern (Fig. S5). These analyses confirmed that patterns of genome-wide variation were highly correlated across this group of taxa. Indeed, PC1 explained 65.9% of the variation in $F_{ST}$ across the 36 pairwise comparisons. Further, all comparisons loaded positively onto PC1 (mean loading = 0.78 s.d. 0.18; Table S4 for all loadings), indicating that peaks and troughs of $F_{ST}$ tended to occur in the same genomic regions across all comparisons. Similarly, patterns of genome-wide divergence ($d_{xy}$) and diversity ($\pi$) were highly correlated across comparisons, with PC1 explaining 69.5% and 84.7% of the variation among the window-based estimates, respectively. Again, all taxa (for $\pi$) and taxon comparisons (for $d_{xy}$) loaded positively onto the first principal component (mean loading for $d_{xy}$ = 0.78 s.d. 0.18; for $\pi$, 0.91 s.d. 0.07). PC1 therefore provides a summary of the original landscapes, and is effectively the same as taking the mean window-based scores for each statistic ($r^2$ between PC1 and mean scores > 0.995 for all three statistics).

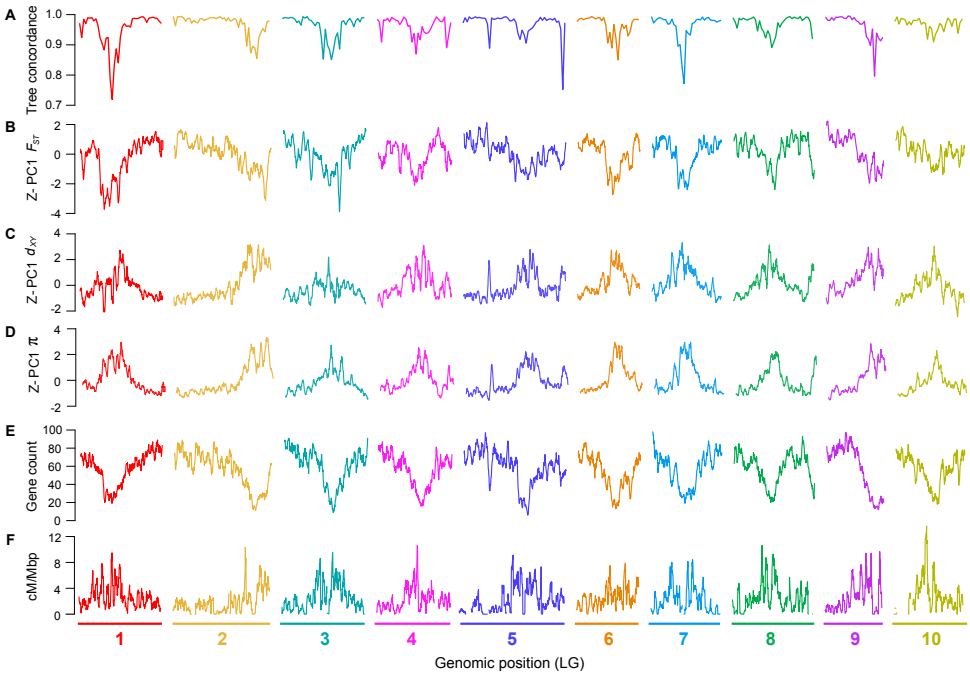



**Figure 2. Common differentiation and diversity landscapes mirror variation in the local properties of the genome.** A) Tree concordance scores for 500 kb non-overlapping genomic windows plotted across the 10 bush monkeyflower chromosomes. B – D) Plots of the *z*-transformed first principal component (PC1) for $F_{ST}$, $d_{xy}$, and $\pi$ in overlapping 500 kb windows (step size = 50 kb). PC1 explains 66%, 70%, and 85% of the variation in $F_{ST}$, $d_{xy}$ and $\pi$, respectively. E – F) Gene count and recombination rate (cM/Mbp) in overlapping 500 kb windows.

Observing similar differentiation, diversity, and divergence 'landscapes' among these taxa suggests that a common mechanism has been responsible for shaping patterns of genome-wide variation across the radiation. Recent studies have observed correlated genomic landscapes among related taxa, concluding that they were generated by a shared pattern of heterogeneous linked selection (48-51). Indeed, if a region experiences a high level of linked selection across the phylogenetic tree, then it will have both lower diversity ($\pi$) within species and lower divergence ($d_{xy}$) between species, because divergence is determined in part by levels of diversity in the common ancestor (15, 16, 22, 24). In agreement with this prediction, we observed a strong positive correlation between PC1 $d_{xy}$ and PC1 $\pi$ ($r$ = 0.84), indicating that regions of the genome with lower diversity tended to be less diverged in all taxa (Fig. 3, Fig. S9 for scatterplots and Fig. S10 for results at 100 kb scale). Regions with reduced diversity also tended to show higher differentiation ($F_{ST}$) ($r$ = -0.84) and higher levels of tree concordance ($r$ = -0.69), both of which are predicted by models of linked selection (52).

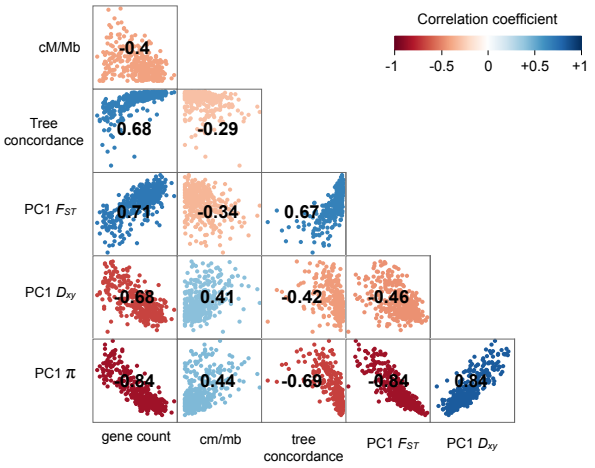



**Figure 3. Correlations reveal the impact of heterogeneous linked selection across the genome.** Matrix of pairwise correlation coefficients between PC1 $F_{ST}$, PC1 $d_{xy}$, PC1 $\pi$, tree concordance, gene density, and recombination rate. The heat map and the shape of the ellipse indicate the strength of the correlation and its sign. All correlations are statistically significant at $p < 0.001$. Detailed scatterplots for each relationship can be found in Fig. S9. See Fig. S10 for a correlation matrix for 100 kb windows.

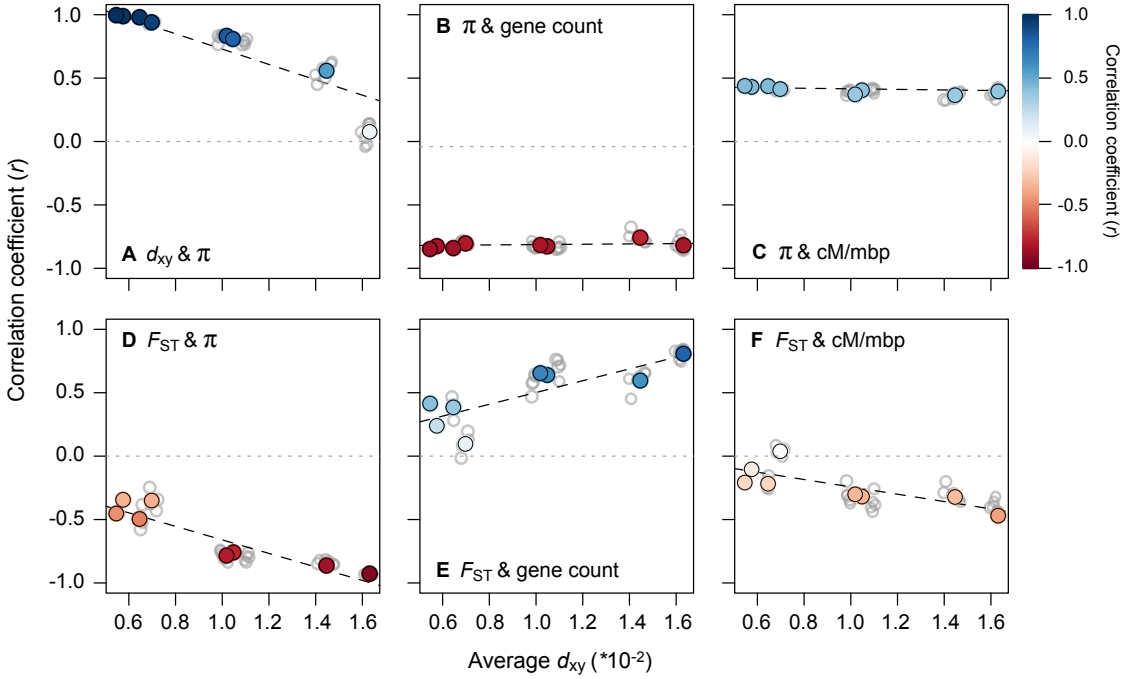
We next identified factors that cause variation in the intensity of linked selection across the bush monkeyflower genome. Its intensity is determined by the local density of targets of selection relative to the recombination rate (19-22). Specifically, a higher intensity of linked selection is predicted in regions of the genome that are enriched for functional elements, because mutations are more likely to have fitness consequences if they arise in areas that are gene rich. The local recombination rate modulates this effect, because regions unlinked to functional sites evolve independently of them.

To test these predictions, we used our annotated genome and genetic map to calculate the number of protein coding genes and the average recombination rate (cM/Mbp) in each 500 kb window (Fig. 2E-F; Fig. S5; Fig. S6). There was a strong negative correlation between gene count and recombination rate ($r$ = -0.40), leading to large variation in the predicted strength of linked selection across the genome. In addition, we observed strong correlations between PC1 $\pi$ and both gene count ($r$ = -0.84) and recombination rate ($r$ = 0.44; Fig. 3; Figs. S9 & S10), both of which indicate that variation in the intensity of linked selection has shaped common patterns of diversity across the genome.

Despite only having a direct estimate of gene density and recombination rate variation from one subspecies (*puniceus*), the presence of a common diversity landscape

309 implies that the genomic distribution of these features has been conserved in all taxa after
310 being inherited from their common ancestor. This scenario is consistent with the recent
311 shared history of the group, and explains how a common pattern of heterogeneous linked
312 selection has become shared among them.
313
314 *The signature of linked selection becomes stronger with increasing divergence time*
315 Our analyses indicate that linked selection has shaped common patterns of
316 diversity and differentiation across this radiation. By using the different levels of
317 divergence between pairs of taxa, we next test predictions about how these signatures
318 should evolve over time (16). When a population first splits, levels of diversity ($\pi$) and
319 divergence ($d_{xy}$) are equal, so $F_{ST}$ will be zero across the genome (Fig. S11 for cartoon
320 explanation assuming a simple model of allopatric divergence, no spatial structure, and
321 large $N_e$). As divergence proceeds, differentiation increases, but due to variation in the
322 intensity of linked selection across the genome, certain regions become differentiated
323 before others. During the early stages of divergence, when lineage-specific linked
324 selection has had a minor impact on the genome, patterns of differentiation should only
325 weakly mirror the footprint of historical linked selection. However, as the divergence
326 time increases, the cumulative effects of linked selection should strengthen the
327 relationships between $F_{ST}$ and $\pi$, gene density, and recombination rate. In contrast, the
328 strength of the correlations between $\pi$ and gene density and recombination rate should
329 remain similar over time, because the shape of the diversity landscape is preserved by
330 recurrent linked selection despite new mutations arising in each lineage.
331 To test these predictions, we determined if the strength of the relationships
332 between these statistics varied with the level of divergence between taxa. As expected for
333 a pair of taxa that recently split, the correlation between $\pi$ and $d_{xy}$ is almost perfect
334 between the least divergent pairs of taxa ($r \sim 1$), but the correlation decays over time as
335 ancestral variants fix and new mutations increase $d_{xy}$ (Fig. 4A). Remarkably, however,
336 the strong correlations between $\pi$ and gene density ($r \sim 0.8$) and $\pi$ and recombination rate
337 ($r \sim 0.4$) barely change with increasing divergence time, as expected if linked selection
338 continues to act on the same regions in each taxon (Fig. 4B-C). By contrast, the
339 relationships between $F_{ST}$ and levels of diversity, gene count, and recombination rate all
340 become stronger with increasing divergence time (Fig. 4D-F), revealing the build up of
341 heterogeneous differentiation due to the accumulating impact of recurrent linked
342 selection.
343 In addition to showing that the footprint of linked selection is dynamic, the
344 sequence of divergence times provides novel insight into when linked selection begins to
345 shape patterns of differentiation, and how long it takes for this signature to develop (Fig.
346 5). In population pairs with the most recent divergence times ($d_{xy}$ = 0.5% – 0.7%), linked
347 selection's effects on the differentiation landscape are already apparent, as genome-wide
348 patterns of $F_{ST}$ are moderately correlated with variation in $\pi$, gene count, and
349 recombination rate ($r \sim 0.4$, Fig. 4D-F). This is true even for the parapatric red and
350 yellow ecotypes of subspecies *puniceus* ($d_{xy}$ = 0.5%), which diverged recently with
351 ongoing gene flow (31, 32). As divergence continues, the effects of linked selection
352 become even more pronounced. In the most divergent comparisons ($d_{xy}$ = 1.5%), the
353 diversity and differentiation landscapes almost perfectly mirror one another ($r$ = -0.94;
354 Fig. 5, Fig. S12). The build-up of such strong correlations in just 1.5 million years is a

355 testament to the power of linked selection to shape genome-wide patterns of variation,
356 both within and between taxa.




357
358 **Figure 4. Time-course analysis reveals static and dynamic signatures of recurrent linked**
359 **selection.** Correlations between variables (500 kb windows) for all 36 taxonomic comparisons
360 (gray dots) plotted against the average $d_{xy}$ as a measure of divergence time. The top row shows
361 how the relationships between π (each window averaged across a pair of taxa) and (A) $d_{xy}$, (B)
362 gene count, and (C) recombination rate vary with increasing divergence time. The bottom row
363 (D-E) shows the same relationships, but with $F_{ST}$. The regressions (dashed lines) in each plot are
364 fitted to the eight independent contrasts (colored points) obtained using a phylogenetic correction.
365 The color gradient shows the strength of the correlation. Details for each regression can be found
366 in Table S5.
367
368 *Conclusions and implications*
369 Facilitated by the first chromosome-level genome assembly for the bush
370 monkeyflower, we show that linked selection has been a primary architect of the common
371 patterns of diversity, differentiation, and lineage sorting across this recent radiation.
372 Genome-wide variation in the intensity of linked selection is conserved among these taxa
373 and is determined by the distribution of functional elements and variation in the local
374 recombination rate. By taking advantage of the range of divergence times between the
375 different pairs of monkeyflower taxa, we provide the first empirical picture of how the
376 signatures of linked selection emerge over time: linked selection occurring within
377 lineages acts to preserve an ancestral pattern of diversity after a population split, while its
378 joint action in separate lineages causes a common differentiation landscape to emerge
379 between them.
380

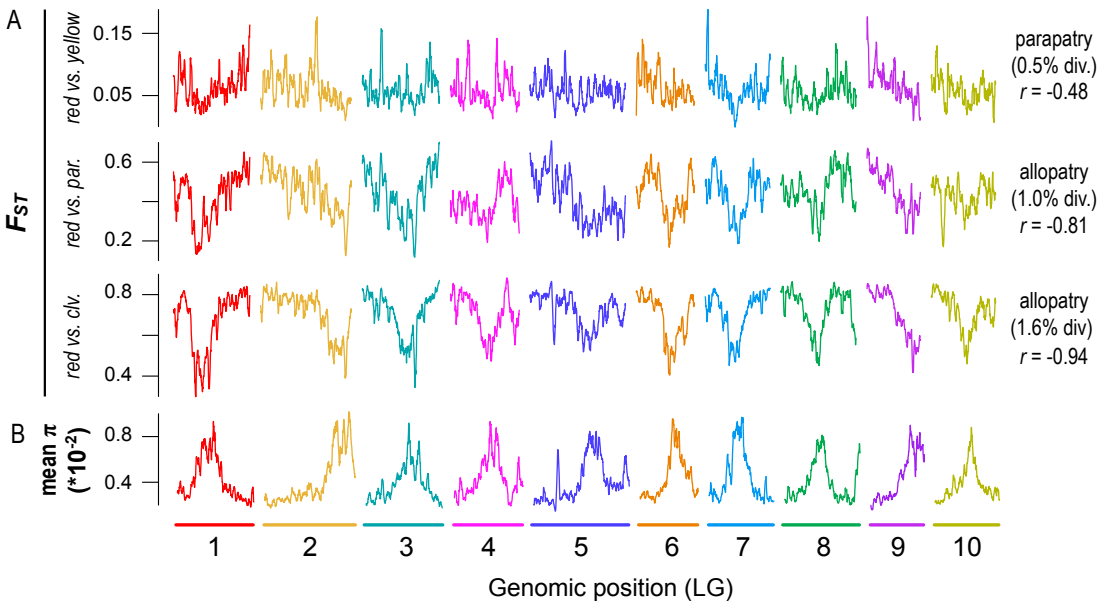



**Figure 5. Emergence of a heterogeneous differentiation landscape across 1.5 million years of divergence.** A) Plots of $F_{ST}$ (500 kb windows) across the genome for pairs of taxa at early (red vs. yellow), intermediate (red vs. *parviflorus*) and late stages (red vs. *M. clevelandii*) of divergence. B) Average nucleotide diversity (for red, *yellow*, *par*., and *clv*.) across the genome in 500 kb windows. The geographic distribution (parapatric or allopatric), percent sequence divergence ($d_{xy}$*$10^{-2}$) and correlation between $F_{ST}$ and mean π are provided next to each taxon pair.

In addition to providing a dynamic picture of how the genomic landscape has evolved over the first 1.5 millions years of divergence, our study has important implications for the fields of molecular evolution and speciation. For example, even though the impact of linked selection might be expected to vary across the tree of life due to factors like differences in genome size, ploidy, mutation rate, recombination rate, and effective population size (53), our findings support previous studies indicating that little of the genome evolves free of its effects (6, 14, 18). This suggests that genome-wide patterns of diversity, differentiation, and lineage sorting cannot be understood without taking the effects of linked selection into account.

Our work also has implications for interpreting the genomic landscape in light of the speciation process. Although it was initially thought that peaks of differentiation would correspond to genomic regions underlying barriers to gene flow between emerging species, it is now clear that differentiation landscapes are also shaped by widespread selection that is unrelated to speciation (17, 49). For example, recent studies have suggested that widespread background selection is likely responsible for generating common differentiation landscapes across groups of closely related taxa due to the conservation of genomic features among them (15, 16, 54). Although background selection would be a straightforward explanation for the correlated patterns of differentiation observed across this radiation, a recent simulation study (55) suggests that it should not impact the differentiation landscape over most of the range of divergence times that we examined. This is especially true for the well-studied red and yellow

411 ecotypes of subspecies *puniceus*, as any effect of background selection should be
412 nullified when divergence occurs with gene flow (55). Therefore, unless these
413 simulations failed to capture some important aspect of our study system, other forms of
414 selection that are relevant to speciation may contribute to the common signature of linked
415 selection that we, and others, have seen. Although further work is clearly needed to
416 understand the causes of linked selection, our study shows that characterizing its
417 signatures is a critical step in understanding the processes that shape genetic variation
418 within and between populations and species.
419

420 **Materials and Methods**

421 *Genome assembly, high-density linkage map, and annotation*

422 We used a combination of short-read Illumina and long-read Single Molecule,
423 Real Time (SMRT) sequencing to assemble the genome of a single individual from the
424 red ecotype of *M. aurantiacus* subspecies *puniceus* (Table S2 for sample collection
425 location). To assemble resulting scaffolds into pseudomolecules, we generated a high-
426 density linkage map from an outbred $F_2$ mapping population (Table S2 for sample
427 collection locations). Restriction-site associated DNA sequencing (RADseq) was used to
428 genotype parents and 269 offspring. Map construction was performed using *Lep-MAP2*
429 (56). After integrating the assembly and genetic map, we made corrections to the map
430 order using the physical position of markers within the assembled scaffolds. Genome
431 annotation was conducted using the MAKER pipeline (57). See supplementary methods
432 for more details.
433

434 *Genome re-sequencing and variant calling*

435 DNA was extracted from 37 individuals (Table S2), and Illumina sequencing was
436 performed using paired-end 150 bp reads. Raw reads were aligned to the reference
437 assembly, and variant calling was executed with the GATK pipeline (58) using the
438 UnifiedGenotyper tool. See supplementary methods for more details.
439

440 *Phylogenetic Analyses*

441 We used *RAxML* v8 to construct genome-wide phylogenies from a concatenated
442 alignment of all variable sites and from genomic windows. We tested for a relationship
443 between node concordance (the number of 500 kb window-based trees that recovered a
444 given node from the genome-wide tree) and internode length using the internode
445 distances from the genome-wide analysis. Tree concordance scores were generated from
446 the correlation between the distance matrix from each window-based tree and the whole-
447 genome tree. Autocorrelation coefficients for tree concordance scores were calculated in
448 *R* using custom scripts, and their significance was tested from 1000 random permutations
449 of the genome-wide data. See supplementary methods for more details.
450

451 *Population genomic analyses*

452 Estimates of nucleotide diversity ($\pi$), differentiation ($F_{ST}$), and divergence ($d_{XY}$)
453 for 100 kb and 500 kb windows were calculated using Python scripts downloaded from
454 https://github.com/simonhmartin/genomics_general. Principal components analysis was
455 used to summarize patterns of variation in these statistics across all taxa (for $\pi$) and taxon
456 comparisons (for $F_{ST}$ and $d_{XY}$). Average recombination rate (cM/Mbp) for each window

was estimated as the average value across three or more adjacent pairs of mapped markers in each genomic window. Gene density was estimated as the number of predicted genes in each window. We used linear regression to test if the strength of the correlations between different statistics changed with the level of divergence using a set of eight statistically independent contrasts (59, 60). See supplementary methods for details.

## Data Accessibility

Raw sequencing reads used for the genome assembly, linkage map construction, and genome resequencing and are available on the Short-Read Archive (SRA) under the bioproject ID xxx. The genetic map, annotation, reference genome sequence, and VCF file have been deposited on DRYAD.

## Acknowledgments

We thank Bill Cresko and Thomas Nelson for advice and stimulating discussion. John Willis performed the Illumina mate-pair library preps used for the genome assembly. Julian Catchen, Clay Small, Susan Bassham, and Janna Fierst provided technical advice. Doug Turnbull and Maggie Weitzman conducted the Illumina sequencing at the University of Oregon Genomics Core facility. Thomas Nelson, Martin Garlovsky, Roger Butlin, Anja Westram and Jeff Ross-Ibarra provided comments on an earlier version of this manuscript. We also thank people who contributed to insightful discussions on twitter. Funding was provided by the National Science Foundation grant DEB-1258199 to MS and the Sloan Foundation to PR.

## References

1. Lewontin RC (1974) *The Genetic Basis of Evolutionary Change* (Columbia University Press, New York).
2. Lewontin RC & Hubby JL (1966) A moleuclar approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* 54(2):595-&.
3. Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217(5129):624-626.
4. Kimura M (1986) DNA and the neutral theory. *Philos Trans Roy Soc Lond B, Biol Sci* 312(1154):343-354.
5. Gillespie JH (1994) *The Causes of Molecular Evolution* (Oxford University Press On Demand).
6. Casillas S & Barbadilla A (2017) Molecular population genetics. *Genetics* 205(3):1003-1035.
7. Ohta T (1973) Slightly deleterious mutant substitutions in evolution. *Nature* 246(5428):96-98.
8. Langley CH & Fitch WM (1974) An examination of the constancy of the rate of molecular evolution. *J Mol Evol* 3(3):161-177.
9. Kreitman M (1996) The neutral theory is dead. Long live the neutral theory. *BioEssays* 18(8):678-683.

503 10. Hahn MW (2008) Toward a selection theory of molecular evolution. *Evolution*
504 62(2):255-265.
505 11. Langley CH*, et al.* (2012) Genomic variation in natural populations of Drosophila
506 melanogaster. *Genetics* 192(2):533-598.
507 12. Begun DJ*, et al.* (2007) Population genomics: Whole-genome analysis of
508 polymorphism and divergence in Drosophila simulans. *Plos Biol* 5(11):2534-
509 2559.
510 13. Charlesworth B & Charlesworth D (2018) Neutral variation in the context of
511 selection. *Mol Biol Evol*.
512 14. Kern AD & Hahn MW (2018) The neutral theory in light of natural selection. *Mol*
513 *Biol Evol*.
514 15. Burri R (2017) Dissecting differentiation landscapes: a linked selection's
515 perspective. *J Evol Biol* 30(8):1501-1505.
516 16. Burri R (2017) Interpreting differentiation landscapes in the light of long-term
517 linked selection. *Evolution Letters* 1:118-131.
518 17. Ravinet M*, et al.* (2017) Interpreting the genomic landscape of speciation: a road
519 map for finding barriers to gene flow. *J Evol Biol* 30(8):1450-1477.
520 18. Corbett-Detig RB, Hartl DL, & Sackton TB (2015) Natural selection constrains
521 neutral diversity across a wide range of species. *Plos Biol* 13(4):e1002112.
522 19. Maynard-Smith J & Haigh J (1974) Hitch-hiking effect of a favorable gene. *Genet*
523 *Res* 23(1):23-35.
524 20. Hudson RR & Kaplan NL (1995) Deleterious background selection with
525 recombination. *Genetics* 141(4):1605-1617.
526 21. Gillespie JH (2000) Genetic drift in an infinite population: The pseudohitchhiking
527 model. *Genetics* 155(2):909-919.
528 22. Charlesworth B, Morgan MT, & Charlesworth D (1993) The effect of deleterious
529 mutations on neutral molecular variation. *Genetics* 134(4):1289-1303.
530 23. Coop G & Ralph P (2012) Patterns of neutral diversity under general models of
531 selective sweeps. *Genetics* 192(1):205-224.
532 24. Charlesworth B (1998) Measures of divergence between populations and the
533 effect of forces that reduce variability. *Mol Biol Evol* 15(5):538-543.
534 25. Pease JB & Hahn MW (2013) More accurate phylogenies inferred from low-
535 recombination regions in the presence of incomplete lineage sorting. *Evolution*
536 67(8):2376-2384.
537 26. Cutter AD & Payseur BA (2013) Genomic signatures of selection at linked sites:
538 unifying the disparity among species. *Nature Rev Gen* 14(4):262-274.
539 27. Chase MA, Stankowski S, & Streisfeld MA (2017) Genomewide variation
540 provides insight into evolutionary relationships in a monkeyflower species
541 complex (Mimulus sect. Diplacus). *Am J Bot* 104(10):1510-1521.
542 28. McMinn HE (1951) Studies in the genus *Diplacus*, Scrophulariaceae. *Madrono*
543 11:33-128.
544 29. Thompson DM (2005) Systematics of *Mimulus* subgenus Schizoplacus
545 (Scrophulariaceae). *Systematic Botany Monographs* 75:1-213.
546 30. Sobel JM & Streisfeld MA (2015) Strong premating reproductive isolation drives
547 incipient speciation in Mimulus aurantiacus. *Evolution* 69(2):447-461.

31. Stankowski S, Sobel JM, & Streisfeld MA (2017) Geographic cline analysis as a tool for studying genome-wide variation: a case study of pollinator-mediated divergence in a monkeyflower. *Mol Ecol* 26(1):107-122.
32. Stankowski S, Sobel JM, & Streisfeld MA (2015) The geography of divergence with gene flow facilitates multitrait adaptation and the evolution of pollinator isolation in Mimulus aurantiacus. *Evolution* 69(12):3054-3068.
33. Stankowski S & Streisfeld MA (2015) Introgressive hybridization facilitates adaptive divergence in a recent radiation of monkeyflowers. *P Roy Soc B-Biol Sci* 282(1814):154-162.
34. Streisfeld MA & Kohn JR (2007) Environment and pollinator-mediated selection on parapatric floral races of Mimulus aurantiacus. *J Evol Biol* 20(1):122-132.
35. Streisfeld MA & Kohn JR (2005) Contrasting patterns of floral and molecular variation across a cline in Mimulus aurantiacus. *Evolution* 59(12):2548-2559.
36. Streisfeld MA, Young WN, & Sobel JM (2013) Divergent Selection Drives Genetic Differentiation in an R2R3-MYB Transcription Factor That Contributes to Incipient Speciation in Mimulus aurantiacus. *Plos Genet* 9(3).
37. Vickery RK (1995) Speciation by Aneuploidy and Polyploidy in Mimulus (Scrophulariaceae). *Great Basin Nat* 55(2):174-176.
38. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, & Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210-3212.
39. Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* 7:44.
40. Tajima F (1983) Evolutionary Relationship of DNA-Sequences in Finite Populations. *Genetics* 105(2):437-460.
41. Maddison WP (1997) Gene trees in species trees. *Syst Biol* 46(3):523-536.
42. Pease JB, Haak DC, Hahn MW, & Moyle LC (2016) Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *Plos Biol* 14(2).
43. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312-1313.
44. Lamichhaney S*, et al.* (2015) Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* 518(7539):371-375.
45. Richards EJ & Martin CH (2017) Adaptive introgression from distant Caribbean islands contributed to the diversification of a microendemic adaptive radiation of trophic specialist pupfishes. *Plos Genet* 13(8).
46. Pamilo P & Nei M (1988) Relationships between Gene Trees and Species Trees. *Mol Biol Evol* 5(5):568-583.
47. Suh A, Smeds L, & Ellegren H (2015) The Dynamics of Incomplete Lineage Sorting across the Ancient Adaptive Radiation of Neoavian Birds. *Plos Biol* 13(8).
48. Poelstra JW*, et al.* (2014) The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* 344(6190):1410-1414.
49. Burri R*, 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 Res* 25(11):1656-1665.

50. Van Doren BM, *et al.* (2017) Correlated patterns of genetic diversity and differentiation across an avian family. *Mol Ecol* 26(15):3982-3997.
51. Vijay N, *et al.* (2017) Genomewide patterns of variation in genetic diversity are shared among populations, species and higher-order taxa. *Mol Ecol* 26(16):4284-4295.
52. Cruickshank TE & Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol Ecol* 23(13):3133-3157.
53. Slotte T (2014) The impact of linked selection on plant genomic variation. *Brief Funct Genomics* 13(4):268-275.
54. Ellegren H & Wolf JBW (2017) Parallelism in genomic landscapes of differentiation, conserved genomic features and the role of linked selection. *J Evol Biol* 30(8):1516-1518.
55. Matthey-Doret R & Whitlock MC (2018) Background selection and the statistics of population differentiation: consequences for detecting local adaptation. *BiorXiv*.
56. Rastas P, Calboli FCF, Guo BC, Shikano T, & Merila J (2016) Construction of Ultradense Linkage Maps with Lep-MAP2: Stickleback F-2 Recombinant Crosses as an Example. *Genome Biol Evol* 8(1):78-93.
57. Holt C & Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC bioinformatics* 12:491.
58. McKenna A, *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297-1303.
59. Felsenstein J (1985) Phylogenies and the comparative method. *American Naturalist* 125:1-15.
60. Coyne JA & A. OH (1989) Patterns of speciation in Drosophila. *Evolution* 43:362-381.

## *Figure captions*

**Figure 1. Evolutionary relationships and patterns of genome-wide variation across the radiation**. A) The black tree was constructed from a concatenated alignment of genome-wide SNPs and is rooted using *M. clevelandii*. The 387 gray trees were constructed from 500 kb genomic windows. The first number associated with each node or taxon is the bootstrap support for that clade in the whole genome tree, and the second number is the percentage of window-based trees in which that clade is present. B) Levels of differentiation ($F_{ST}$), C) divergence ($d_{xy}$), and D) diversity ($\pi$) within and among taxa based on the same 500 kb windows. For simplicity, $F_{ST}$ and $d_{xy}$ are shown only for comparisons with the red ecotype of subspecies *puniceus*.

**Figure 2. Common differentiation and diversity landscapes mirror variation in the local properties of the genome.** A) Tree concordance scores for 500 kb non-overlapping genomic windows plotted across the 10 bush monkeyflower chromosomes. B – D) Plots of the *z*-transformed first principal component (PC1) for $F_{ST}$, $d_{xy}$, and $\pi$ in overlapping

500 kb windows (step size = 50 kb). PC1 explains 66%, 70%, and 85% of the variation in $F_{ST}$, $d_{xy}$ and $\pi$, respectively. E – F) Gene count and recombination rate (cM/Mbp) in overlapping 500 kb windows.

**Figure 3. Correlations reveal the impact of heterogeneous linked selection across the genome.** Matrix of pairwise correlation coefficients between PC1 $F_{ST}$, PC1 $d_{xy}$, PC1 $\pi$, tree concordance, gene density, and recombination rate. The heat map and the shape of the ellipse indicate the strength of the correlation and its sign. All correlations are statistically significant at $p < 0.001$. Detailed scatterplots for each relationship can be found in Fig. S9. See Fig. S10 for a correlation matrix for 100 kb windows.

**Figure 4. Time-course analysis reveals static and dynamic signatures of recurrent linked selection.** Correlations between variables (500 kb windows) for all 36 taxonomic comparisons (gray dots) plotted against the average $d_{xy}$ as a measure of divergence time. The top row shows how the relationships between $\pi$ (each window averaged across a pair of taxa) and (A) $d_{xy}$, (B) gene count, and (C) recombination rate vary with increasing divergence time. The bottom row (D-E) shows the same relationships, but with $F_{ST}$. The regressions (dashed lines) in each plot are fitted to the eight independent contrasts (colored points) obtained using a phylogenetic correction. The color gradient shows the strength of the correlation. Details for each regression can be found in Table S5.

**Figure 5. Emergence of a heterogeneous differentiation landscape across 1.5 million years of divergence.** A) Plots of $F_{ST}$ (500 kb windows) across the genome for pairs of taxa at early (red vs. yellow), intermediate (red vs. *parviflorus*) and late stages (red vs. *M. clevelandii*) of divergence. B) Average nucleotide diversity (for red, *yellow*, *par*., and *clv*.) across the genome in 500 kb windows. The geographic distribution (parapatric or allopatric), percent sequence divergence ($d_{xy}*10^{-2}$) and correlation between $F_{ST}$ and mean $\pi$ are provided next to each taxon pair.

*Supplementary information for:*

# The tempo of linked selection: rapid emergence of a heterogeneous genomic landscape during a radiation of monkeyflowers

Sean Stankowski[1,2,§], Madeline A. Chase[1,3,§], Allison M. Fuiten[1], Peter L. Ralph[1], and Matthew A. Streisfeld[1*].

[1]Institute of Ecology and Evolution, 335 Pacific Hall 5289, University of Oregon, Eugene, OR 97405, USA
[2]Present address: Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK
[3]Present address: Department of Evolutionary Biology, Evolutionary Biology Centre (EBC), Uppsala University, Uppsala, Sweden

[§]These authors contributed equally to this work

*Corresponding author: mstreis@uoregon.edu

## ORCID ids:

SS: https://orcid.org/0000-0003-0472-9299; MAC: https://orcid.org/0000-0002-7916-3560; AMF: https://orcid.org/0000-0003-3926-5455 PLR: https://orcid.org/0000-0002-9459-6866; MAS: https://orcid.org/0000-0002-2660-8642

## Supplementary Materials and Methods

*Genome Assembly*

We used a combination of short-read Illumina and long-read Single Molecule, Real Time (SMRT) sequencing to assemble the genome of a single individual from the red ecotype of *M. aurantiacus* subspecies *puniceus* (population UCSD; Table S1). Genomic DNA was isolated from leaf tissue using either ZR plant/seed DNA miniprep kits (Zymo Research) or GeneJet Plant Genomic DNA purification kits (Thermo Fisher). Illumina libraries were generated following the *Allpaths-LG* assembly pipeline (Gnerre et al. 2011), which included a single fragment library with average 180 bp insert size and three mate pair libraries (average insert sizes: 3.5-5 kb, 5-7 kb, and 7-13 kb). Libraries were sequenced on the Illumina HiSeq 2500 using paired-end 100 bp reads. An initial scaffold-level assembly was performed with *Allpaths-LG* using default parameters and the *haploidify* function enabled. This assembly yielded 11,123 contigs (N50 = 40.5 kb) and 2,299 scaffolds (N50 = 1,310 kb), for a total assembly size of 193.3 Mbp. Long-read sequencing was performed from the same individual using 12 SMRT cells sequenced on the Pacific Biosystems RS II machine at Duke University. We obtained a total of 6.4 Gbp of sequence, which corresponds to ~21 × coverage of the genome. The PacBio reads were used to re-scaffold the *Allpaths-LG* scaffolds using *Opera-LG* (Gao et al. 2016). This reduced the number of scaffolds to 1,547 (N50 = 1,578 kb).

We then manually improved the scaffold containing the flower color gene *MaMyb2* (Streisfeld et al. 2013). We first aligned this scaffold to a previously published draft sequence assembly from this same individual (Stankowski et al. 2017), which was generated using Illumina short-reads and the *Velvet* assembler (Zerbino and Birney 2008). We used long range PCR and cloning to generate Sanger sequences across three regions within 20 kb of *MaMyb2* that did not assemble well. Genomic DNA was amplified using Phusion high fidelity polymerase (NEB). PCR products were cloned into the pCR2.1 TOPO-TA vector (Life Technologies), and purified plasmids were sequenced with Sanger technology. Resulting sequences were aligned to the scaffold containing *MaMyb2*, and new PCR primers were designed to sequence internal fragments until the entire insert was sequenced. Using this approach, we sequenced a total of 9,824 bp across the three regions. The reference sequence in the assembly was corrected manually to match the Sanger data.

Finally, we gap filled the assembly using the PacBio data and the program *PBJelly* (English et al. 2012). Resulting scaffolds were assembled into pseudomolecules using *Chromonomer* (http://catchenlab.life.illinois.edu/chromonomer/), according to the online manual. This software anchored and oriented scaffolds based on the order of markers in a high-density linkage map (see below) and made corrections to scaffolds when differences occurred between the genetic and physical positions of markers in the map. A final round of gap filling with *PBJelly* was performed to fill any gaps that were created by broken scaffolds in *Chromonomer*. To assess the completeness of the gene space in the assembly, we used both the BUSCO and CEGMA pipelines to estimate the proportion of 956 single copy plant genes (BUSCO) or 248 core eukaryotic genes (CEGMA) that were completely or partially assembled (Parra et al. 2007; Simao et al. 2015). The proportion of these genes present in an assembly has been shown to be correlated with the total proportion of assembled gene space, and thus serves as a good predictor of assembly completeness.

*Construction of high-density linkage map*

We generated an outbred $F_2$ mapping population by crossing two $F_1$ individuals, each the product of crosses between different greenhouse-raised red and yellow ecotype plants collected from one red ecotype and one yellow ecotype population (populations UCSD and LO, respectively; Table S1). We then used restriction-site associated DNA sequencing (RADseq) to genotype $F_1$ and $F_2$ individuals. DNA was extracted from leaf material using Zymo ZR plant/seed DNA miniprep kits, and RAD library preparation followed the protocol outlined in Sobel and Streisfeld (Sobel and Streisfeld 2015). Libraries were sequenced on the Illumina HiSeq 2000 platform using single-end 100 bp reads at the Genomics Core Facility, University of Oregon.

Reads were filtered based on quality, and errors in the barcode sequence or RAD site were corrected using the *process_radtags* script in *Stacks v. 1.35* (Catchen et al. 2011; Catchen et al. 2013). Loci were created using the *denovo_map.pl* function of *Stacks*, with three identical raw reads required to create a stack, two mismatches allowed between loci for an individual, and two mismatches allowed when processing the catalog. Single nucleotide polymorphisms (SNPs) were determined and genotypes called using a maximum-likelihood (ML) statistical model implemented in *Stacks* and a stringent $\chi^2$ significance level of 0.01 to distinguish between heterozygotes and homozygotes. We then used the *genotypes* program implemented in *Stacks* to identify a set of 9,029 mappable markers. We specified a 'CP' cross design ($F_1$ individuals coded as the parents), requiring that a marker was present in at least 85% of progeny at a minimum depth of 12 reads per individual, and we allowed automated corrections to be made to the data.

Linkage map construction was performed using *Lep-MAP2* (Rastas et al. 2016). The data were filtered using the *Filtering* module to include only individuals with less than 15% missing data and excluded markers that showed evidence for extreme segregation distortion ($\chi^2$ test, $P < 0.01$). To assign markers to linkage groups, we used the *SeparateChromosomes* module with a logarithm of odds (LOD) score limit of 20 and no minimum size for linkage groups (LG). This assigned 7,217 markers to 10 linkage groups, which matches the number of chromosomes in *M. aurantiacus*. The *JoinSingles* module was executed again with a LOD limit of 10 to join an additional 877 ungrouped markers to the 10 previously formed LGs. Fifty-seven singles that were not joined at this stage were discarded from the dataset. Initial marker orders were determined using sex-averaged and sex-specific recombination rates using the *OrderMarkers* module. For each LG, we conducted 10 independent runs using the Kosambi mapping function (*useKosambi=1*), with the dataset split into seven pseudofamilies to take advantage of parallel processing. When multiple markers had identical genotypes, only the duplicate marker with the least missing data was used in marker ordering. We retained the marker order from the run with the best likelihood. After removing markers with an error rate > 0.05, the ML order was re-evaluated using the *evaluateOrder* flag. The map contained 8,094 informative loci from 269 F2 individuals, with an average of 3.5% ± SD 3.86 missing data per individual.

After the integration of our assembly and genetic map using the *Chromonomer* software (Amores et al. 2014), we made corrections to the map order based on the physical position of markers within assembled scaffolds. Using the output of

*Chromonomer*, we identified markers that were out of order in the map compared to their local assembly order and aligned these markers to the assembly from *Chromonomer* using *Bowtie2* v. 2.2.5 (Langmead and Salzberg 2012) with the *very_sensitive* settings to obtain their physical order. We then re-estimated the map using the *evaluateOrder* flag in *Lep-MAP2* as described above, but with the marker order constrained to the physical order (*improveOrder=0*) and with all duplicate markers included in the analysis (*removeDuplicates=0*). After initial map construction, we removed 17 markers with an estimated error rate greater than 5% and estimated the map one last time using the same settings. The final map contained 7,589 markers across the 10 linkage groups.

*Genome annotation*

Prior to genome annotation, the assembly was soft-masked for repetitive elements and areas of low complexity with *RepeatMasker* (RepeatMasker Open-4.0) using a custom *Mimulus aurantiacus* library created by *RepeatModeler* (RepeatModeler Open-1.0), Repbase repeat libraries (Jurka et al. 2005), and a list of known transposable elements provided by *MAKER* (Holt and Yandell 2011). In total, 30.99% of the genome assembly was masked by *RepeatMasker*. Repetitive elements were annotated with *RepeatModeler*. Hidden Markov Models for gene prediction were generated by *SNAP* (Korf 2004) and *Augustus* (Stanke and Waack 2003)and were trained iteratively to the assembly using *MAKER*, as described by Cantarel et al. (Cantarel et al. 2008). Training was performed on the 14.5 Mbp sequence from LG9. Evidence used by *MAKER* for annotation included protein sequences from *Arabidopsis thaliana*, *Oryza sativa*, *Solanum lycopersicum*, *Solanum tuberosum*, *Daucus carota*, *Vitis vinifera* (all downloaded from EnsemblPlants on 9 August 2016), *Salvia miltiorrhiza* (downloaded from Herbal Medicine Omics Database on 9 August 2016), *Mimulus guttatus v. 2* (downloaded from JGI Genome Portal on 9 August 2016), and all Uniprot/swissprot proteins (downloaded on 18 August 2016) (Goodstein et al. 2012; Nordberg et al. 2013; Kersey et al. 2016)(Herbal Medicine Omics Database; Uniprot). We filtered the annotations with *MAKER* to include: 1) only evidence-based information that also contained assembled protein support, and 2) those *ab initio* gene predictions that did not overlap with the evidence-based annotations and that contained protein family domains, as detected with InterProScan (Quevillon et al. 2005).

*Genome re-sequencing and variant calling*

We collected leaf tissue from four to five individuals from seven subspecies of *M. aurantiacus*, including both ecotypes of subspecies *puniceus* (Table S3; Fig. S2). In addition, we collected leaf tissue from three individuals of *M. clevelandii*. We extracted DNA from dried tissue using the Zymo Plant/Seed MiniPrep DNA kit following the manufacturer's instructions. We prepared sequencing libraries using the Kapa Biosystems HyperPrep kit, and libraries with an insert size between 400-600 bp were sequenced on the Illumina HiSeq 4000 using paired-end 150 bp reads at the Genomics Core Facility, University of Oregon.

We filtered raw reads using the *process_shortreads* script in *Stacks* v1.46 to remove reads with uncalled bases or poor quality scores. We then aligned the retained reads to the reference assembly using the BWA-MEM algorithm in *BWA* v0.7.15 (Li 2013). An average of 91.7% of reads aligned (range: 82.6-96.0%), and the average

sequencing depth was 21x (range: 15.16 – 30.86x). We then marked PCR duplicates using *Picard* (http://broadinstitute.github.io/picard). We performed an initial run of variant calling using the UnifiedGenotyper tool in *GATK* v3.8 (McKenna et al. 2010) and filtered the data to remove variants with a mapping quality < 50, a quality depth < 4, and a Fisher Strand score > 50. We then used these variants to perform base quality score recalibration for each individual, before performing another run of the UnifiedGenotyper to call final variants. After the second run of variant calling, we removed variants with a mapping quality < 40, a quality depth < 2, and a Fisher Strand score > 60. The final dataset contained 13,233,829 SNPs across the nine taxa. Finally, we ran UnifiedGenotyper with the EMIT_ALL_SITES option to output all variant and invariant genotyped sites.

*Phylogenetic analyses*

Initially, we used *RAxML* v8 to reconstruct the evolutionary relationships among the nine taxa by concatenating variant sites from across the genome. To investigate patterns of phylogenetic discordance across the genome, we also built trees from windows across the genome. We phased SNPs using *BEAGLE* v4.1 (Browning and Browning 2007), using a window size of 100 kb and an overlap of 10 kb. We incorporated information on recombination rate from the genetic map and did not impute missing genotypes. After phasing, we used *MVFtools* (https://www.github.com/jbpease/mvftools) to run *RAxML* from 100 kb and 500 kb nonoverlapping windows, with the *M. clevelandii* samples set as the outgroups. We then visualized the window trees in *DensiTree* v2.01 (Bouckaert 2010).

To assess concordance between the window-based trees and the whole-genome tree, we converted trees to distance matrices using the *Ape* package in R (Paradis et al. 2004). We then calculated the Pearson's correlation coefficient between the distance matrix from each window and the whole-genome tree, with a stronger correlation indicating higher concordance with the whole-genome tree. We used one-dimensional autocorrelation analysis to determine if variation in tree concordance was randomly distributed across the genome. This involved estimating the autocorrelation between genomic position and tree concordance for each LG with a lag size of 2 Mbp. The significance of the observed value for each LG was determined from a null distribution of autocorrelation coefficients estimated from 1000 random permutations of the genome-wide data.

We also conducted a Principal Components Analysis (PCA) based on all variant sites from across the entire genome using *Plink* v. 1.90 (Chang et al. 2015). Initially, we ran the PCA with all 37 samples, but we consecutively re-ran the analysis by removing different taxa in order to assess clustering patterns among more closely related samples.

*Population genomic analyses*

To examine how genome-wide patterns of diversity, differentiation, and divergence varied among taxa, we calculated within-taxon nucleotide diversity ($\pi$), between-taxon relative differentiation ($F_{ST}$), and between-taxon absolute divergence ($d_{xy}$) across non-overlapping 100 kb and 500 kb windows using custom Python scripts downloaded from https://github.com/simonhmartin/genomics_general. We calculated measures of differentiation and divergence across all 36 pairwise comparisons among the nine taxa,

and diversity was estimated separately for each taxon. These scripts estimated π and $d_{xy}$ by dividing the number of sequence differences within a window (either within or between taxa) by the total number of sites in that window. To account for missing data, the script counted the number of differences between each sample, divided by the total number of variant sites that were genotyped within those samples, and then averaged across all pairs of samples. To provide an unbiased estimate of diversity and divergence, we incorporated invariant sites into the calculation by dividing the number of pairwise differences (within and between taxa, respectively) by the total number of genotyped sites (variant and invariant) within a window. $F_{ST}$ was calculated following the measure of $K_{ST}$ (Hudson et al. 1992), equation 9), but was modified to incorporate missing data using the same approach as π and $d_{xy}$. We filtered the data separately for each taxonomic comparison, so that each site was required to be genotyped in at least three individuals for comparisons within the *M. aurantiacus* complex or at least two individuals for each comparison involving *M. clevelandii.*

We summarized the variation in each statistic across comparisons using a Principal Components Analysis (PCA), with taxon or taxon pair as the variables. Thus, across each window, the first principal component of π, $F_{ST}$, and $d_{xy}$ provided multivariate measures that explained the greatest covariance in the data. We used a one-dimensional autocorrelation analysis and permutation test to determine if the genome-wide patterns of PC1 π, $F_{ST}$, and $d_{xy}$ departed from a random expectation, as described above for tree concordance (see section 'phylogenetic analyses').

To examine the relationships among PC1 diversity, differentiation, and divergence, we estimated Pearson's correlation coefficient among all three statistics across genomic windows. Further, we estimated correlations among these three statistics and tree concordance, gene density, and recombination rate. Recombination rate was estimated by comparing the genetic and physical distance (in cM/Mbp) between all pairs of adjacent markers on each LG from the genetic linkage map described above. We removed the top 5% of recombination rates, as these represented unrealistically high values of recombination. A minimum of three estimates was required to obtain an average recombination rate estimate within each window. Gene density was calculated from the number of predicted genes in each window, as determined from the annotation described above.

To determine how the correlations among the statistics (diversity, differentiation, divergence, recombination rate, gene count) changed with increasing divergence time, we examined the correlation coefficient among all pairs of statistics individually for each of the 36 pairwise comparisons. Because diversity was measured within taxa rather than between them, we calculated the mean value of π between each pair of taxa. Also, because many of the pairwise comparisons are non-independent, we applied the phylogenetic correction outlined by (Felsenstein 1985; Coyne and A. 1989) to produce a statistically independent set of data points for this analysis.

As a measure of the divergence time between *M. clevelandii* and *M. aurantiacus*, we estimated the percent sequence divergence ($d_{xy}$) between individuals of *M. clevelandii* and all subspecies of *M. aurantiacus* combined. We then converted this value into a divergence time $T$ (in generations) using the equation: $T = d_{xy}/(2\mu)$, where μ is the mutation rate, $1.5 \times 10^{-8}$ (Koch et al. 2001; Brandvain et al. 2014). This value was then converted into years by multiplying by a generation time of two years.

**Supplemental References**


Amores, A., J. Catchen, I. Nanda, W. Warren, R. Walter, M. Schartl, and J. H. Postlethwait. 2014. A RAD-tag genetic map for the platyfish (Xiphophorus maculatus) reveals mechanisms of karyotype evolution among teleost fish. Genetics 197:625-641.

Bouckaert, R. R. 2010. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics 26:1372-1373.

Brandvain, Y., A. M. Kenney, L. Flagel, G. Coop, and A. L. Sweigart. 2014. Speciation and Introgression between Mimulus nasutus and Mimulus guttatus. Plos Genet 10.

Browning, S. R., and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-1097.

Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. S. Alvarado, and M. Yandell. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18:188-196.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124-3140.

Catchen, J. M., A. Amores, P. Hohenlohe, W. Cresko, and J. H. Postlethwait. 2011. Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. G3-Genes Genom Genet 1:171-182.

Chang, C. C., C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4.

Coyne, J. A., and O. H. A. 1989. Patterns of speciation in Drosophila. Evolution 43:362-381.

English, A. C., S. Richards, Y. Han, M. Wang, V. Vee, J. Qu, X. Qin, D. M. Muzny, J. G. Reid, K. C. Worley, and R. A. Gibbs. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PloS one 7:e47768.

Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist 125:1-15.

Gao, S., D. Bertrand, B. K. Chia, and N. Nagarajan. 2016. OPERA-LG: efficient and exact scaffolding of large, repeat-rich eukaryotic genomes with performance guarantees. Genome biology 17:102.

Gnerre, S., I. Maccallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences of the United States of America 108:1513-1518.

Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam, and D. S. Rokhsar. 2012. Phytozome: a comparative platform for green plant genomics. Nucleic acids research 40:D1178-1186.

Holt, C., and M. Yandell. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC bioinformatics 12:491.

Hudson, R. R., D. D. Boos, and N. L. Kaplan. 1992. A Statistical Test for Detecting Geographic Subdivision. Mol Biol Evol 9:138-151.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research 110:462-467.

Kersey, P. J., J. E. Allen, I. Armean, S. Boddu, B. J. Bolt, D. Carvalho-Silva, M. Christensen, P. Davis, L. J. Falin, C. Grabmueller, J. Humphrey, A. Kerhornou, J. Khobova, N. K. Aranganathan, N. Langridge, E. Lowy, M. D. McDowall, U. Maheswari, M. Nuhn, C. K. Ong, B. Overduin, M. Paulini, H. Pedro, E. Perry, G. Spudich, E. Tapanari, B. Walts, G. Williams, M. Tello-Ruiz, J. Stein, S. Wei, D. Ware, D. M. Bolser, K. L. Howe, E. Kulesha, D. Lawson, G. Maslen, and D. M. Staines. 2016. Ensembl Genomes 2016: more genomes, more complexity. Nucleic acids research 44:D574-580.

Koch, M., B. Haubold, and T. Mitchell-Olds. 2001. Molecular systematics of Brassicaceae: evidence from plastidic matK and nuclear Chs sequences. Am J Bot 88:534-544.

Korf, I. 2004. Gene finding in novel genomes. BMC bioinformatics 5:59.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nature methods 9:357-359.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997v1.

Nordberg, H., M. Cantor, S. Dusheyko, S. Hua, A. Poliakov, I. Shabalov, T. Smirnova, I. V. Grigoriev, and I. Dubchak. 2013. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic acids research 42:D26-D31.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289-290.

Parra, G., K. Bradnam, and I. Korf. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genornes. Bioinformatics 23:1061-1067.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. 2005. InterProScan: protein domains identifier. Nucleic acids research 33:W116-W120.

Rastas, P., F. C. F. Calboli, B. C. Guo, T. Shikano, and J. Merila. 2016. Construction of Ultradense Linkage Maps with Lep-MAP2: Stickleback F-2 Recombinant Crosses as an Example. Genome Biol Evol 8:78-93.

Simao, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210-3212.

Sobel, J. M., and M. A. Streisfeld. 2015. Strong premating reproductive isolation drives incipient speciation in Mimulus aurantiacus. Evolution 69:447-461.

Stanke, M., and S. Waack. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19:Ii215-Ii225.

Stankowski, S., J. M. Sobel, and M. A. Streisfeld. 2017. Geographic cline analysis as a tool for studying genome-wide variation: a case study of pollinator-mediated divergence in a monkeyflower. Mol Ecol 26:107-122.

Streisfeld, M. A., W. N. Young, and J. M. Sobel. 2013. Divergent Selection Drives Genetic Differentiation in an R2R3-MYB Transcription Factor That Contributes to Incipient Speciation in Mimulus aurantiacus. Plos Genet 9.

Zerbino, D. R., and E. Birney. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821-829.

## Supplementary tables

**Table S1. Summary of the genetic linkage map constructed using an F2 intercross between the red and yellow ecotypes of subspecies *puniceus*.** The table includes map length in cM for each linkage group (LG), the number of markers associated with each LG, the number of unique map positions, and the average genetic distance in cM between each unique map position. Standard deviations are given in parentheses.

| LG | Map length (cM) | Number of markers | Unique map positions | Avg. genetic dist. between unique markers in cM (sd) |
|---|---|---|---|---|
| **1** | 93.9 | 969 | 335 | 0.28 (0.33) |
| **2** | 71.37 | 893 | 253 | 0.28 (0.88) |
| **3** | 76.3 | 912 | 256 | 0.30 (0.65) |
| **4** | 70.2 | 851 | 257 | 0.28 (0.41) |
| **5** | 78.7 | 778 | 295 | 0.27 (0.29) |
| **6** | 69.1 | 741 | 247 | 0.28 (0.88) |
| **7** | 59.8 | 738 | 234 | 0.26 (0.33) |
| **8** | 65.6 | 674 | 246 | 0.27 (0.41) |
| **9** | 68.6 | 623 | 182 | 0.37 (0.58) |
| **10** | 71.1 | 410 | 150 | 0.48 (0.96) |
| **Avg.** | 72.74 (8.69) | 758.90 (155.75) | 245.50 (49.04) | 0.31 (0.06) |

**Table S2. Analysis of gene space completeness in the *M. aurantiacus* genome using CEGMA and BUSCO.** The number and percent of core genes found in the final assembly are shown for each analysis (CEGMA, *n* = 248; BUSCO, *n* = 1440).

| Analysis | # Genes | % Found in Assembly |
|---|---|---|
| **CEGMA Complete** | 233 | 93.95 |
| **CEGMA Partial** | 244 | 98.39 |
| **BUSCO total complete (duplicated)** | 1340 (61) | 93 (4.2) |
| **BUSCO Fragmented** | 29 | 2.0 |
| **BUSCO Missing** | 71 | 5.0 |

**Table S3. Sample information for the 37 sequenced samples.** Includes their taxon identity, sampling location, percent read alignment, and total sequencing depth.

| Sample | Taxon | Latitude | Longitude | % Reads aligned | Seq. Depth |
|---|---|---|---|---|---|
| **159_83** | ssp. *aridus* | 32.6630 | -116.2230 | 91.7 | 21.12 |
| **159_84** | ssp. *aridus* | 32.6630 | -116.2230 | 89.3 | 21.98 |
| **195_1** | ssp. *aridus* | 32.6300 | -116.1429 | 92.6 | 20.20 |
| **T84** | ssp. *aridus* | 32.6526 | -116.2449 | 87.2 | 21.75 |
| **T102** | ssp. *aurantiacus* | 39.0424 | -122.7727 | 94.9 | 23.74 |
| **T104** | ssp. *aurantiacus* | 39.2045 | -123.7646 | 94.6 | 25.09 |
| **T50** | ssp. *aurantiacus* | 35.9865 | -121.4928 | 88.3 | 24.36 |
| **T92** | ssp. *aurantiacus* | 37.8459 | -120.6110 | 94.0 | 15.16 |
| **T144** | ssp. *calycinus* | 34.1929 | -117.2784 | 93.2 | 26.00 |
| **T150** | ssp. *calycinus* | 33.8564 | -116.8481 | 94.7 | 24.02 |
| **T90** | ssp. *calycinus* | 35.5918 | -118.5052 | 91.3 | 19.97 |
| **T91** | ssp. *calycinus* | 35.3172 | -118.5871 | 95.5 | 27.91 |
| **T101** | ssp. *grandiflorus* | 39.5536 | -121.4301 | 92.0 | 16.05 |
| **T61** | ssp. *grandiflorus* | 39.5590 | -120.8243 | 91.6 | 17.31 |
| **T96** | ssp. *grandiflorus* | 39.0122 | -120.7552 | 92.0 | 28.21 |
| **T99** | ssp. *grandiflorus* | 39.4376 | -121.0599 | 91.4 | 23.84 |
| **DPR** | ssp. *longiflorus* | 33.7459 | -117.4485 | 96.0 | 26.88 |
| **SS** | ssp. *longiflorus* | 34.2722 | -118.6100 | 94.2 | 30.86 |
| **T33** | ssp. *longiflorus* | 34.3438 | -118.5099 | 94.6 | 18.87 |
| **T8** | ssp. *longiflorus* | 34.1347 | -118.6452 | 82.6 | 25.11 |
| **KK168** | ssp. *parviflorus* | 34.0180 | -119.6730 | 91.8 | 23.66 |
| **KK161** | ssp. *parviflorus* | 34.0180 | -119.6730 | 92.0 | 19.11 |
| **KK180** | ssp. *parviflorus* | 34.0180 | -119.6730 | 92.4 | 18.18 |
| **KK182** | ssp. *parviflorus* | 34.0193 | -119.6802 | 91.3 | 19.46 |
| **ELF** | ssp. *puniceus*, red | 33.0860 | -117.1453 | 93.0 | 18.20 |
| **JMC** | ssp. *puniceus*, red | 32.7373 | -116.9541 | 93.8 | 19.06 |
| **LH** | ssp. *puniceus*, red | 33.0609 | -117.1188 | 87.1 | 19.77 |
| **MT** | ssp. *puniceus*, red | 32.8210 | -117.0618 | 93.7 | 20.85 |
| **UCSD** | ssp. *puniceus*, red | 32.8894 | -117.2362 | 87.0 | 18.23 |
| **BCRD** | ssp. *puniceus*, yellow | 32.9496 | -116.6380 | 94.6 | 20.85 |
| **INJ** | ssp. *puniceus*, yellow | 33.0979 | -116.6643 | 93.1 | 18.83 |
| **LO** | ssp. *puniceus*, yellow | 32.6767 | -116.3312 | 93.4 | 18.04 |
| **PCT** | ssp. *puniceus*, yellow | 32.7326 | -116.4698 | 92.3 | 19.68 |
| **POTR** | ssp. *puniceus*, yellow | 32.6038 | -116.6339 | 90.5 | 19.27 |
| **CLV_GH** | *M. clevelandii* | 33.1589 | -116.8122 | 92.3 | 21.31 |
| **CLV_11** | *M. clevelandii* | 33.3391 | -116.9325 | 84.4 | 15.52 |
| **CLV_4** | *M. clevelandii* | 33.3391 | -116.9325 | 89.3 | 17.31 |

**Table S4. Loadings for principal component 1 calculated across all 36 pairwise comparisons (for $F_{ST}$ and $d_{xy}$) or all nine taxa (for $\pi$)**

| Comparison | $F_{ST}$ PC1 | $d_{xy}$ PC1 |
|---|---|---|
| **AUR_ARI** | 0.85707 | 0.95575 |
| **BIF_ARI** | 0.89606 | 0.86182 |
| **CAL_ARI** | 0.88738 | 0.95423 |
| **CLV_ARI** | 0.8969 | 0.39278 |
| **FLE_ARI** | 0.85342 | 0.94563 |
| **LON_ARI** | 0.9046 | 0.96057 |
| **AUR_BIF** | 0.81191 | 0.85563 |
| **AUR_CAL** | 0.32407 | 0.90376 |
| **AUR_CLV** | 0.90725 | 0.46698 |
| **AUR_FLE** | 0.79434 | 0.94778 |
| **LON_AUR** | 0.40962 | 0.90381 |
| **ARI_AUS** | 0.90097 | 0.94262 |
| **AUR_AUS** | 0.451 | 0.90512 |
| **BIF_AUS** | 0.91203 | 0.90305 |
| **CAL_AUS** | 0.49255 | 0.88634 |
| **CLV_AUS** | 0.90576 | 0.54264 |
| **FLE_AUS** | 0.86511 | 0.94318 |
| **LON_AUS** | 0.59046 | 0.88521 |
| **CAL_BIF** | 0.89043 | 0.88319 |
| **CLV_BIF** | 0.91181 | 0.32045 |
| **FLE_BIF** | 0.89579 | 0.86769 |
| **LON_BIF** | 0.88549 | 0.89344 |
| **CAL_FLE** | 0.83543 | 0.94124 |
| **LON_CAL** | 0.3786 | 0.87337 |
| **CAL_CLV** | 0.90548 | 0.52607 |
| **FLE_CLV** | 0.91991 | 0.45233 |
| **LON_CLV** | 0.90738 | 0.51482 |
| **LON_FLE** | 0.84197 | 0.94507 |
| **PUN_ARI** | 0.90783 | 0.94574 |
| **PUN_AUR** | 0.51788 | 0.89861 |
| **PUN_AUS** | 0.51423 | 0.84798 |
| **PUN_BIF** | 0.91936 | 0.90695 |
| **PUN_CAL** | 0.62051 | 0.88316 |
| **PUN_CLV** | 0.90758 | 0.55149 |
| **PUN_FLE** | 0.87615 | 0.93825 |
| **PUN_LON** | 0.66912 | 0.88264 |

| Taxon | $\pi$ PC1 |
|---|---|
| **ARI** | 0.9251 |
| **AUR** | 0.86508 |
| **AUS** | 0.97255 |
| **BIF** | 0.77181 |
| **CAL** | 0.96734 |
| **CLV** | 0.92017 |
| **FLE** | 0.91044 |
| **LON** | 0.95166 |

**Table S5. Details for the linear regressions presented in Figure 4 of the main text.**

| Variables | Pearson's r | Regression equation | *p* |
|---|---|---|---|
| $d_{xy}$ & $\pi$ | -0.93 | $y = -60.5x + 1.33$ | < 0.001 |
| $\pi$ & gene count | 0.59 | $y = 3.9x - 0.85$ | 0.130 |
| $\pi$ & cM/mbp | -0.79 | $y = -5.4x + 0.46$ | 0.020 |
| $F_{ST}$ & $\pi$ | -0.89 | $y = -53.3x - 0.13$ | 0.003 |
| $F_{ST}$ & gene count | 0.81 | $y = 46.5x + 0.04$ | 0.016 |
| $F_{ST}$ cM/mbp | -0.73 | $y = -25.5x + 0.01$ | 0.041 |

## Supplementary figures

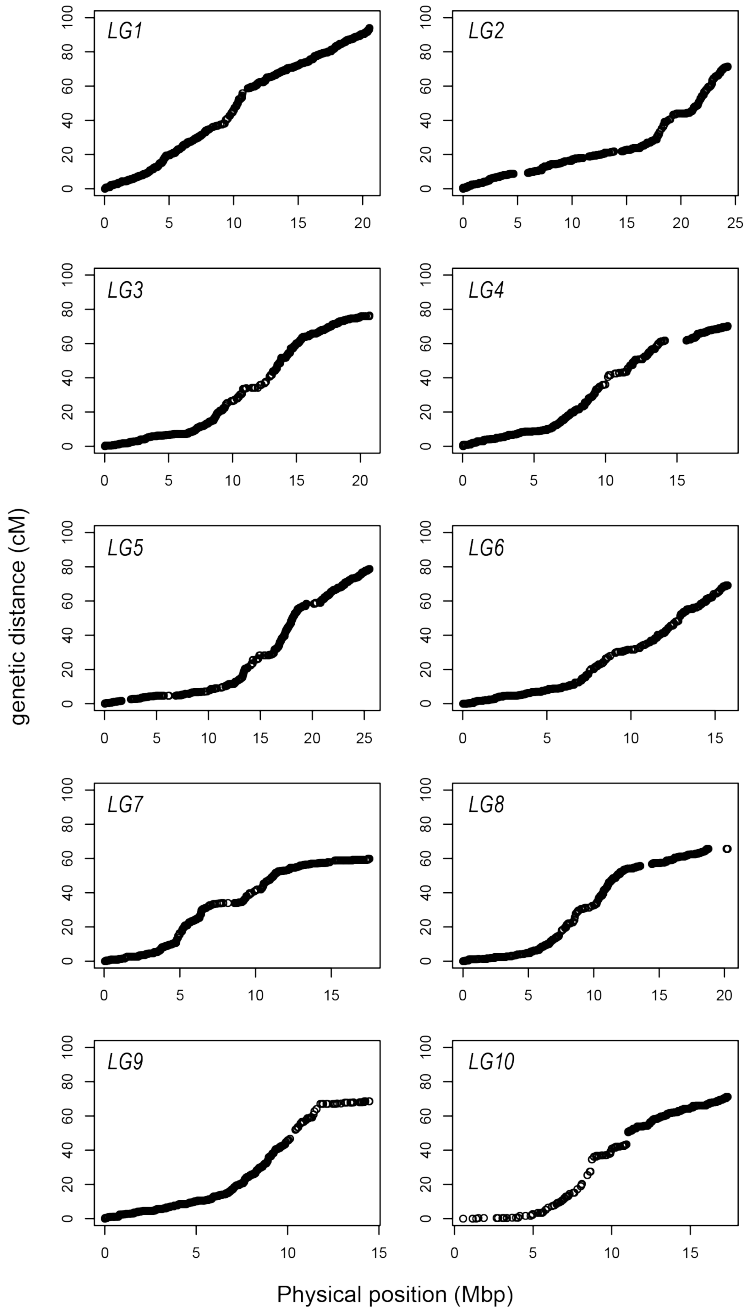



**Figure. S1. Map distance (cM/Mbp) vs. physical distance across the 10 linkage groups.** Recombination for each marker is estimated relative to the start of the linkage group and plotted at its physical location on each chromosome in the reference assembly.

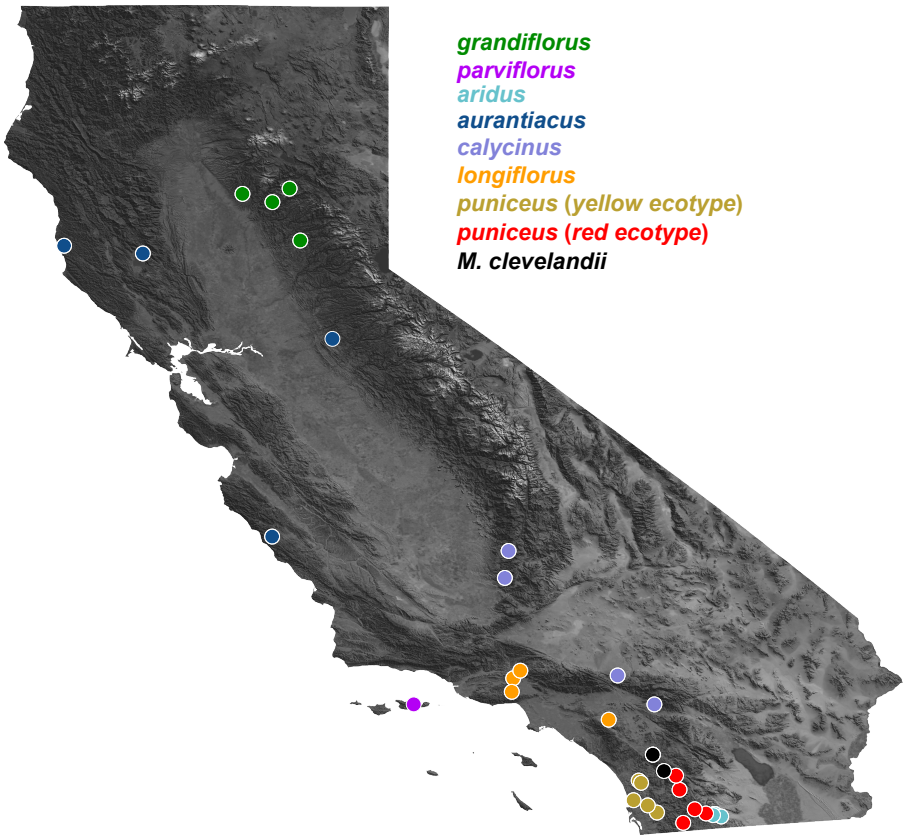



**Figure S2. Geographic distribution of sampling locations for each sample sequenced in this study.** Detailed position information for each population can be found in Table S3.

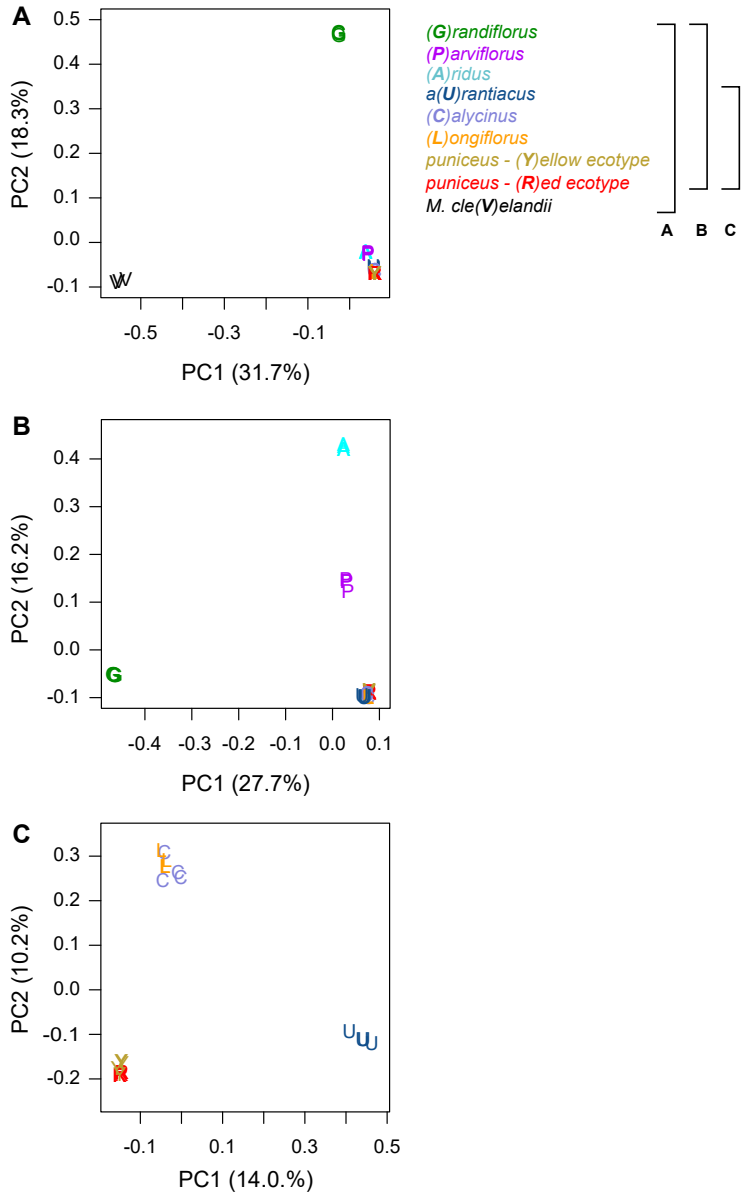



**Figure S3. Genome-wide Principal Components Analysis (PCA).** Each plot is a separate PCA performed using different sets of taxa. The legend to the right describes the set of taxa included in each analysis, with the capital letter in parentheses and the color representing the specific taxon. A) All taxa; B) all subspecies of *M. aurantiacus*, but excluding *M. clevelandii*; C) only subspecies *aurantiacus*, *longiflorus*, *calycinus*, and the red and yellow ecotypes of subspecies *puniceus*. The percent variation explained by each principal component is given in parentheses.

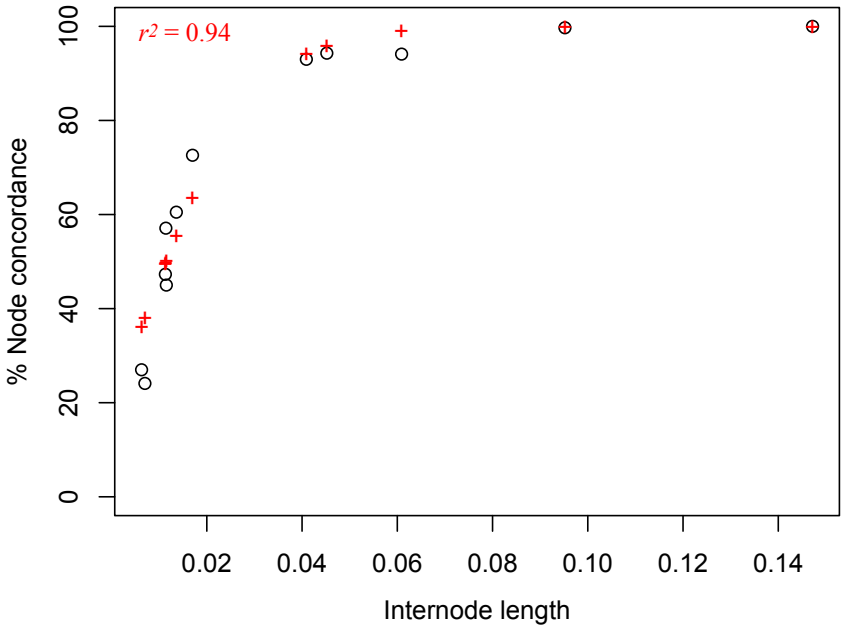



**Figure S4. Incomplete lineage sorting due to rapid diversification.** Clades separated by shorter internodes (i.e., separated by less time) are recovered less frequently in window-based trees (500 kb windows). This indicates a strong role for incomplete lineage sorting in areas of the tree where diversification is rapid. The % node concordance is the percentage of window-based of trees that contain a given node in the genome-wide tree, and is plotted against the length of the internode separating each clade. Only clades at and above the level of taxon were included. The red points are the predicted values from a 4-parameter logistic function fitted to the data using an iterative least-squares method.

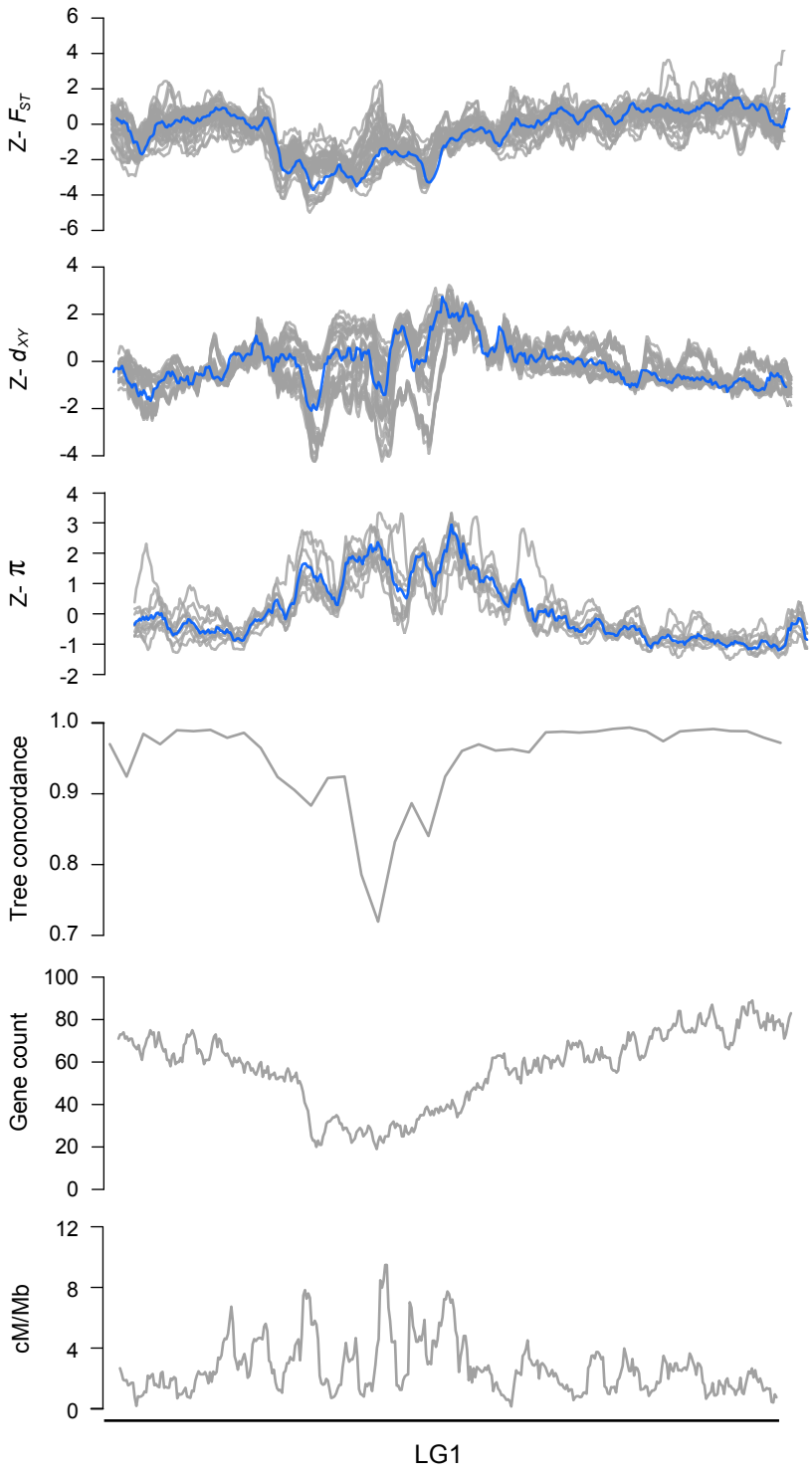



**Figure S5. Patterns of variation plotted across each bush monkeyflower linkage group.** Z-transformed $F_{ST}$, $d_{xy}$, and π in overlapping 500 kb windows (step size = 50 kbp). The gray lines are z-transformed scores for each of the 36 pairwise comparisons ($F_{ST}$ and $d_{xy}$) or nine taxa (π), and the blue line is the z-transformed score for the first principal component (PC1). Estimates of tree concordance, gene count and recombination rate (cM/Mbp) are also shown.

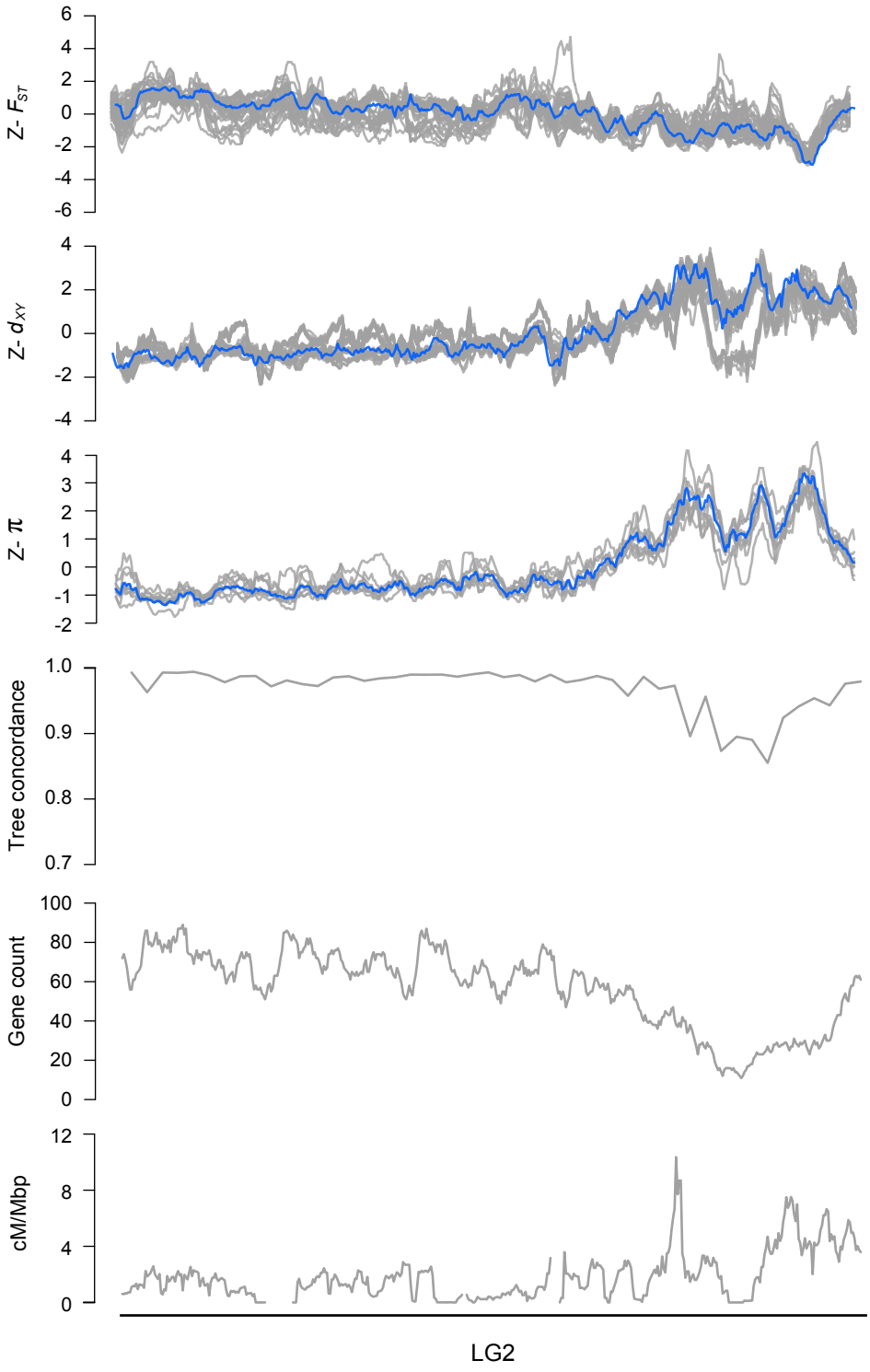



**Figure S5. continued**

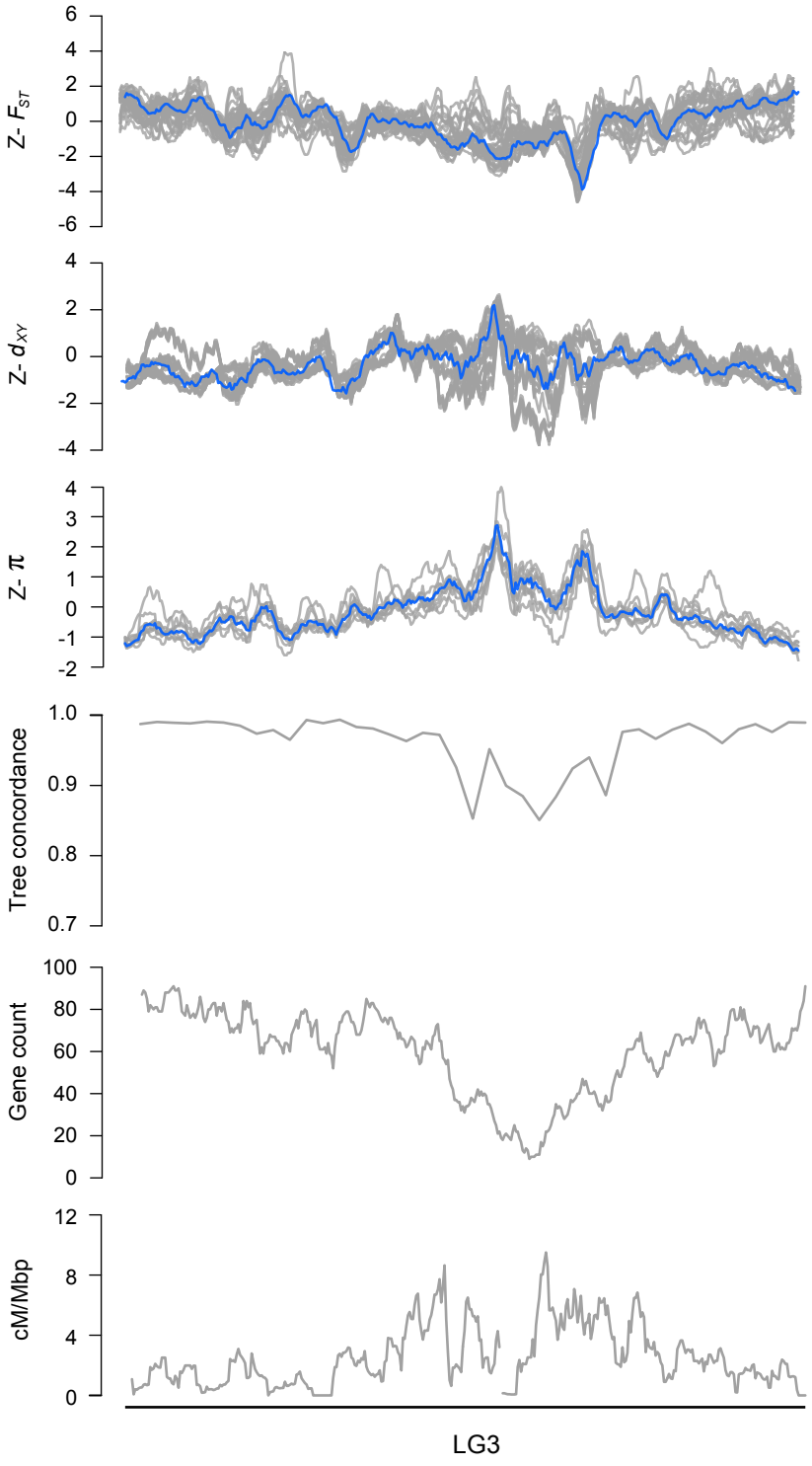



**Figure S5. continued**

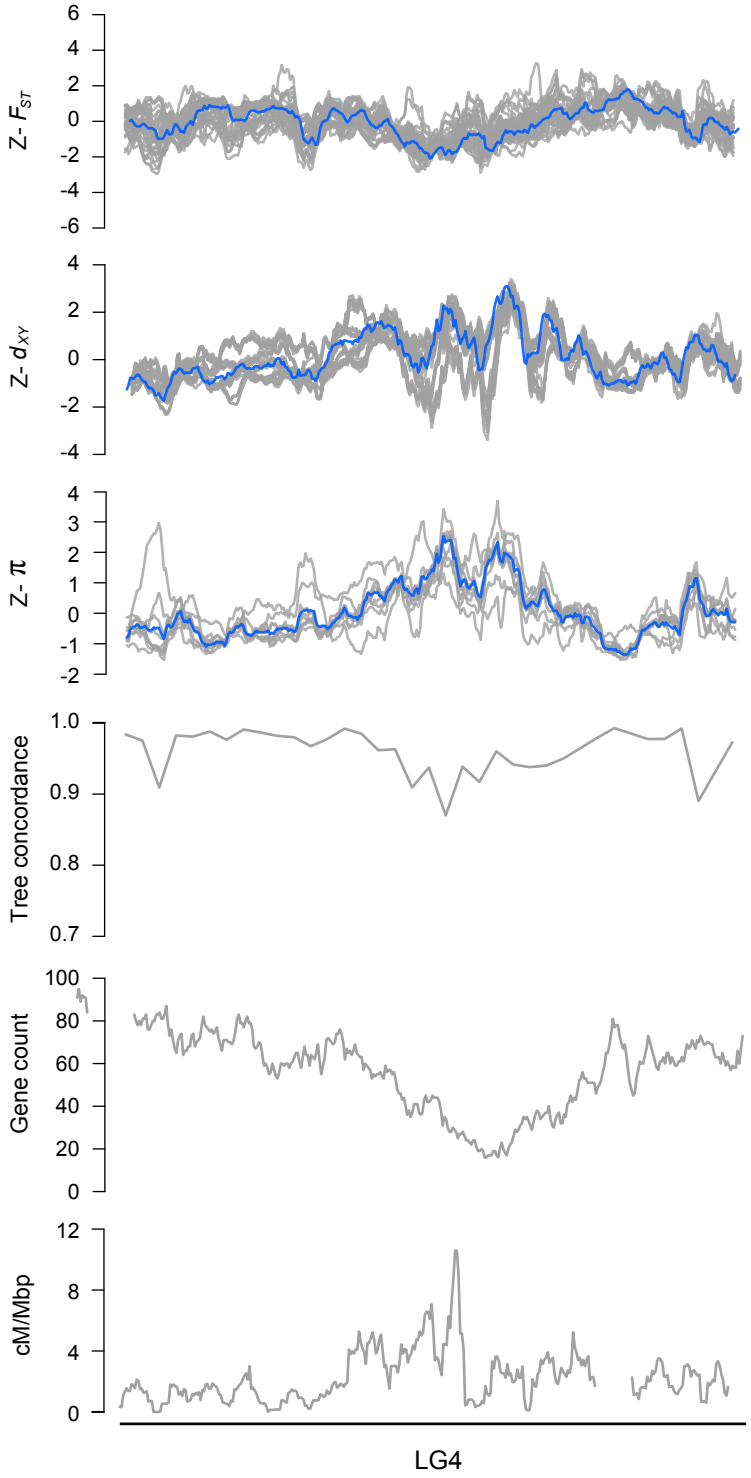



**Figure S5. continued**

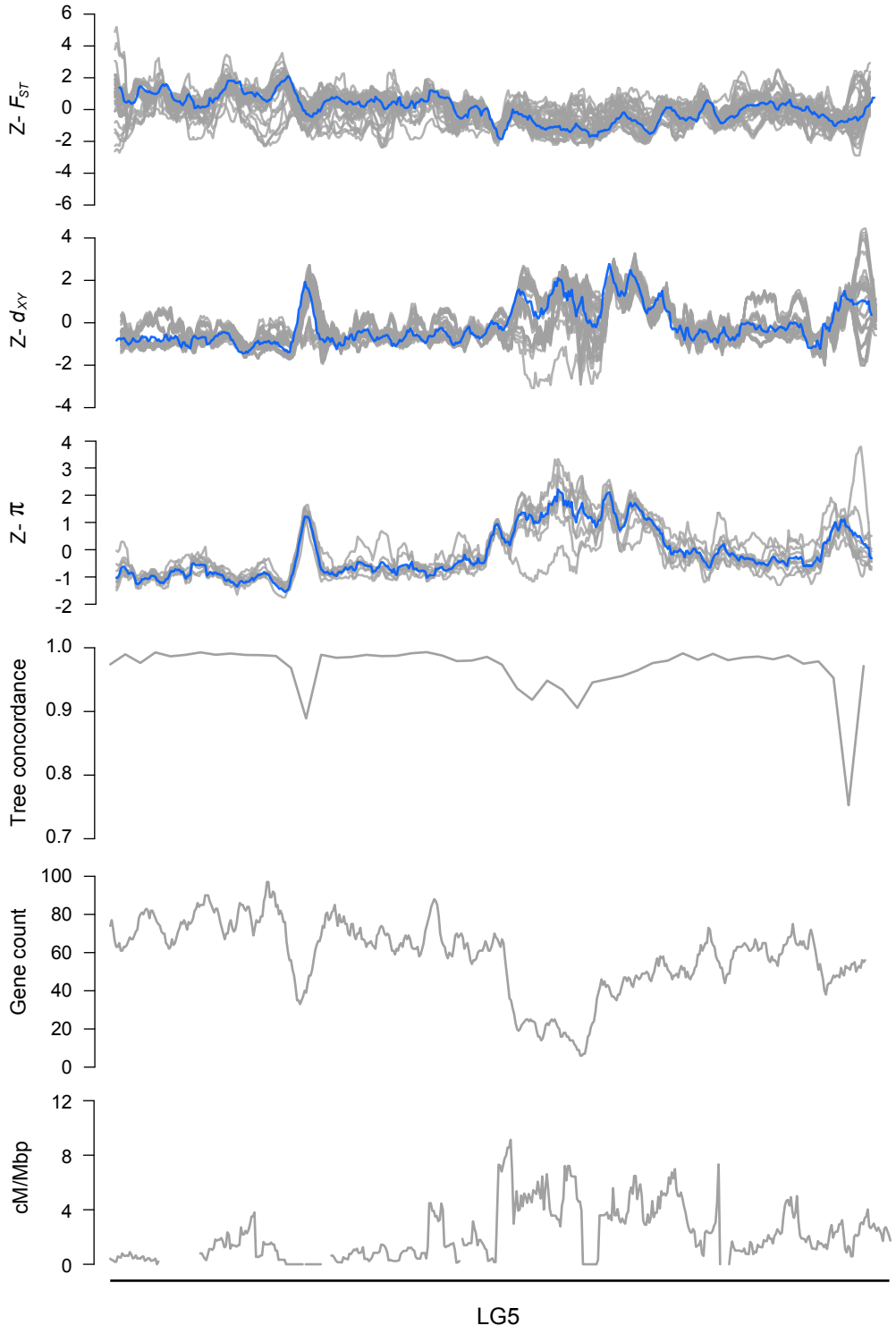



**Figure S5. continued**

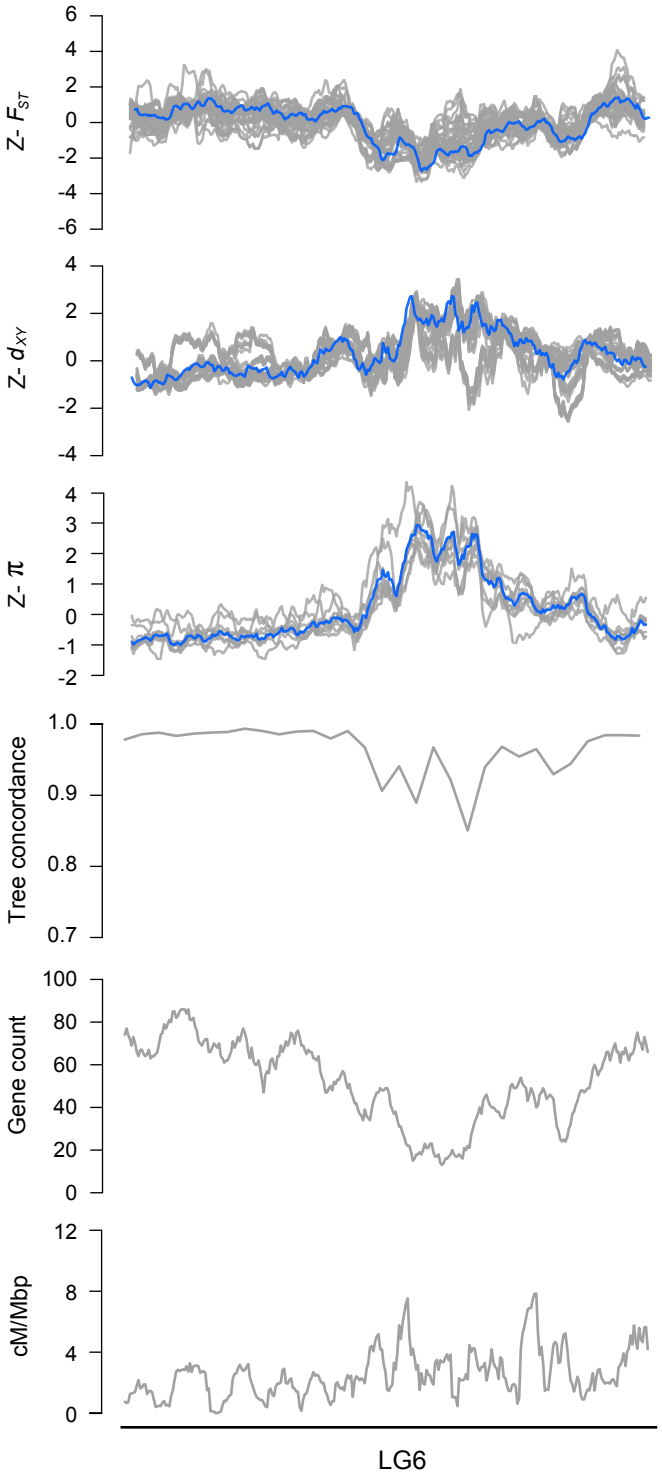



**Figure S5. continued**

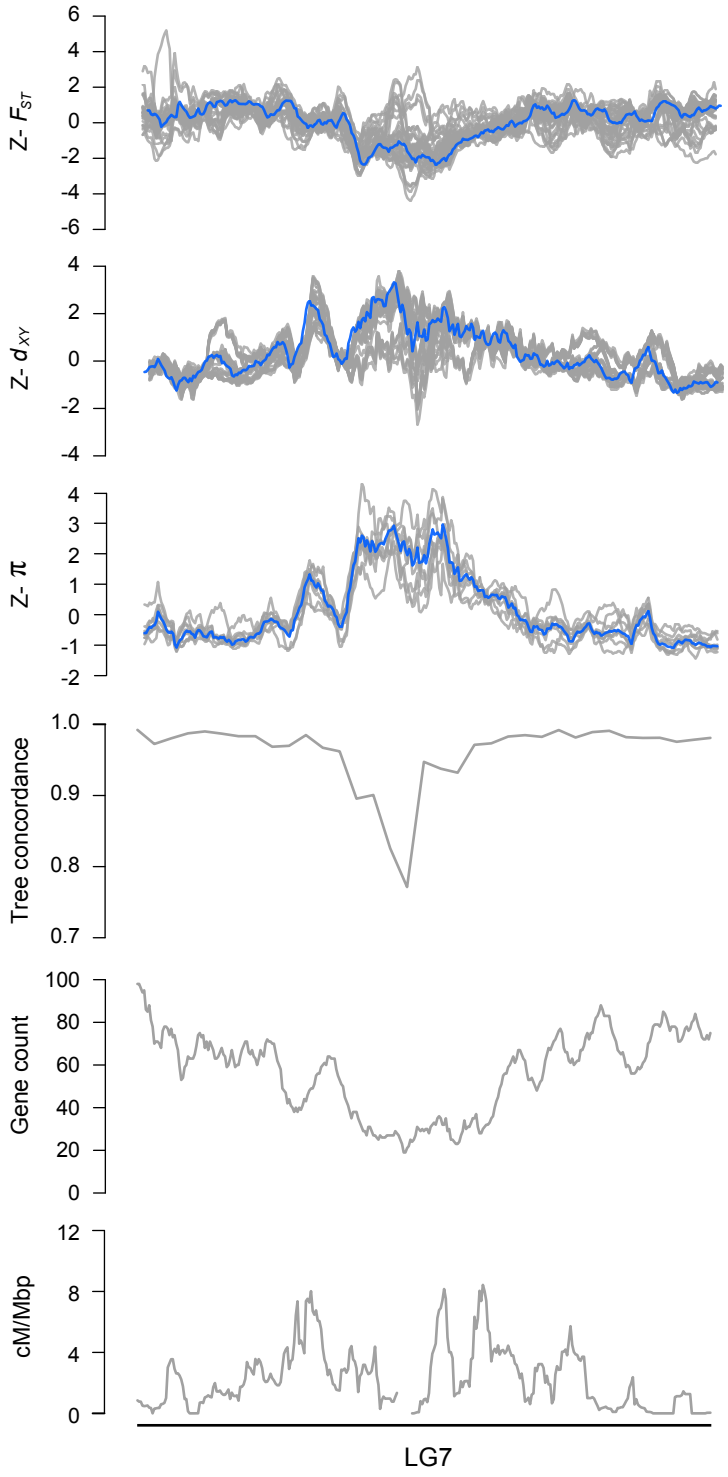



**Figure S5. continued**

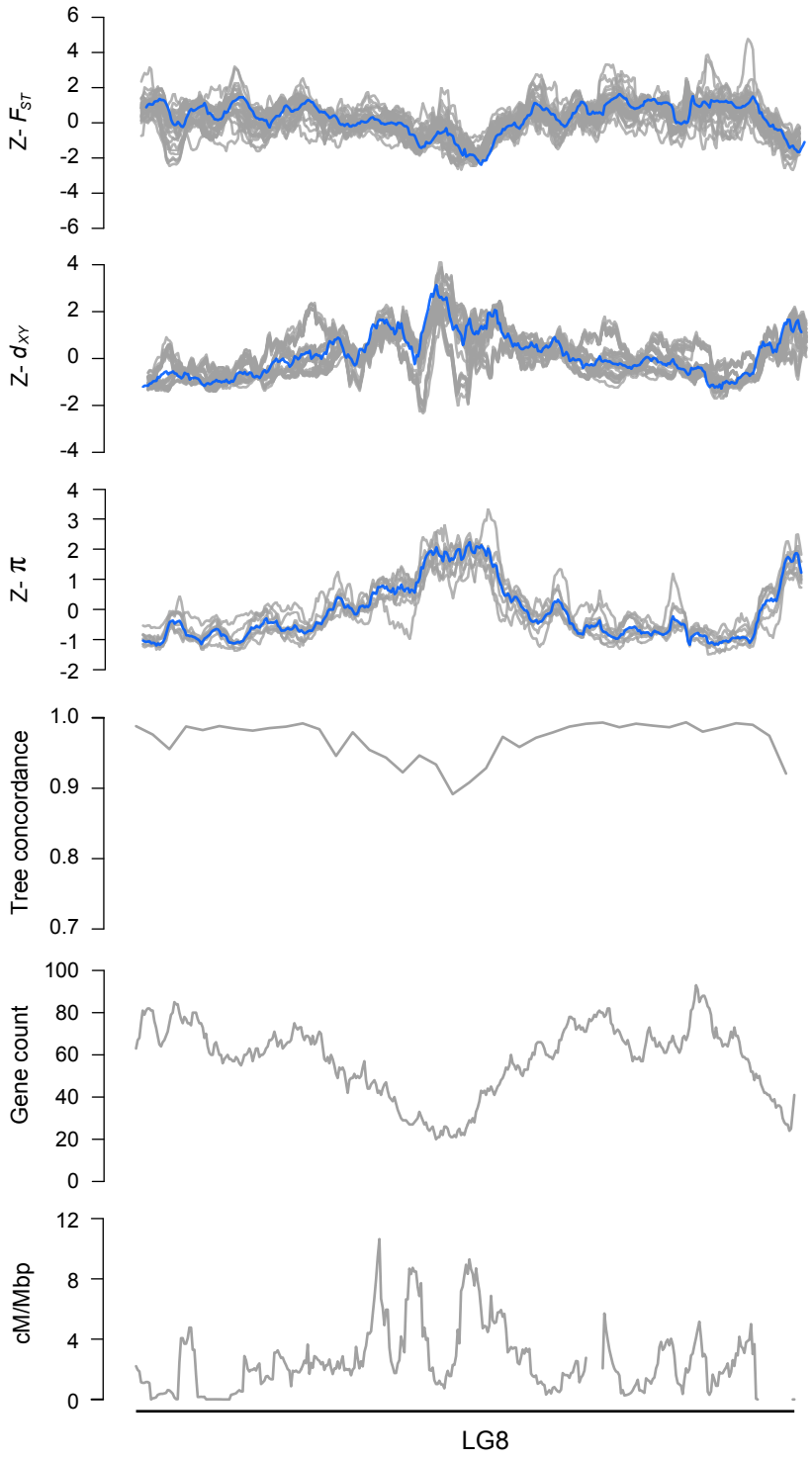



**Figure S5. continued**

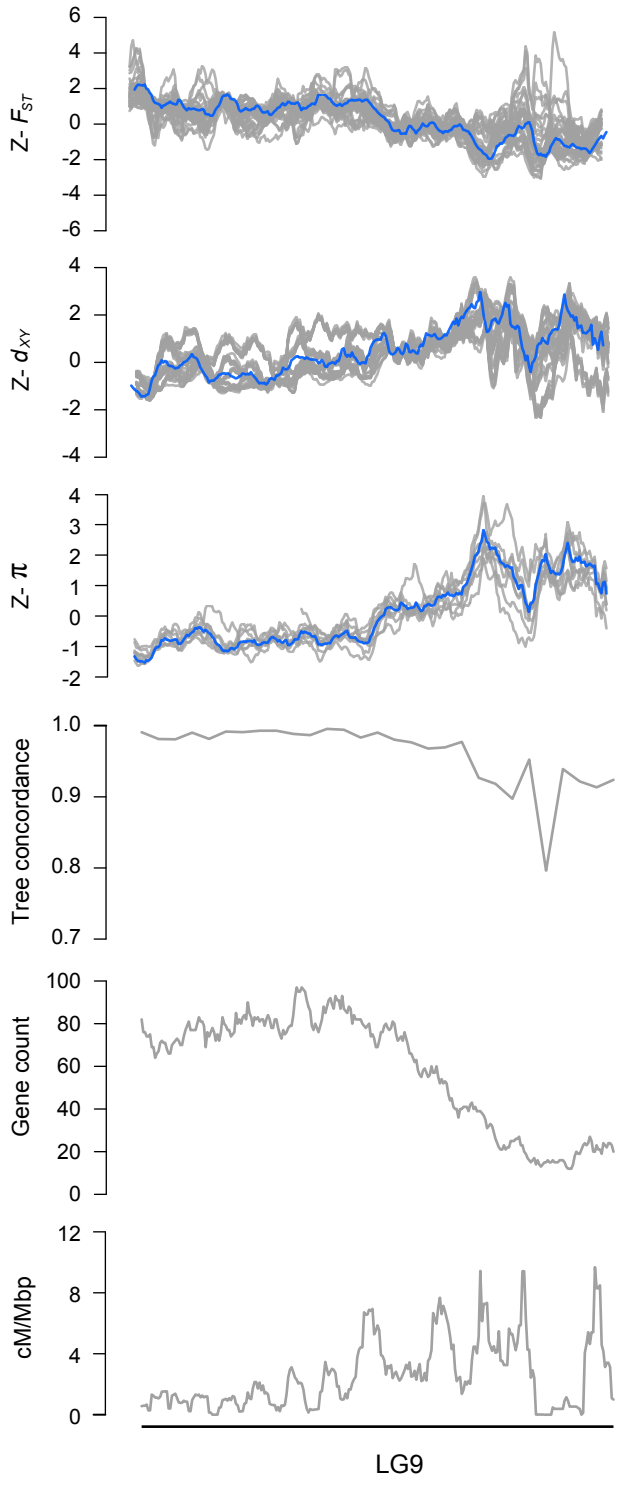



**Figure S5. continued**

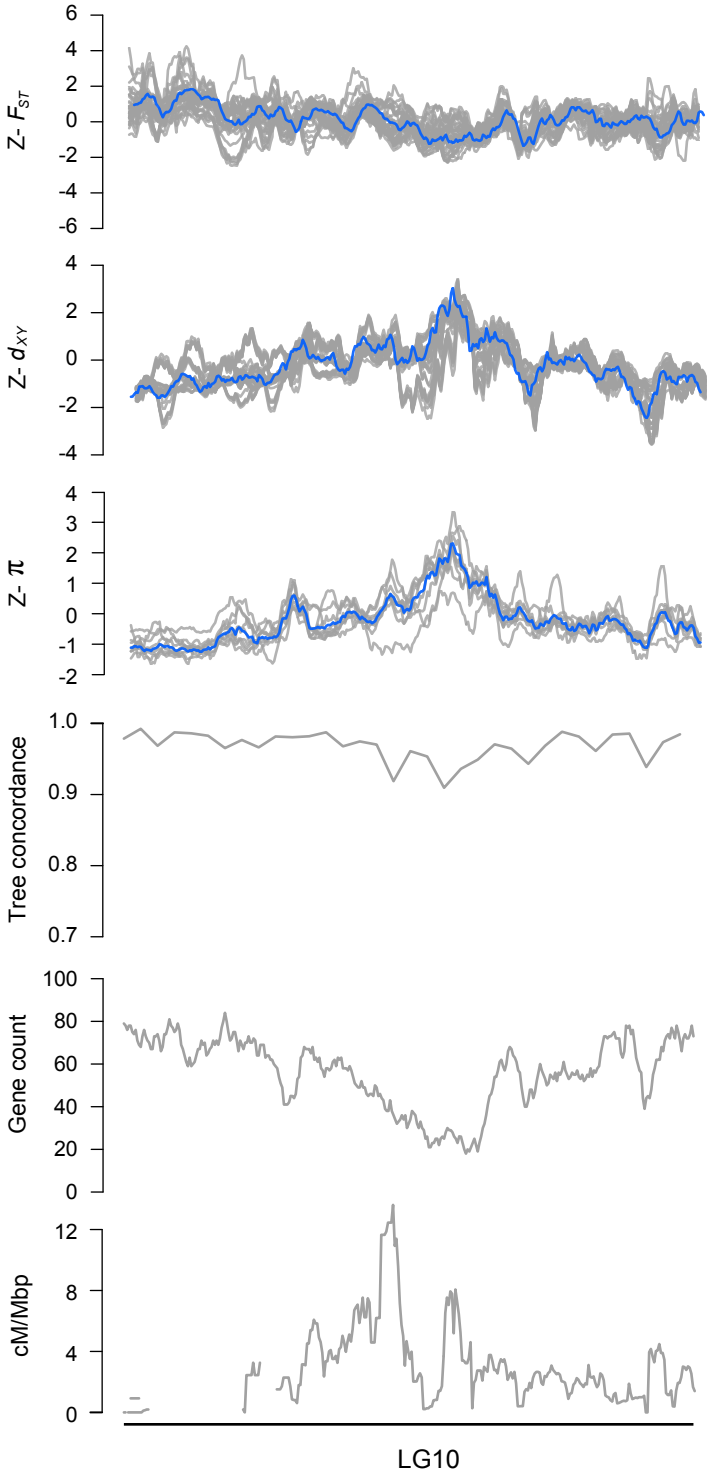



**Figure S5. continued**

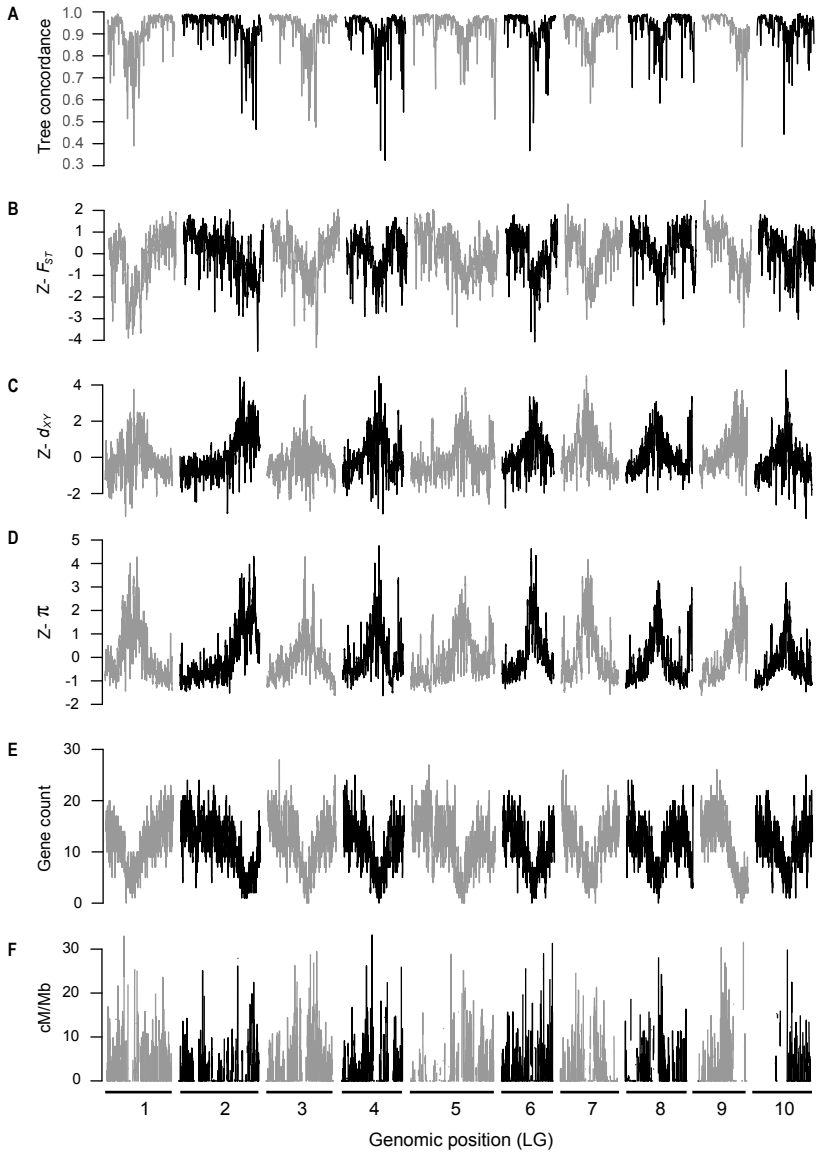



**Figure S6. Common patterns of genome-wide variation mirror variation in the local properties of the genome.** Plots are the same as in Fig. 2 of the main text, but for 100 kb windows (step size 10 kb). A) tree concordance; B-D) Z-transformed PC1 for $F_{ST}$, $d_{xy}$ and π, respectively; E) gene count; and F) recombination rate (cM/Mbp) are ploted across the 10 monkeyflower LGs.

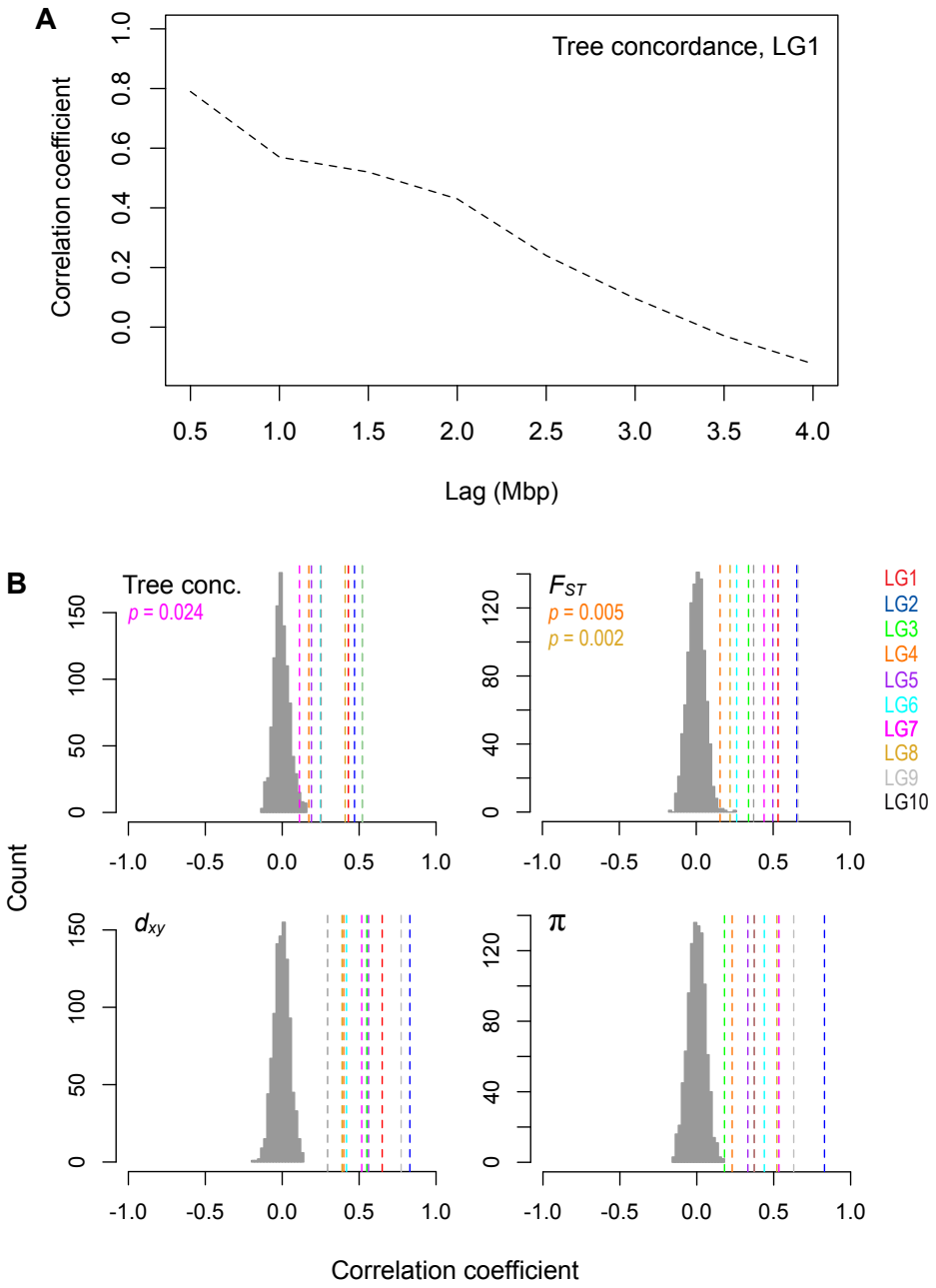



**Figure S7. Patterns of variation are non-randomly distributed across the genome.** A) Strong Autocorrelation of tree concordance scores on LG1 over a Mbp scale. B) Levels of tree concordance, $F_{ST}$, $d_{xy}$, and $\pi$ all show significant autocorrelation at the 2 Mbp scale. The dashed vertical lines show the observed autocorrelation coefficients for each LG with a 2 Mbp lag. The histogram shows the null distribution of autocorrelation coefficients (same lag) generated from 1000 random permutations of the genome-wide values. The observed data are significant at $p = 0.001$ unless stated otherwise.

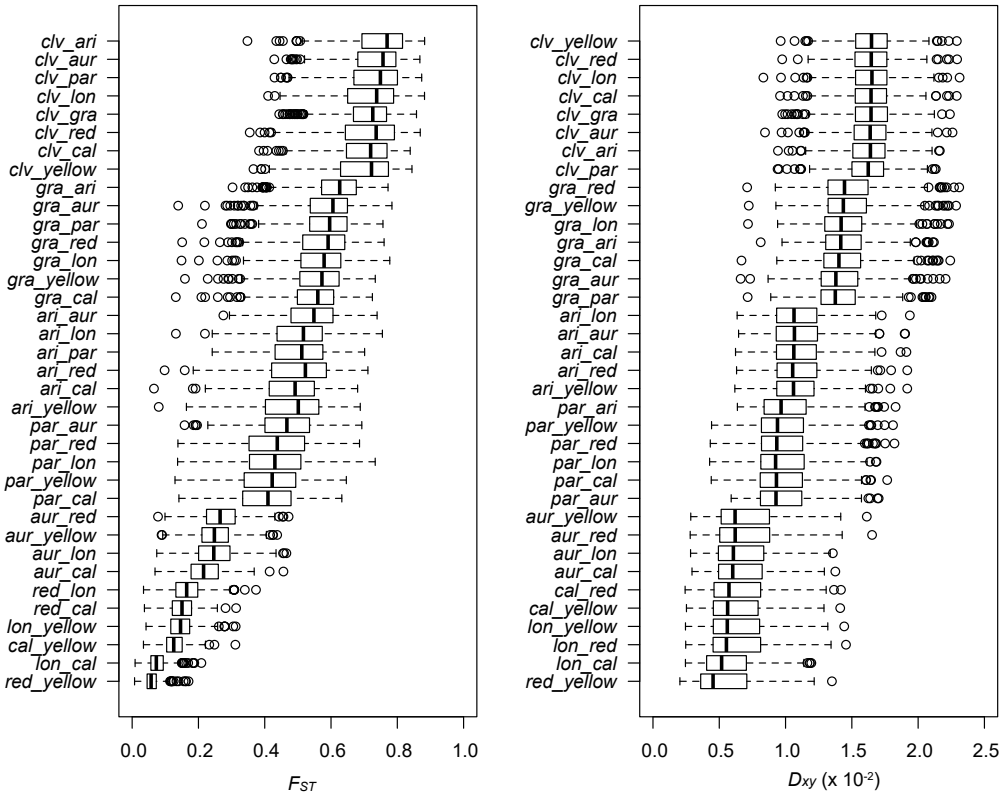



**Figure S8. Patterns of differentiation and divergence for all 36 pairs of taxa.** Box plots for each of the 36 pairwise taxonomic comparisons reveal the range of variation in $F_{ST}$ and $d_{xy}$ across the radiation. Moreover, the data show extensive variance among genomic windows within each comparison. Vertical black lines indicate the median, boxes represent the lower and upper quartiles, and whiskers extend to 1.5 times the interquartile range. Taxon abbreviations: *cal, calycinus; lon, longiflorus; aur, aurantiacus; par, parviflorus; ari, aridus; gra, grandiflorus; clv, M. clevelandii*.

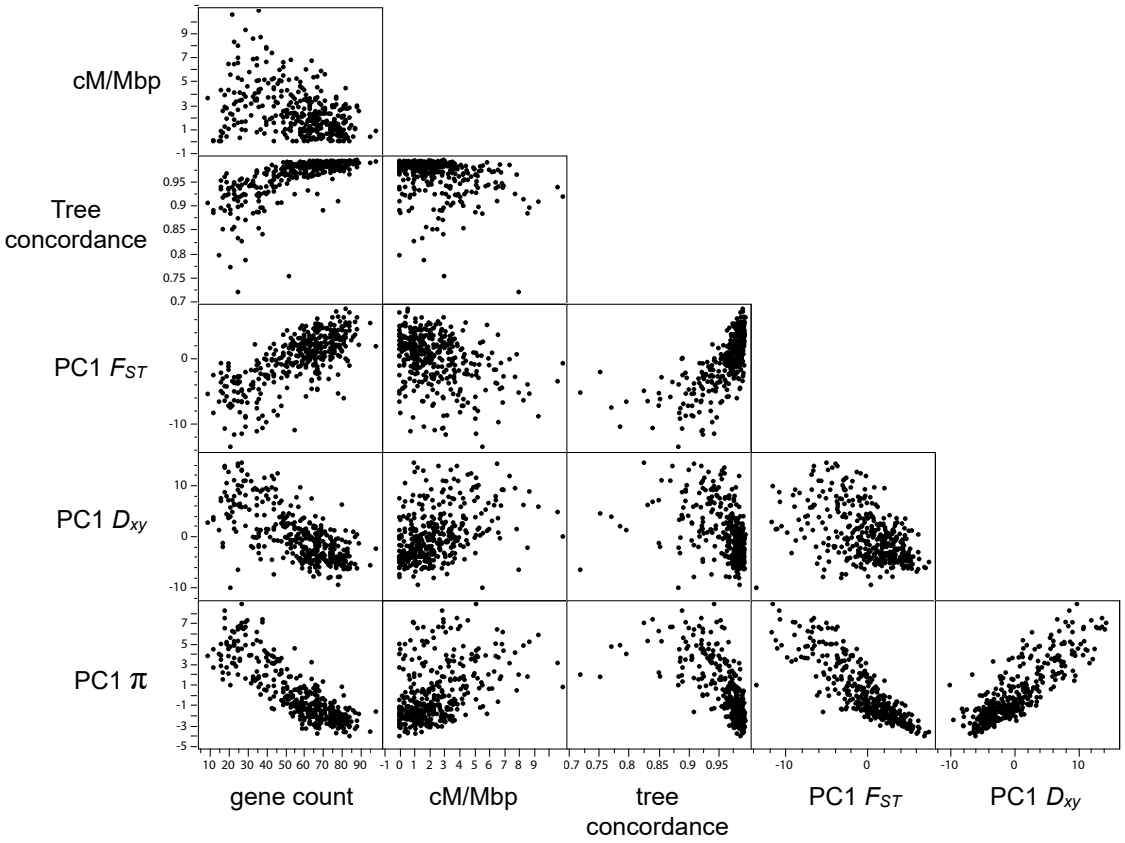



**Figure S9. Bivariate plots among measures of variation and genomic features across 500 kb genomic windows.** Note that this is the same as Figure 3 but with axes units. Also note that the axes are different across rows and columns of the matrix.

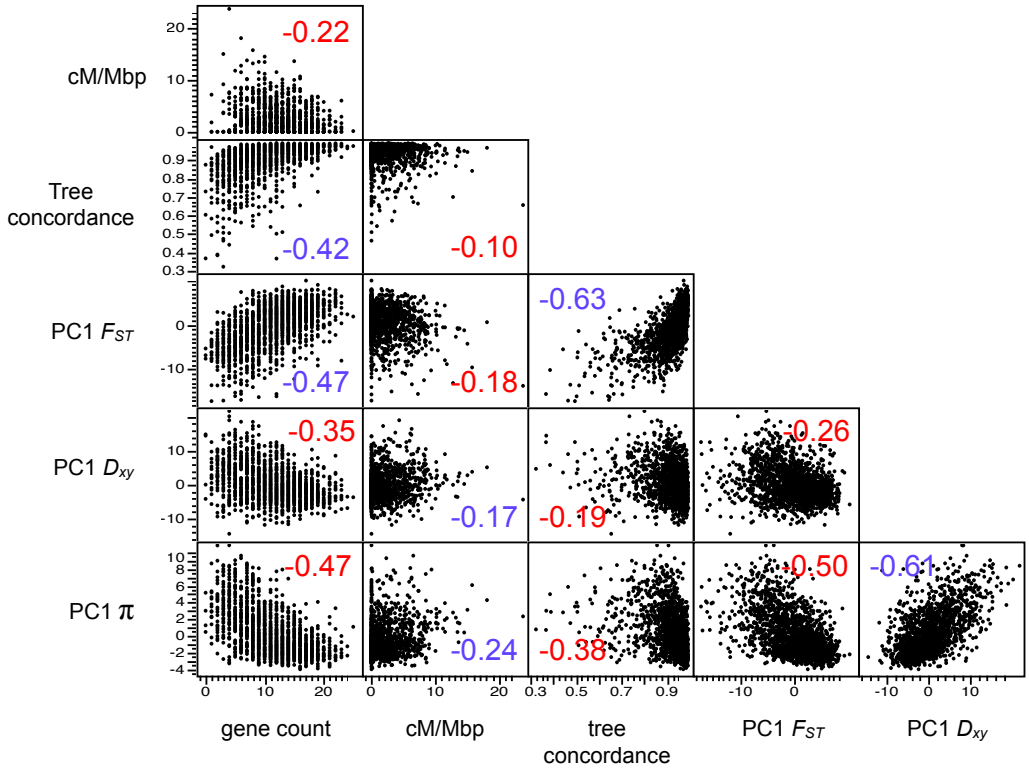



**Figure S10. Bivariate plots among measures of variation and genomic features across 100 kb genomic windows.** The number is the correlation coefficient. Positive correlation coefficients are colored blue and negative coefficients are colored red.

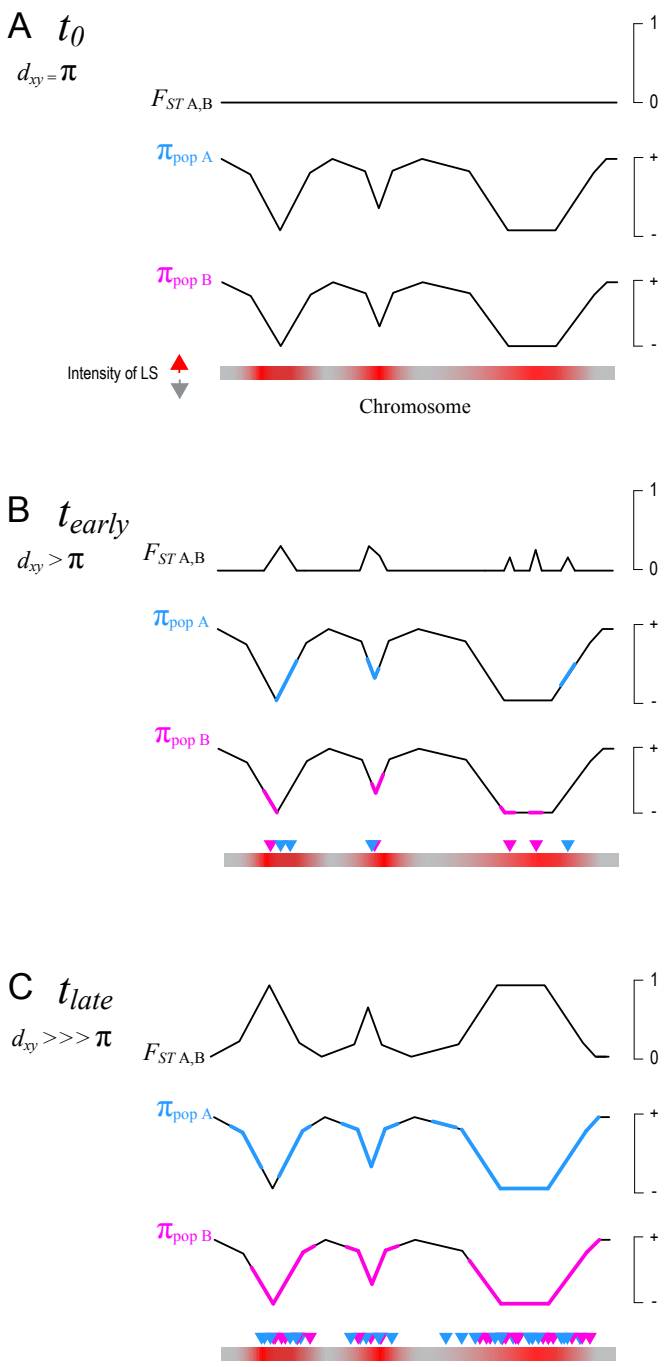



**Figure S11. A cartoon depicting the gradual build-up of a heterogeneous differentiation landscape by lineage-specific LS.** A) When a population first splits (allopatric divergence with large population sizes), patterns of genome-wide diversity ($\pi$ pops A and B) are identical due to the complete sharing of ancestral variation among them. Thus, there is no pattern of heterogeneous differentiation between them. B) As time passes, LS begins to act separately within each lineage (LS events are indicated by arrows across the chromosome). This functions to maintain the diversity landscape in the face of new neutral mutations (affected areas in each lineage are shown in color) and also causes ancestral variants to be fixed among them. The result is an increase in $F_{ST}$ in affected areas. C) Long after the split, many LS events have occurred in each lineage. Because the heterogeneous patterns of LS have been conserved since prior to the common ancestor of these taxa (redder areas of the chromosome experience a higher rate of LS), the affected areas are similar between the lineages. As a result, the diversity and differentiation landscapes come to perfectly mirror one another.

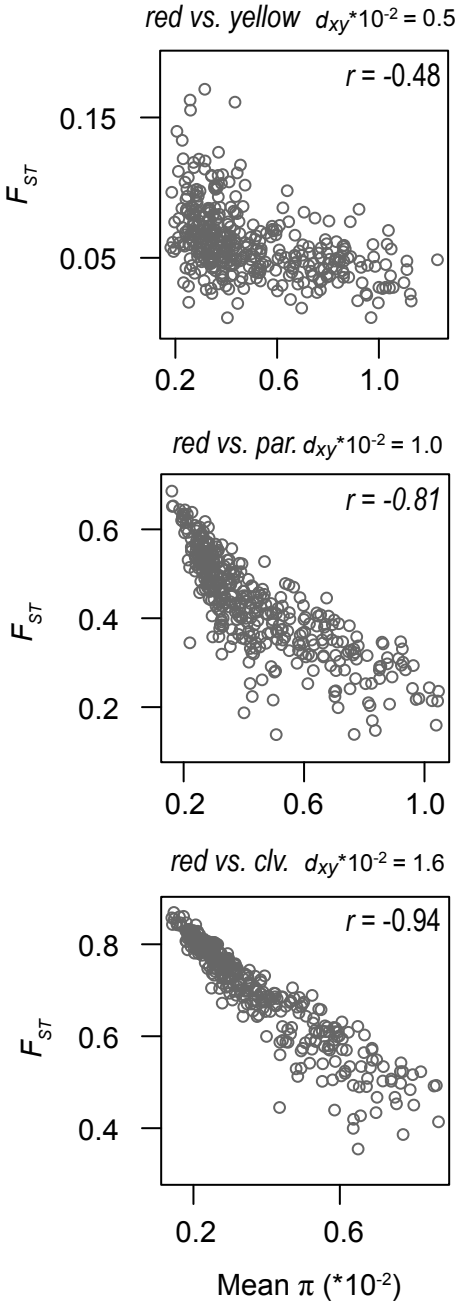



**Figure S12. Negative correlation between nucleotide diversity and differentiation becomes stronger with increasing divergence time.** A) Bivariate plots of the correlation between $F_{ST}$ and π at varying levels of sequence divergence ($d_{xy}$).

Bonus Haiku!

---

Peaks and troughs of $\pi$,

Static, yet ever-changing.

Their reflection grows.

---