Recurrent selection shapes the genomic landscape of differentiation between a pair of host-specialized haplodiploids that diverged with gene flow

Running title: Genomic divergence in pine sawflies

Ashleigh N. Glover[1], Vitor C. Sousa[2], Ryan D. Ridenbaugh[1], Sheina B. Sim[3], Scott M. Geib[3], and Catherine R. Linnen[1]

[1]Department of Biology, University of Kentucky, Lexington, KY 40506, United States

[2]CE3C – Center for Ecology, Evolution and Environmental Changes, Department of Animal Biology, Faculdade de Ciências da Universidade de Lisboa, University of Lison, Lisboa, Portugal

[3]USDA-ARS Daniel K. Inouye US Pacific Basin Agricultural Research Center Tropical Pest Genetics and Molecular Biology Research Unit, Hilo, HI 96270, United States

Corresponding author: Ashleigh N. Glover, angl226@uky.edu

ABSTRACT

Understanding the genetics of adaptation and speciation is critical for a complete picture of how biodiversity is generated and maintained. Heterogeneous genomic differentiation between diverging taxa is commonly documented, with genomic regions of high differentiation interpreted as resulting from differential gene flow, linked selection, and reduced recombination rates. Disentangling the roles of each of these non-exclusive processes in shaping genome-wide patterns of divergence is challenging but will enhance our knowledge of the repeatability of genomic landscapes across taxa. Here, we combine whole-genome resequencing and genome feature data to investigate the processes shaping the genomic landscape of differentiation for a sister-species pair of haplodiploid pine sawflies, *Neodiprion lecontei* and *Neodiprion pinetum*. We find genome-wide correlations between genome features and summary statistics are consistent with pervasive linked selection, with patterns of diversity and divergence more consistently predicted by exon density and recombination rate than the neutral mutation rate (approximated by dS). We also find that both global and local patterns of F_{ST}, d_{XY}, and π provide strong support for recurrent selection as the primary selective process shaping variation across pine sawfly genomes, with some contribution from balancing selection and lineage-specific linked selection. Because inheritance patterns for haplodiploid genomes are analogous to those of sex chromosomes, we hypothesize that haplodiploids may be especially prone to recurrent selection, even if gene flow occurred throughout divergence. Overall, our study helps fill an important taxonomic gap in the genomic landscape literature and contributes to our understanding of the processes that shape genome-wide patterns of genetic variation.

Keywords: genomic landscape, linked selection, speciation genomics, *Neodiprion*, haplodiploid

INTRODUCTION

A core goal in speciation genomics is to use genome-scale data from closely related populations and species to make inferences about the genetic underpinnings, evolutionary mechanisms, and demographic context of speciation (Wolf and Ellegren 2016; Han et al. 2017; Stankowski et al. 2019; Shang et al. 2023). Technological advances over the last two decades have made it possible to characterize the genome-wide landscape of genetic variation in virtually any organism. A deluge of such genomic landscapes has revealed highly heterogeneous patterns of genetic variation across genomes as well as variable landscapes across taxa (e.g., Martin et al. 2013; Burri et al. 2015; Irwin et al. 2016, 2018; Ma et al. 2018; Han et al. 2017; Talla et al. 2019; Stankowski et al. 2019; Sun et al. 2022; Bendall et al. 2022; Jiang et al. 2023; Shang et al. 2023). As more and more data accrue, an emerging challenge is to identify factors that reliably predict genomic landscape characteristics. To this end, three general frameworks have emerged to interpret patterns of genomic variation: (1) exploration of relationships between genome features and genetic summary statistics; (2) exploration of relationships among different genetic summary statistics; and (3) comparison among taxa differing in features that are likely to affect genomic landscapes of differentiation.

The first framework for interpreting genomic landscapes incorporates decades of theoretical work that describe how the interplay of mutation, gene flow, selection, and recombination affects levels of genetic variation and differentiation (Ravinet et al. 2017; Burri 2017a). For example, purifying selection purges deleterious mutations while positive selection fixes advantageous mutations (Charlesworth et al. 1993; Maynard Smith and Haigh 1974; Rodrigues et al. 2023). Importantly, neutral loci that are physically linked to the deleterious and/or beneficial mutations are also affected: the haplotype background on which the mutation

arose is either lost (background selection) or sweeps to high frequency or fixation (selective sweep), resulting in the loss of most or all other haplotypes. Thus, both background selection and selective sweeps, collectively referred to as "hitchhiking" or "linked selection", lead to a loss of local within-species genetic diversity (π) and increased differentiation (F_{ST}) between species (Charlesworth et al. 1993; Charlesworth 1998; Fay and Wu 2000; Via and West 2008; Cutter and Payseur 2013; Rougemont et al. 2019; Han et al. 2017). That said, recent theoretical work suggests that selective sweeps lead to a greater loss of genetic diversity compared to background selection (Matthey-Doret and Whitlock 2019; Schrider 2020).

The abundance of different types of polymorphic sites will also be affected by linked selection. For example, both background selection and selective sweeps are expected to produce a deficit of mid-frequency neutral variants surrounding selected sites (Tajima 1989; Charlesworth et al. 1995; Cutter and Payseur 2013; Burri et al. 2015). Moreover, an expectation unique to selective sweeps is an excess of high-frequency derived variants (Fay and Wu 2000; Burri et al. 2015). Because linked selection is expected to be most pronounced in low-recombination and gene-dense regions of the genome, widespread linked selection predicts genome-wide correlations between genetic variation summaries (π, F_{ST}, d_{XY}, and summaries of the site-frequency spectrum such as Tajima's D and Fay and Wu's H) and both recombination rate and gene density (Table 1A, Cutter and Payseur 2013; Comeron 2017; Rougemont et al. 2019; Stankowski et al. 2019; Chase et al. 2021).

Variable mutation rates can also contribute to heterogeneity in diversity and differentiation across genomes (Ravinet et al. 2017). On the one hand, as the ultimate source of new variation, mutation rate is expected to correlate positively with diversity (Castellano et al. 2019). On the other hand, because most non-neutral mutations are deleterious (Muller 1950),

genomic regions with high mutation rates may be more likely to be targeted by background selection, resulting in low diversity and high differentiation (Ohta 1992; Eyre-Walker and Keightley 2007; Ravinet at al. 2017). Other features of the genomic landscape can have similarly nuanced effects on genetic variation. For example, centromeric regions are often subject to unique selection pressures (Henikoff et al. 2001; Padmanabhan et al. 2008; Fishman and Saunders 2008; Hofstatter et al. 2021) and often have higher mutation rates and lower recombination rates and gene densities than non-centromeric regions (Bensasson 2011; Ravinet et al. 2017). Accordingly, some studies found that within-species genetic diversity is lower and differentiation is higher near centromeres (e.g., Begun et al. 2007; Gore et al. 2009), while other studies found that within-species genetic diversity is higher near centromeres (e.g., Branca et al. 2011; Clark et al. 2007). Overall, many genomic variables—which are also expected to correlate to varying degrees with one another—are predicted to affect genome-wide patterns of genetic variation. Evaluating their contribution to observed genomic landscapes requires high quality, annotated genomes with complementary data for inferring mutation and recombination rates. When such data are available, multiple linear regression can be used to evaluate the explanatory power of each genomic predictor variable relative to other sources of variation for each genetic summary statistic (e.g., Burri et al. 2015; Samuk et al. 2017; Stankowski et al. 2019; Kartje et al. 2020).

The second framework for interpreting patterns of genomic variation examines relationships among genetic summary statistics [differentiation (F_{ST}); between-population pairwise nucleotide distance, commonly referred to as absolute divergence (d_{XY}); and within-species pairwise nucleotide diversity (π)] at both local and genome-wide scales to make inferences about the selection scenario(s) shaping observed genomic landscapes (Table 1B-C).

Due to recombination, each bout of selection should affect only selected sites plus tightly linked sites. Thus, different selection scenarios are expected to give rise to different local patterns of F_{ST}, d_{XY}, and π. Although most genomic landscapes are almost certainly shaped by multiple types of selection, examining genome-wide relationships among genetic summary statistics can reveal whether a particular type of selection scenario has tended to predominate. The four primary evolutionary scenarios considered under this framework are divergence-with-gene-flow, allopatric selection, recurrent selection, and balancing selection (Han et al. 2017; Irwin et al. 2016, 2018; Shang et al. 2023).

In the first scenario, divergence-with-gene-flow, loci that contribute to reproductive isolation, as well as any linked neutral loci, experience restricted gene exchange between the diverging lineages. This results in locally reduced π and elevated F_{ST} (Han et al. 2017). Due to the locally restricted gene flow, average coalescent times between lineages of the two populations are older, leading to locally elevated d_{XY} (Charlesworth 1998; Cruickshank and Hahn 2014). Other parts of the genome are homogenized by gene flow, thereby reducing both F_{ST} and d_{XY} (Irwin et al. 2018). The remaining three scenarios are similar in that they do not explicitly include gene flow, but they differ in the primary selection pressures shaping genome-wide variation.

Under what has been referred to as the allopatric selection scenario, selective sweeps associated with local environmental conditions and background selection occur independently in geographically separated populations. This results in local reductions of π and elevated F_{ST} at selected and linked neutral loci (Irwin et al. 2016; Han et al. 2017). In contrast to F_{ST}, d_{XY} should be unaffected by reduced π caused by selection in isolated populations (Nachman and Payseur 2012; Burri 2017b). For this reason, d_{XY} is expected to be similar on average between areas of

high F_{ST} and low F_{ST} under the allopatric selection scenario (Cruickshank and Hahn 2014; Han et al. 2017; Irwin et al. 2018).

The last two scenarios, recurrent selection and balancing selection, are expected to produce genome-wide correlations among summary statistics in the same directions (Table 1B), but they produce very different local patterns of F_{ST}, d_{XY}, and π (Table 1C, Shang et al. 2023). Under the recurrent selection scenario, standing genetic variation is reduced at some regions of the genome via selective sweeps or background selection in the common ancestor. After the common ancestor splits into two independent lineages, these same regions that were under selection in the ancestral population are also targeted by selection in the descendant lineages (Irwin et al. 2018). This scenario is expected to produce localized reductions of π and elevated F_{ST} due to selection in descendant lineages. The recurrent selection scenario is also expected to produce localized reductions in d_{XY} due to selection in the common ancestor, which reduces coalescence times between descendant lineages (Han et al. 2017; Irwin et al. 2018; Stankowski et al. 2019). These effects will be most pronounced in gene-dense and low-recombination regions, predicting correlations between d_{XY} and these genome features that are unique to linked selection in ancestral populations (Table 1A).

Finally, under the balancing selection scenario, ancestral polymorphism is maintained in the two descendant lineages (Charlesworth 2006). Under this scenario, selected loci will have elevated π and reduced F_{ST}. Maintenance of ancestral polymorphism in the descendant lineages will also increase coalescence times, causing elevated d_{XY} at loci with a history of balancing selection. Importantly, this balancing selection scenario does not refer to lineage-specific balancing selection, which does not necessarily generate elevated d_{XY}, or the case where divergent ancestral haplotypes are sorted rather than maintained in the descendant lineages,

which would cause elevated d_{XY} and F_{ST}, thereby mirroring the pattern produced by differential gene flow (Han et al. 2017; Guerrero and Hahn 2017).

The third framework for interpreting genomic landscapes uses a comparative approach to identify whether patterns of genomic variation differ consistently across taxa and divergence time points (e.g., Burri et al. 2015; Stankowski et al. 2019; Chase et al. 2021; Shang et al. 2023) and/or between autosomes and sex chromosomes (e.g., Irwin et al. 2016; Wong Miller et al. 2017; Talla et al. 2019; Fiteni et al. 2022). For example, many studies have found that patterns of variation differ markedly between sex chromosomes and autosomes: on sex chromosomes, π is usually lower, and thus F_{ST} is usually higher, which could simply reflect the lower effective sizes of sex chromosomes (e.g., Nachman and Payseur 2012; Irwin et al. 2016; Wong Miller et al. 2017; Talla et al. 2019; Fiteni et al. 2022; Moreira et al. 2023). Yet, this observation is consistent with the hypothesis that sex chromosomes are predisposed to experience recurrent selection because all recessive mutations are expressed in the heterogametic sex, increasing the probability that they will fix (if beneficial) or be purged (if deleterious) (Charlesworth et al. 1987; Ellegren et al. 2012; Oyler-McCance et al. 2015; Irwin et al. 2016; Miller and Sheehan 2023). Haplodiploid inheritance patterns resemble those of sex chromosomes: males develop from unfertilized eggs and are haploid across their entire genome (not recombining), while females develop from fertilized eggs, are diploid and recombine (Nouhaud et al. 2020). Thus, like sex chromosomes, haplodiploids may be especially prone to experience repeated bouts of linked selection, even if there is gene flow throughout divergence. As approximately 15% of all invertebrates are haplodiploid (de la Filia et al. 2015; Blackmon et al. 2017), this taxon-specific factor may have a widespread effect on patterns of genomic divergence in nature. However, compared to diploid organisms, very few genomic landscapes are currently available for

haplodiploid organisms (but see Wallberg et al. 2015; Christmas et al. 2021; Mozhaitseva et al. 2023; Everitt et al. 2023). As a first step to evaluating the broader effects of haplodiploid inheritance on genome-wide patterns of genetic variation, we combine whole-genome resequencing and genome feature data from a sister-species pair of haplodiploid pine sawflies, *Neodiprion lecontei* and *Neodiprion pinetum*.

For several reasons, *N. lecontei* and *N. pinetum* are an outstanding model for speciation genetics and genomics (Figure 1). Like all *Neodiprion* (Order: Hymenoptera; Family: Diprionidae), these sister species are conifer feeders that depend on their host plant for all stages of their life cycle (Coppel and Benjamin 1965; Knerer and Atwood 1973; Linnen and Farrell 2008a; Herrig et al. 2024). *Neodiprion lecontei* and *N. pinetum* have largely overlapping ranges in eastern North America (Linnen and Farrell 2008a, 2010; Glover et al. 2023), and a recent demographic analysis suggests that they diverged in sympatry with continuous gene flow (Bendall et al. 2022). Although the two species are frequently found at the same geographical locations, they are adapted to different pine species with dissimilar needle morphologies. While *N. pinetum* specializes on the thin-needled white pine (*Pinus strobus*), *N. lecontei* avoids white pine and uses a variety of *Pinus* species that have thicker needles (Wilson et al. 1992; Linnen and Farrell 2010; Bendall et al. 2017). In addition to their divergent host preferences, *N. lecontei* and *N. pinetum* also differ in adult body size, female ovipositor morphology, and additional egg-laying traits (Bendall et al. 2017; Glover et al. 2023). These divergent host-use traits contribute to both prezygotic and postzygotic barriers to gene flow (Bendall et al. 2017, 2023; Glover et al. 2023) and map to many regions across the genome (Bendall 2020). Although reproductive isolation between *N. lecontei* and *N. pinetum* is strong, it is incomplete: these species occasionally hybridize in nature and viable, fertile hybrids for both sexes and cross directions

can be produced in the lab (Bendall et al. 2017, 2022, 2023). Finally, in addition to their experimental tractability and well-characterized ecological differences and demographic history, *N. lecontei* and *N. pinetum* have excellent genomic resources, including annotated, chromosome-level genome assemblies for both species and a high-quality recombination map for *N. lecontei* (Linnen et al. 2018; Vertacnik et al. 2023; Herrig et al. 2024).

As host-specialized haplodiploids that diverged with gene flow and have complementary genome-feature data available, *N. lecontei* and *N. pinetum* offer a unique opportunity to evaluate how the demography and ecology of speciation, haplodiploid transmission genetics, and genome features shape the genomic landscape of differentiation. To this end, we first use multiple linear regression to determine which genomic features predict genetic diversity and differentiation and to ask whether the data are consistent with pervasive linked selection (Table 1A). We then examine both genome-wide and local patterns of genetic summary statistics to determine which, if any, of the four selection scenarios have shaped patterns of variation within and between *N. lecontei* and *N. pinetum* (Table 1B and 1C). On the one hand, based on their ecological differences and divergence history, we might expect our genomic data to exhibit patterns consistent with divergence-with-gene-flow. On the other hand, due to the increased efficacy of selection in hemizygous males (Avery 1984; Charlesworth et al. 1987; Presgraves 2018; Bendall et al. 2022), we might instead expect our genomic data to support a recurrent selection scenario. A third possibility is that a mixture of selection scenarios obscures genome-wide patterns but is evident in local patterns of variation (Irwin et al. 2016, 2018; Stankowski et al. 2019; Shang et al. 2023). When interpreted in light of these *a priori* predictions, our findings have implications for the predictability of genomic landscapes of differentiation, which we consider further in the discussion.

MATERIALS AND METHODS

Sampling, sequencing and read processing

We collected *Neodiprion lecontei* and *N. pinetum* mid- to late-instar larval colonies from Lexington, Kentucky and surrounding areas (Table S1). To confirm sex (and therefore ploidy), we reared the larvae to adults in the lab using standard lab protocols (Harper et al. 2016; Bendall et al. 2017). The adults were either preserved in 100% ethanol (stored at -20 °C) or flash frozen and stored at -80 °C. To avoid sampling close relatives, we selected one individual from each larval colony (each colony typically represents a group of siblings). In total, we sampled 20 *N. lecontei* females and 18 *N. pinetum* females. We extracted DNA from head and thorax tissue with a Qiagen DNeasy Blood & Tissue Kit. We followed the standard manufacturer protocol for insects, including an optional RNase A step. DNA concentration was measured with a Quant-iT dsDNA High-Sensitivity fluorescence assay (Invitrogen). A single library was prepared using a Tn5 tagmentation protocol following Bendall (2020). Whole-genome resequencing was performed with 150bp paired-end sequencing technology in an Illumina Hi-Seq X sequencer at Admera Health (Plainfield, NJ). The library was first run in a single lane of the sequencer. A subset of samples that had low read counts were then re-pooled and run on a second lane.

Demultiplexed reads were cleaned using trimmomatic v0.39 (Bolger et al. 2014) with the following criteria: (a) remove adapters, (b) perform sliding window trimming where a sequence is cut when a window (4bp) drops below a quality threshold (15), (c) remove low quality bases (quality score < 3) from the beginning and end of each read. Then, for each lane separately, the cleaned reads were mapped to the high-quality chromosome-level *N. lecontei* reference genome (iyNeoLeco1.1, GCA_021901455.1; Herrig et al. 2024) using the BWA-MEM algorithm in bwa

v0.7.17 (Li and Durbin 2009). We used samtools v1.13 (Li et al. 2009; Danecek et al. 2021) to mark PCR duplicates ('markdup') and remove ambiguously mapped reads/secondary alignments (*'view -F 1284 -f 0x02'*). We then used samtools to merge the filtered reads from the two lanes ('merge') and index the resulting bam files ('index').

Estimates of summary statistics

All genetic summary statistics were calculated in ANGSD v0.933 (Korneliussen et al. 2014), a program that is suitable for low coverage whole-genome resequencing data because it incorporates genotype likelihoods (rather than hard-called genotypes) and information contained in the site frequency spectrum (which contains variant and invariant sites) to calculate summary statistics following equations in Fumagalli et al. (2013) and Korneliussen et al. (2013). To estimate F_{ST} across the genome using the Hudson et al. (1992) estimator, we first created a sample allele frequency file for each species (*'-dosaf 1 -uniqueOnly 1 -remove_bads 1 -only_proper_pairs 1 -trim 0 -baq 1 -minMapQ 20 -minQ 20 -setMinDepth 14 (*for *N. lecontei*)*/11* (for *N. pinetum*) *-setMaxDepth 210 (*for *N. lecontei*)*/165* (for *N. pinetum*) *-minInd 14 (*for *N. lecontei*)*/11* (for *N. pinetum*) *-doCounts 1 -GL 1'*). We included the genome of another *Neodiprion* species (*N. virginiana*; Herrig et al. 2024) as an outgroup in the ANGSD runs so that we could polarize SNPs and generate unfolded site frequency spectra. The resultant outputs were used to generate an unfolded site frequency spectrum for each species separately as well as an unfolded pairwise joint site frequency spectrum using the 'realSFS' function. The sample allele frequency files and joint site frequency spectrum were then used as input files to calculate per-site F_{ST} using 'realSFS', which was then used as an input file to calculate F_{ST} in 50kbp non-overlapping windows (*'-win 50000 -step 50000'*). We chose this window size because it is larger

than the distance at which linkage disequilibrium has decayed to approximately zero (Figure S1). These and all other scripts for subsequent analyses can be found on DRYAD (Glover et al. 2024).

To estimate window-based within-species pairwise nucleotide diversity (π), Tajima's D, and Fay and Wu's H across the genome, we first performed ANGSD runs for each species separately to calculate per-site "pairwise differences" theta. These runs included the same parameters above but with an additional '*-doThetas 1*' command and the unfolded site frequency spectrum generated for the F_{ST} analysis as input ('*-pest*'). The resulting per-site theta files were then used to calculate 50kbp windowed statistics (*'-win 50000 -step 50000'*) using the 'thetaStat' function.

To estimate d_{XY} across the genome, we used a custom script by Josh Peñalba (https://github.com/mfumagalli/ngsPopGen/blob/master/scripts/calcDxy.R) with the following modification: we removed the SNP-calling flags in the ANGSD runs so that invariant sites were included in addition to variant sites. We first calculated allele frequencies in ANGSD (*'-uniqueOnly 1 -remove_bads 1 -only_proper_pairs 1 -trim 0 -baq 1 -minMapQ 20 -minQ 20 -setMinDepth 25 -setMaxDepth 375 -minInd 25 -doCounts 1 -GL 1 -doMajorMinor 1 -doMaf 1 -doGlf 4'*) and then used the resultant output as input for the custom R script. We averaged the resulting per-site d_{XY} across 50kbp windows in R version 4.1.0 (R Core Team 2021).

To ensure that our results and conclusions were not biased by choice of window size, we also calculated F_{ST}, d_{XY}, and π in 25kbp and 100kbp non-overlapping windows. Finally, to reduce biases due to poor mapping/genotyping error, we filtered out windows where the number of called sites (invariant + variant) fell below the 10^{th} percentile (i.e., windows that had < 921 sites). We chose this cutoff based on the distribution of site counts per window (Figure S2).

Finally, note that of all summary statistics considered, only Fay and Wu's H depends on polarizing ancestral states. Hence, misidentification of ancestral and derived states is not expected to affect our conclusions, which was confirmed by repeating the analyses with the folded SFS for all other summary statistics.

Genome features

We measured recombination rate, exon density (a measure of gene density), and synonymous substitution rate (dS; a proxy for the neutral mutation rate) in each 50kbp window and measured the distance of each window from the centromere. We obtained recombination rates (cM/Mb) from a previously published high density linkage map generated from a cross between divergent *N. lecontei* populations (Linnen et al. 2018; Herrig et al. 2024).

To estimate exon density, we first extracted all exons from the *Neodiprion lecontei* genome annotation (GCF_021901455.1_iyNeoLeco1.1_genomic.gtf; Herrig et al. 2024). Because many genes have more than one transcript, we retained the transcript with the most exons. We then used a custom R script from Samuk et al. (2017; https://github.com/ksamuk/gene_flow_linkage/blob/master/ev_prep_scripts/gene_density_calc_build.R) to calculate the proportion of each 50kbp window containing exon sequence.

To estimate mutation rate, we calculated the synonymous substitution rate (dS) in 50kbp windows. We first inferred orthologous gene groups between *Neodiprion lecontei*, *N. pinetum*, and *N. virginiana* using Broccoli v1.2 (Derelle et al. 2020). For this analysis, we used annotated genes from reference-quality genomes for the three species (GCF_021901455.1_iyNeoLeco1.1_genomic.gff; GCF_021155775.1_iyNeoPine1.1_genomic.gff; GCF_021901495.1_iyNeoVirg1.1_genomic.gff;

318 Herrig et al. 2024). Genes from these three species were matched to 10,686 orthogroups using
319 maximum likelihood ('-phylogenies ml'). We excluded 24 orthogroups that were located on
320 unplaced scaffolds prior to downstream analysis. Before alignment, orthogroups were checked
321 for the presence of multiple isoforms. If multiple isoforms were present, only the longest isoform
322 for each species was retained. Then, filtered orthogroups were aligned using the L-INS-I method
323 in MAFFT v7.509 (Katoh and Standley 2013). With these aligned sequences as input, we
324 calculated the synonymous substitution rate for each orthogroup using codeml with model = 0
325 and Nsites = 0 (one omega ratio for all branches) in PAML v4.10.6 (Yang 2007). We then
326 filtered out orthogroups where the estimated dS for any species was greater than or equal to two
327 standard deviations above the mean. Next, we used a custom script from Samuk et al. (2017;
328 https://github.com/ksamuk/gene_flow_linkage/blob/master/ev_prep_scripts/ds_dn_7_recomb_an
329 alysis.R) to assign the mutation rate estimates to 50kbp windows for each species. Finally, we
330 calculated the average dS between *N. lecontei* and *N. pinetum* for each window to produce a
331 single mean dS value for each window.

332 To estimate distance from the centromere, we first used Juicebox v1.11.08 (Durand et al.
333 2016) to visualize the HiC contacts for each chromosome (Figures S3-S9, Herrig et al. 2024) and
334 estimated the midpoint of each centromere by identifying the local maximum delta in the number
335 of contacts between adjacent loci within each chromosome. These points were also identified
336 visually and were supported by depressed levels of repeat density, HiC read coverage, GC
337 content, and gene density (Figures S3-S9). We then calculated the distance of each window from
338 the centromere by taking the absolute value of the midpoint of each window subtracted from the
339 midpoint of the centromere.

Finally, we also wanted to consider the potential effect of local genotyping error on patterns of genetic variation. Due to variation in base composition, repetitive sequence content, and sequence divergence, some regions of the genome are more prone to sequencing, mapping, and genotyping error. As a rough metric for this error, we used site counts for each 50kbp window, which are the number of invariant and variable sites that were called after quality filtering. Because ANGSD directly calculates windowed F_{ST}, π, Tajima's D, and Fay and Wu's H as well as provides the called site count for each window (i.e., the number of sites used to calculate the statistic in each window), site counts were taken directly from ANGSD outputs for each of these windowed summary statistics. However, ANGSD does not directly calculate d_{XY}. Therefore, to obtain the number of called sites used to calculate d_{XY} in each window, we used a custom R script that takes the per-site d_{XY} file and counts the number of called sites in each 50kbp window. Our assumption here is that more error-prone regions of the genome would have lower site counts. As noted above, windows with the lowest site counts were filtered out prior to analysis, but the remaining windows still exhibited substantial variation in site counts.

Correlation and regression analyses of summary statistics and genome features

To explore evolutionary processes shaping genome-wide patterns of genetic variation, we first examined the genome-wide correlations among differentiation, absolute divergence, diversity, and features of the genome. We estimated Pearson's correlation coefficients between pairs of these statistics and calculated *p*-values using the *correlation* v0.7.1 package (Makowski et al. 2020) in R version 4.1.0 (R Core Team 2021); *p*-values were adjusted for multiple testing using the Holm (1979) method. We then used the *corrplot* v0.90 package (Wei and Simko 2021) to visualize the correlation matrix. Second, to investigate relationships between summary

363 statistics and genome features, we used a multiple regression approach in R version 4.1.0 (R
364 Core Team 2021). We first normal-quantile transformed all predictor variables to ensure they
365 were on the same scale. Then, using the 50kbp windows as data points, we fit multiple linear
366 regression models ("lm" function) for our genetic summary statistics (F_{ST}, d_{XY}, π, Tajima's D,
367 and Fay and Wu's H) using the following form: summary statistic ~ recombination rate + exon
368 density + mutation rate + distance from the centromere + window site count. After fitting each
369 model, we used a type II analysis of variance (ANOVA) implemented in the *car* v3.1.0 package
370 (Fox and Weisberg 2019) to evaluate the significance of model terms.
371
372 **Local patterns of summary statistics**
373 In addition to the global analyses, we also examined local patterns via identifying
374 windows that exhibited patterns of F_{ST}, d_{XY}, and π matching one of the four evolutionary
375 scenarios: divergence-with-gene-flow, allopatric selection, recurrent selection, and balancing
376 selection (Table 1C, Han et al. 2017; Irwin et al. 2016, 2018; Shang et al. 2023). For F_{ST}, d_{XY},
377 and π, we considered windows with elevated values to be those with estimates above the 95th
378 percentile and windows with decreased values to be those with estimates below the 5th percentile.
379 We considered windows with average d_{XY} to be those with estimates that fell within the
380 interquartile range (for the allopatric selection scenario) following Shang et al. (2023) and
381 Piatkowski et al. (2023).
382
383 **RESULTS**
384 **Genomic landscape of differentiation between *Neodiprion lecontei* and *N. pinetum***

Whole-genome resequencing of 38 *Neodiprion lecontei* and *N. pinetum* individuals resulted in an average read count of 5,876,752 (range: 3,114,858 – 11,143,016). After excluding females with < 5 million reads (to balance data quality and sample size), we retained 14 *N. lecontei* and 11 *N. pinetum* individuals with an average site depth of 4.97x. Despite their recent divergence and continued gene exchange (Bendall et al. 2022), *N. lecontei* and *N. pinetum* exhibited substantial genomic divergence. Average genome-wide F_{ST} was 0.61 and d_{XY} was 3.16 x 10^{-3}. Average π was 1.78 x 10^{-3} for *N. lecontei* and 1.51 x 10^{-3} for *N. pinetum*. Average Tajima's D was -1.19 for *N. lecontei* and -1.25 for *N. pinetum*. Average Fay and Wu's H was -0.20 for *N. lecontei* and -0.36 for *N. pinetum*. Overall, all statistics were highly heterogeneous across the genome, with the most extreme values tending to occur in the putative centromeres (Figure 2). Notably, these putative centromeric regions also had the lowest number of called sites per window (Figure S10) and thus comprised the majority of the 570 windows that were filtered out and excluded from analysis for low site counts.

Relationships between genomic features and genetic summary statistics

The effects of linked selection on genetic variation are expected to be most pronounced in gene-dense and low-recombination regions (Ravinet et al. 2017; Stankowski et al. 2019), producing genome-wide correlations among recombination rate, gene density, within-species genetic diversity, and differentiation (Table 1A). While most of our genomic variables are themselves correlated (Figure 3), our multiple regression models were nevertheless able to tease apart some of their individual contributions to variation in each summary statistic (Table 2). Looking first at patterns of within-species variation, we found significant and negative genome-wide correlations between π and exon density in both *N. lecontei* and *N. pinetum* (Figure 3) and a

significant and negative relationship between exon density and π in multiple regression models for both species (Table 2). Although we did not detect a significant genome-wide correlation between recombination rate and π for *N. lecontei* (Figure 3), our multiple regression model revealed a significant and positive relationship between recombination rate and π in *N. lecontei* after taking into account other genomic variables (Table 2). For *N. pinetum*, however, the relationship between within-species diversity and recombination rate was significant and negative (Figure 3; Table 2). We consider possible explanations for this and other results that did not fit predictions of linked selection (Table 1A) in the discussion.

We also examined the relationship between genome features and two summary statistics derived from the site-frequency spectrum within each species. For Tajima's D, we found a significant and negative relationship between exon density and Tajima's D in both *N. lecontei* and *N. pinetum* (Table 2), supporting the prediction that signatures of linked selection (negative Tajima's D) are more pronounced in gene-dense regions. As expected, we also found a significant and positive relationship between recombination rate and Tajima's D in *N. lecontei* (Table 2). We did not recover a significant relationship between recombination rate and Tajima's D in *N. pinetum* (Table 2). For Fay and Wu's H, for which negative values are indicative of selective sweeps (Burri et al. 2015), we found a significant and *positive* relationship between exon density and Fay and Wu's H in both species (Table 2), implying that signatures of selective sweeps are most pronounced in gene-poor regions. Our multiple regression models also recovered a significant and *negative* relationship between recombination rate and Fay and Wu's H in *N. lecontei*, indicating more evidence of selective sweeps in high-recombination regions. We did not recover a significant relationship between recombination rate and Fay and Wu's H in *N. pinetum*.

Looking at patterns of differentiation between species, we found a significant and positive genome-wide correlation between F_{ST} and exon density (Figure 3) and a significant and positive relationship between exon density and F_{ST} in our multiple regression model (Table 2). Additionally, we found a significant and negative genome-wide correlation between F_{ST} and recombination rate (Figure 3) and a significant and negative relationship between recombination rate and F_{ST} in our multiple regression model (Table 2). These findings support the prediction that F_{ST} will be highest in gene-rich and low-recombination regions of the genome (Table 1A). For d_{XY}, we found that absolute divergence was highest in gene-poor and low-recombination regions. These findings provide mixed support for linked selection in ancestral populations (Table 1A, see below).

In addition to exploring the relationship between gene density and recombination rate and genetic summary statistics, we also examined the impact of site count, distance from the centromere, and mutation rate (approximated with dS). Site count predicted variation in all summary statistics except F_{ST} (Table 2). Whereas Fay and Wu's H tended to increase as site count increased, π, d_{XY}, and Tajima's D decreased as site count increased. We detected significant relationships between distance from the centromere and all summary statistics except F_{ST} and *N. lecontei* π; when significant, all relationships were negative except for *N. lecontei* Tajima's D: divergence and genetic diversity declined as distance from the centromere increased and signatures of linked selection tended to be found on chromosome arms (Table 2). We note, however, that when significant relationships were detected for site count and distance from the centromere, their effect sizes tended to be smaller than those estimated for exon density (Table 2). Finally, for mutation rate (dS), we only detected a significant relationship for *N. lecontei* π and *N. lecontei* Tajima's D, with both relationships being positive (Table 2). Estimated effect

454 sizes for mutation rate (dS) also tended to be smaller than those of all other predictor variables
455 (Table 2).

456 Taken together, our results suggest that background selection has played an important but
457 not exclusive role in shaping the heterogeneous landscape of differentiation between *N. lecontei*
458 and *N. pinetum*: there are numerous regions across the genome for both species that exhibit
459 signatures of selective sweeps (i.e., extremely negative Fay and Wu's H values; Figure 2), with
460 Fay and Wu's H tending to be lower when F_{ST} is high (Figure S11). Additionally, there are
461 multiple windows for both species where estimates of Tajima's D and Fay and Wu's H are
462 strongly negative, suggesting that the low values for Tajima's D in those windows is driven by
463 selective sweeps and not background selection (Figure S12).
464

465 **Genome-wide correlations between summary statistics**

466 Genome-wide, we found a significant and negative correlation between F_{ST} and d_{XY}, a
467 significant and negative correlation between F_{ST} and mean π, and a significant and positive
468 correlation between d_{XY} and mean π (Figure 3). These correlations were not sensitive to our
469 choice of window size (Figure S13). Collectively, these results are consistent with either the
470 recurrent selection or balancing selection evolutionary scenarios (Table 1B). Lending further
471 support to the recurrent selection scenario, we also found a significant and negative relationship
472 between exon density and d_{XY} in both our genome-wide correlations (Figure 3) and multiple
473 regression model (Table 2). Together, these results are consistent with widespread linked
474 selection in the ancestral population reducing coalescent times—and therefore d_{XY}—between
475 descendant lineages.
476

Local patterns of variation

As an additional method for distinguishing between recurrent selection and balancing selection—as well as to identify windows with variation patterns that deviate from the primary genome-wide pattern—we also examined local patterns of F_{ST}, d_{XY}, and π. We found that most windows with unusually high or low values for our focal summary statistics fit the recurrent selection evolutionary scenario (60.7% for *N. lecontei* π and 49.5% for *N. pinetum* π; Figures 4, S14-S20). Genomic regions matching the balancing selection evolutionary scenario were the second most frequent (31.5% for *N. lecontei* π and 48.5% for *N. pinetum* π), and these windows were located primarily in or near centromeres (Figures 4, S14-S20). The remainder of the detected windows (7.9% for *N. lecontei* π and 2.0% for *N. pinetum* π) fit the allopatric selection evolutionary scenario, in which high F_{ST} was the result of low π and not high d_{XY}. We did not detect any windows that fit the divergence-with-gene-flow evolutionary scenario for either species. In the discussion, we consider possible limitations of the predictions outlined in Table 1C.

DISCUSSION

Genome-wide patterns of variation in pine sawflies support pervasive linked selection

Over 50 years ago, Kimura (1968) proposed the neutral theory of molecular evolution, which posits that differences between species are due to neutral substitutions and that within-species polymorphism is governed by mutation-drift equilibrium dynamics. As genome assemblies and population genomic datasets became available, genome-wide patterns of polymorphism and divergence and their relationship with genome features have featured prominently in debates about whether modern data support neutral theory. Under neutrality,

variation in the neutral mutation rate across the genome is expected to produce a strong positive correlation between interspecific divergence and intraspecific polymorphism. By contrast, except for the potential mutagenic effect of recombination (Pratto et al. 2014; Arbeithuber et al. 2015), recombination environment is expected to have no effect on levels of polymorphism under neutrality (Hudson 1983). Under selection, however, polymorphism levels should relate to recombination rate (with lower recombination rates having stronger effects on linked neutral sites; Stankowski et al. 2019). For this reason, the positive correlations between polymorphism and recombination that have been observed in several taxa (Cutter and Payseur 2013) have been interpreted as evidence that neutral theory does not adequately explain modern genomic data (Hahn 2008; Kern and Hahn 2018; but see Jensen et al. 2018). Meanwhile, others have argued that the number of species with relevant data—i.e., high quality reference genomes and independent estimates of recombination rate (e.g., from crosses)—remain too limited to support this claim (Jensen et al. 2018). Additionally, while it is uncontroversial that linked selection affects patterns of polymorphism (Maynard Smith and Haigh 1974; Charlesworth et al. 1993), the proportion of the genome affected and the relative importance of background selection versus selective sweeps remains debated (Jensen et al. 2018; Kern and Hahn 2018; Pouyet et al. 2018). Ultimately, more data are needed to characterize genomic landscapes across diverse taxa. Interpreting these landscapes in light of neutral theory requires well-annotated reference genomes, recombination rate data, and adequate controls for genotyping error.

Consistent with neutral expectations, we found a strong positive correlation between interspecific divergence (d_{XY}) and intraspecific polymorphism (π) in both *N. lecontei* and *N. pinetum* (Figure 3). However, the synonymous substitution rate (dS; our proxy for the neutral mutation rate) tended to have the smallest estimated effect size among our genomic predictor

variables and was not significant after accounting for other variables in most of our multiple regression models (Table 2). By contrast, and as observed in other taxa including in other invertebrates (Wallberg et al. 2015; Christmas et al. 2021; Herrig et al. 2024), vertebrates (Burri et al. 2015; Rettelbach et al. 2019; Kartje et al. 2020; Chase et al. 2021; Rougemont et al. 2019; Rodrigues et al. 2023; Moreira et al. 2023), and plants (Flowers et al. 2012; Stankowski et al. 2019; Shang et al. 2023), we found significant relationships between intraspecific polymorphism and recombination rate, although not always in the direction predicted under simple models of linked selection (Figure 3; Tables 1A, 2; see below). Another line of evidence consistent with expectations under linked selection is that exon density—which should correlate positively with the density of selected sites (Payseur and Nachman 2002)—was negatively correlated with intraspecific polymorphism in both species (Figure 3; Table 2). In fact, exon density often had the largest effect size of all variables in our multiple regression models (Table 2). Overall, these data suggest that neutral mutation rate alone cannot explain our observed positive correlations between divergence and polymorphism. Instead, we argue that these patterns result from selection repeatedly targeting the same regions in ancestral and descendant populations (i.e., recurrent selection, see below). In further support of pervasive linked selection in ancestral populations, a recent analysis of genealogical variation across the genomes of 19 eastern North American *Neodiprion* species (including the focal species here) revealed that concordance with the estimated species tree tended to be highest in low-recombination and gene-dense regions (Herrig et al. 2024). These patterns are expected under linked selection because when ancestral polymorphism is reduced, phylogenetic discordance via incomplete lineage sorting will also be reduced (Pease and Hahn 2013).

Although most of our results fit the patterns expected from widespread linked selection, some patterns deviate from expectations. Specifically, we found that both π in *N. pinetum* and d_{XY} tended to be lower in high-recombination regions of the genome. Indeed, studies in other animal and plant taxa have revealed a mixture of correlations between π and recombination rate ranging from strong positive correlations (e.g., Wallberg et al. 2015; Burri et al. 2015; Rougemont et al. 2019; Stankowski et al. 2019), to minimal or no correlations (e.g., Payseur and Nachman 2002; Flowers et al. 2012; Kartje et al. 2020), to negative correlations (e.g., Flowers et al. 2012). Here, we consider three non-mutually exclusive explanations for unexpected negative correlations between recombination rate and polymorphism (Table 1A). Before doing so, we first note that absolute divergence (d_{XY}) between two species is the combination of the amount of variation that existed in the ancestral population at the time of the speciation event (i.e., ancestral π) and the accumulation of substitutions post-speciation (Cruickshank and Hahn 2014). For recently diverged species, d_{XY} largely reflects ancestral diversity.

One potential explanation for the negative correlation between recombination rate and diversity in *N. pinetum* (π) and in the *N. lecontei/N. pinetum* ancestor (d_{XY}) is that we quantified exon density using a *N. lecontei* genome annotation and used a genetic map from a *N. lecontei* cross to quantify recombination rate across the genome, potentially leading to incorrect inferences about local gene density and recombination rate in *N. pinetum* and in the shared ancestor. However, chromosome-level assemblies for *N. lecontei* and *N. pinetum* indicate that the two genomes are colinear (Herrig et al. 2024), and recombination rate estimates from interspecific genetic maps recapitulate patterns in our *N. lecontei* genetic map (unpublished data). For these reasons, we expect very similar patterns of local gene density and recombination rates in the two species, making it unlikely that the unexpected correlations with recombination

rate can be explained by our use of *N. lecontei* genome feature data. Nevertheless, local recombination rate sometimes varies between closely related species (McGaugh et al. 2012, but see Burri et al. 2015; Rodrigues et al. 2023; Wang et al. 2022; Moreira et al. 2023; Shang et al. 2023), so an intraspecific recombination rate map in *N. pinetum* would be necessary to rule out this potential explanation.

A second potential explanation for the negative correlation between diversity metrics involving *N. pinetum* (π and d_{XY}) and recombination rate is that we mapped all sequencing reads to the *N. lecontei* reference genome. Although synteny plots constructed from annotated gene sets indicate that *N. lecontei* and *N. pinetum* genomes are colinear (Herrig et al. 2024), alignment errors for the non-reference species could be elevated in the repeat-rich centromeric regions, which also have the lowest recombination rates in the genome. Indeed, the highest d_{XY} and *N. pinetum* π values tended to be observed in and around centromeres (Figure 2). To control for variation in genotyping error across the genome, we excluded genomic windows with unusually low site count and included both distance from centromere and site count (a proxy for local genotyping error) in our regression models (Table 2). Despite these efforts, it is possible that our data remain insufficient to fully tease apart the effects of minimal recombination and increased genotyping error in centromeric regions. Long-read population genomic data and re-analysis using a *N. pinetum* reference genome would be informative for evaluating the accuracy of diversity estimates obtained using the *N. lecontei* reference genome.

A third potential explanation for the negative correlation between genetic diversity and recombination rate in *N. pinetum* and the shared ancestor, but not *N. lecontei*, is that the former lineages experienced more intense Hill-Robertson interference, which refers to the reduction in the efficacy of selection stemming from selection on two or more linked sites (Hill and

591 Robertson 1966). Because this reduced efficacy of selection is expected to be most pronounced
592 in low-recombination regions, increased recombination rates should reduce Hill-Robertson
593 interference and increase fixation rates for beneficial alleles (Felsenstein 1974; Comeron et al.
594 2008; Flowers et al. 2012). This mechanism has been proposed to explain the stronger negative
595 correlation between π and recombination rate (driven by lower π on chromosome arms) in
596 domesticated strains of rice compared to a wild strain (Flowers et al. 2012). Similarly, we
597 propose that pervasive Hill-Robertson interference associated with novel host-associated
598 selection pressures could explain the observed negative correlation between diversity and
599 recombination rate in *N. pinetum* and in the *N. pinetum/N. lecontei* ancestor.
600 The eastern North American *Neodiprion* clade radiated rapidly and recently onto a
601 variety of pine species (Linnen and Farrell 2008a,b; Herrig et al. 2024), making it likely that
602 ancestral populations experienced abundant novel host-associated selection pressures. Similarly,
603 *N. pinetum* recently shifted onto eastern white pine, a novel host with much thinner and less
604 resinous needles than other eastern North America pines (Linnen and Farrell 2010; Herrig et al.
605 2024). Adaptation to this novel host—which all other eastern North American *Neodiprion*
606 avoid—required changes to behavioral, physiological, and morphological traits expressed in
607 eggs, larvae, and adults (Bendall et al. 2017, 2023; Glover et al. 2023). By contrast, based on
608 host associations in other *Neodiprion* species, host-use patterns in *N. lecontei* likely resemble
609 those of the ancestral population (Linnen and Farrell 2010; Herrig et al. 2024). When adaptation
610 is highly polygenic, as is likely the case for host adaptation in these specialist pine feeders,
611 selection can affect a substantial proportion of the genome (Stankowski et al. 2019). Thus,
612 compared to *N. lecontei*, we might expect more of *N. pinetum*'s genome (and potentially the
613 ancestral genome) to have experienced positive selection. Consistent with this prediction, there

are numerous regions across the genome that exhibit signatures of selective sweeps (i.e., strongly negative Fay and Wu's H values) in both species, with these regions appearing to be more numerous in *N. pinetum* (Figures 2, S12). Moreover, average genome-wide Fay and Wu's H is lower in *N. pinetum* (-0.36) compared to *N. lecontei* (-0.20). Also consistent with the hypothesis that Hill-Robertson interference affects genome-wide patterns of variation in *Neodiprion* is the observation that in both *N. lecontei* and *N. pinetum*, signatures of selective sweeps (i.e., more negative Fay and Wu's H estimates) tend to be found in gene-poor regions (Table 2). Overall, these findings add to the growing body of empirical (e.g., Irwin et al. 2016; Rettelbach et al. 2019; Stankowski et al. 2019; Chase et al. 2021; Wang et al. 2022; Shang et al. 2023; Jiang et al. 2023) and theoretical (Matthey-Doret and Whitlock 2019; Schrider 2020) literature demonstrating that background selection alone may not be sufficient to explain the observed patterns of diversity and differentiation. Regardless of the explanation, our study shows that the effects of linked selection can vary in important ways even among closely related taxa, such as *N. lecontei* and *N. pinetum*.

Pine sawfly genomic landscapes support recurrent selection, not divergence-with-gene-flow

A unique feature of the pine sawfly system is that their ecology, demography, and haplodiploid transmission genetics (Figure 1) allow us to make competing *a priori* predictions about the predominant evolutionary scenarios shaping their genomic landscape of differentiation. Specifically, we expected to find evidence of recurrent selection, divergence-with-gene-flow, or some mixture of the two. Examination of both global and local patterns of F_{ST}, d_{XY}, and π provided strong support for recurrent selection—i.e., when selection repeatedly targets the same regions of the genome in both ancestral populations and descendant lineages—as the primary

process shaping variation across pine sawfly genomes. These findings are consistent with a growing number of studies in diverse animal and plant taxa that have also identified recurrent selection as the primary driver of heterogeneous genomic differentiation (e.g., Irwin et al. 2016, 2018; Stankowski et al. 2019; Chase et al. 2021; Jiang et al. 2023).

While most of our local "outlier" windows fit the recurrent selection evolutionary scenario (~50-60%), we did detect at least some windows that had patterns consistent with either the balancing selection or the allopatric selection evolutionary scenarios (Figures 4, S14-S20). This suggests that although many of the same regions of the genome have been targeted by selection in the common ancestor and in both *N. lecontei* and *N. pinetum*, there are some regions of the genome where ancestral polymorphism has been maintained and where lineage-specific selection has occurred due to local ecological adaptation (Han et al. 2017; Shang et al. 2023). Intriguingly, patterns of balancing selection were most evident near the centromeres (Figures 4, S14-S20). As noted above, one potential explanation for elevated π and d_{XY} surrounding centromeres is that these are artifacts of increased genotyping error in and around repeat-rich centromeres.

Signatures of balancing selection near centromeres could also be due to unique evolutionary dynamics near centromeres. For example, in *Mimulus* monkeyflowers, strong female meiotic drive has been linked to a centromere-associated repeat domain locus. Individuals that are homozygous for the driving allele suffer reduced pollen viability; thus, balancing selection prevents the fixation or loss of the driving allele within the population (Fishman and Saunders 2008). More generally, there are several mechanisms by which balancing selection maintains polymorphism, including heterozygote advantage, frequency-dependent selection, sexual antagonism, and variation in fitness across time (including between larval and adult

stages) and space (Llaurens et al. 2017). However, trans-species polymorphisms can also persist due to neutral processes (Wiuf et al. 2004). Thus, additional tests (e.g., simulations) are required to rule out neutral processes and demonstrate that trans-species polymorphisms have been maintained by long-term balancing selection. Overall, more work is required to determine the underlying mechanisms maintaining polymorphism in *N. lecontei* and *N. pinetum*, particularly in or near centromeres.

Surprisingly, we did not recover any windows that fit the divergence-with-gene-flow evolutionary scenario. We do not think this finding is due to incorrect inferences about the demography and ecology of speciation in *N. lecontei* and *N. pinetum* because all existing data strongly and consistently support ecological speciation driven by a recent host shift with substantial gene exchange throughout divergence (Linnen and Farrell 2007, 2010; Bendall et al. 2017, 2022, 2023; Glover et al. 2023). One potential explanation for this finding is that our observed lack of divergence-with-gene-flow windows is an artifact of how we selected and categorized outlier windows. Recurrent selection, allopatric selection, and divergence-with-gene-flow all predict some outlier windows with high F_{ST} and low π, corresponding to regions under selection. These scenarios differ only in their predictions about whether absolute divergence (d_{XY}) will be high (divergence-with-gene-flow), average (allopatric selection), or low (recurrent selection) relative to the rest of the genome. Categorization is therefore influenced by arbitrary cutoffs for what we consider low (bottom 5%), average (interquartile range), or high (top 5%).

Additionally, the predictions for divergence-with-gene-flow windows assume that outside of barrier loci, the rest of the genome is evolving neutrally. Based on the evidence discussed above supporting pervasive linked selection, this assumption is almost certainly violated. Importantly, when a locus experiences selection in the ancestral population, the statistical power

683 to detect increased d_{XY} between the descendant lineages at this locus due to restricted gene flow
684 is drastically reduced unless very high levels of gene flow occur across the rest of the genome
685 (Cruickshank and Hahn 2014). Thus, while examining local genomic patterns of variation
686 provides some additional context when interpreting genome-wide correlations—especially
687 distinguishing between recurrent and balancing selection—such qualitative categorizations
688 should be interpreted with caution.

689 Regardless of how we identify and categorize putative outlier windows, it is nevertheless
690 true that the predominant patterns in our genome-wide data do not fit published expectations for
691 divergence-with-gene-flow (Table 1B). One potential explanation for this finding is that genome-
692 wide signatures of divergence-with-gene-flow are likely to be ephemeral and *N. lecontei* and *N.*
693 *pinetum* are already too diverged to recover this pattern (Figure 2). In early stages of primary
694 speciation-with-gene-flow (i.e., no periods of allopatry throughout divergence), theory predicts
695 that gene flow will be reduced only at loci involved in reproductive isolation (i.e., "speciation
696 genes") and tightly linked loci, forming localized "islands of differentiation", while the rest of
697 the genome is homogenized by gene flow (Wu 2001; Turner et al. 2005; Via and West 2008; Via
698 2012; Nosil and Feder 2012). As additional loci diverge, effective gene flow is reduced across
699 more of the genome, eventually leading to widespread genomic divergence (Nosil and Feder
700 2012). When this occurs, islands of differentiation become harder to detect (Via 2012; Han et al.
701 2017; Gauthier et al. 2018; Jiang et al. 2023) and genome-wide correlations between F_{ST}, d_{XY},
702 and π may become less pronounced. Moreover, as reproductive isolation increases and gene flow
703 declines further, diverging populations will increasingly behave as semi-independent
704 populations, diverging via drift and independent bouts of selection. Eventually, an initial pattern
705 of divergence-with-gene-flow may get "overwritten" by subsequent bouts of selection,

increasingly producing genome-wide correlations consistent with whatever selection scenario predominated post-divergence.

Evaluating our hypothesis that divergence-with-gene-flow signatures are ephemeral will require characterizing the genomic landscape of divergence across multiple timepoints in the *Neodiprion* speciation continuum. Indeed, this strategy is increasingly applied in other taxa and is becoming a promising tool to investigate how genomic landscapes "evolve" as speciation proceeds (e.g., Burri et al. 2015; Stankowski et al. 2019; Shang et al. 2023). As a complementary approach, simulations under a wide range of selection scenarios, demographic histories, and divergence time scales would be very useful for evaluating the robustness and temporal stability of the qualitative predictions outlined in Tables 1B, 1C (e.g., Matthey-Doret and Whitlock 2019; Rettelbach et al. 2019; Stankowski et al. 2019). From these studies, we can better understand the evolutionary forces shaping the genomic landscape, further enhancing our understanding of the genetics of adaptation and speciation.

Does haplodiploidy predict recurrent selection?

Sex chromosomes are hypothesized to be particularly prone to recurrent selection due to expression of all recessive mutations in the heterogametic sex (Charlesworth et al. 1987; Ellegren et al. 2012; Oyler-McCance et al. 2015; Irwin et al. 2016; Miller and Sheehan 2023). Because haplodiploid inheritance patterns are similar to that of sex chromosomes (Nouhaud et al. 2020), we hypothesize that haplodiploids may be especially prone to recurrent selection, even if there is gene flow throughout divergence. In support of this hypothesis, we found that recurrent selection is likely the primary process shaping the heterogeneous landscape of differentiation between *N. lecontei* and *N. pinetum*. Unfortunately, a taxonomic bias in the literature precludes

us from assessing the prevalence of this pattern in haplodiploids. Although other taxa such as birds (e.g., Burri et al. 2015; Irwin et al. 2016, 2018; Han et al. 2017; Rettelbach et al. 2019; Chase et al. 2021; Jiang et al. 2023; Moreira et al. 2023), fish (e.g., Rougemont et al. 2019; Wang et al. 2022; Sun et al. 2022), plants (e.g., Flowers et al. 2012; Ma et al. 2018; Stankowski et al. 2019; Piatkowski et al. 2023; Shang et al. 2023), and other insects (e.g., Lindtke et al. 2017; Wong Miller et al. 2017; Talla et al. 2019; Fiteni et al. 2022) are well represented in genomic landscape studies, we only found one other comparable study (i.e., that estimates and compares patterns of F_{ST}, d_{XY}, and π across the genome so that these patterns can be matched to one or more of the four selection-based evolutionary scenarios) in a haplodiploid system (Christmas et al. 2021).

As in our study, Christmas et al. (2021) found that genomic windows with elevated differentiation in *Bombus* alpine bumblebees tended to be found in genomic regions with high gene density, low recombination rates, and low π. Additionally, they found evidence of recurrent selection (low d_{XY} in F_{ST} outlier windows) in an intraspecific comparison and in an allopatric species pair. Conversely, and unlike our findings, they found evidence of divergence-with-gene-flow (high d_{XY} in F_{ST} outlier windows) in a sympatric species pair. Notably, average genome-wide F_{ST} in the *Bombus* sympatric species pair (0.41) is considerably lower than in our *Neodiprion* pine sawflies (0.61). Also, it is unclear whether these sympatric *Bombus* species diverged with continuous gene flow or whether they experienced periods of allopatry and subsequent gene flow upon secondary contact. The power to detect locally elevated d_{XY} due to restricted gene flow is much higher when gene flow occurs upon secondary contact compared to divergence with continuous gene flow at short divergence times (Cruickshank and Hahn 2014). Collectively, the overall lower genomic differentiation and possibly different demographic

752 history between the sympatric *Bombus* species compared to *N. lecontei* and *N. pinetum* could
753 explain the different findings between our study and Christmas et al. (2021). Ultimately,
754 however, more studies in diverse haplodiploid taxa are needed to determine whether this mode of
755 reproduction has a predictable impact on genome-wide patterns of genetic differentiation.
756

757 **CONCLUSIONS**

758 Collectively, our study makes several important contributions to the study of genomic
759 landscapes. First, our study adds to the growing body of literature documenting evidence of
760 pervasive linked selection across the genome. Second, our study highlights that even when there
761 is widespread linked selection, genomic predictors of variation can differ even between closely
762 related species. Third, although gene density and recombination rate appear to be the primary
763 biological sources of variation in genetic summary statistics across the genome, our study
764 demonstrates that it is important to consider other factors, such as genotyping error and
765 proximity to centromeres. Fourth, by focusing on a haplodiploid species pair, our study fills an
766 important taxonomic gap in the speciation genomics literature and supports the hypothesis that
767 patterns of variation in haplodiploids will be heavily influenced by recurrent selection.
768 Nevertheless, it is also clear that more genomic landscape studies from a broader range of taxa
769 are required to determine whether there are consistent differences in landscape features among
770 different taxonomic groups and divergence scenarios. Given the paucity of data from
771 haplodiploids, we suggest that investigation of such taxa should be a high priority for future
772 studies. As demonstrated here, genomic landscape studies are perhaps most informative when
773 there are sufficient genomic resources (e.g., high-quality reference genomes, genome

774 annotations, recombination maps) and information about the study system (e.g., ecology and
775 divergence history) to aid hypothesis generation and data interpretation.
776
777 **ACKNOWLEDGEMENTS**
778 We thank members of the Linnen lab for assistance with pine sawfly collection and
779 rearing, Andres Bendesky for providing the reagents used in the Tn5 tagmentation library
780 preparation protocol, Emily Bendall for advice on library preparation, and Danielle Herrig for
781 providing the genome of an outgroup taxon. We also thank three anonymous reviewers whose
782 comments helped us improve our analyses and interpretations and the clarity of our arguments.
783 This work was supported by the USDA National Institute of Food and Agriculture Predoctoral
784 Fellowship (2022-67011-36550) to ANG and the National Science Foundation DEB-1257739
785 and DEB-CAREER-1750946 to CRL. For computing resources, we thank the University of
786 Kentucky Center for Computational Sciences, the Morgan Compute Cluster, and the SCINet
787 project and the AI Center of Excellence of the USDA Agricultural Research Service, ARS
788 project numbers 0201-88888-003-000D and 0201-88888-002-000D. The US Department of
789 Agriculture, Agricultural Research Service is an equal opportunity/affirmative action employer
790 and all agency services are available without discrimination.
791

792 **REFERENCES**

793 Arbeithuber, B., Betancourt, A. J., Ebner, T., & Tiemann-Boege, I. (2015). Crossovers are
794 associated with mutation and biased gene conversion at recombination hotspots.
795 *Proceedings of the National Academy of Sciences of the United States of America*, 112,
796 2109-2114. https://doi.org/10.1073/pnas.1416622112
797 Avery, P. J. (1984). The population genetics of haplo-diploids and X-linked genes. *Genetical*
798 *Research*, 44(3), 321-341. https://doi.org/10.1017/S0016 67230 0026550
799 Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y. P., Hahn, M. W., ... &
800 Langley,
801 C. H. (2007). Population genomics: whole-genome analysis of polymorphism and
802 divergence in *Drosophila simulans*. *PLoS Biology*, 5(11), e310.
803 https://doi.org/10.1371/journal.pbio.0050310
804 Bendall, E. E. (2020). From Genes to Species: Ecological Speciation with Gene Flow in
805 *Neodiprion pinetum* and *N. lecontei*. *Theses and Dissertations—Biology*, 62,
806 https://uknowledge.uky.edu/biology_etds/62
807 Bendall, E. E., Vertacnik, K. L., & Linnen, C. R. (2017). Oviposition traits generate extrinsic
808 postzygotic isolation between two pine sawfly species. *BMC Evolutionary Biology*,
809 17(1), 1-15. https://doi.org/10.1186/s1286 2-017-0872-8
810 Bendall, E., E., Bagley, R. K., Sousa, V. C., & Linnen, C. R. (2022). Faster-haplodiploid
811 evolution under divergence-with-gene-flow: Simulations and empirical data from pine-
812 feeding hymenopterans. *Molecular Ecology*, 00, 1-19. https://doi.org/10.1111/mec.16410
813 Bendall, E. E., Mattingly, K. M., Moehring, A. J., & Linnen, C. R. (2023). A test of Haldane's
814 rule in *Neodiprion* sawflies and implications for the evolution of postzygotic isolation in

haplodiploids. *The American Naturalist*, 202(1), 40-54. https://doi.org/10.1086/724820

Bensasson, D. (2011). Evidence for a high mutation rate at rapidly evolving yeast centromeres. *BMC Evolutionary Biology*, 11, 1-11. https://doi.org/10.1186/1471-2148-11-211

Blackmon, H., Ross, L., & Bachtrog, D. (2017). Sex determination, sex chromosomes, and karyotype evolution in insects. *Journal of Heredity*, 108(1), 78-93. https://doi.org/10.1093/jhered/esw047

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Branca, A., Paape, T. D., Zhou, P., Briskine, R., Farmer, A. D., Mudge, J., ... & Tiffin, P. (2011). Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proceedings of the National Academy of Sciences*, 108(42), E864-E870. https://doi.org/10.1073/pnas.1104032108

Burri, R. (2017a). Dissecting differentiation landscapes: A linked selection's perspective. *Journal of Evolutionary Biology*, 30(8), 1501-1505. https://doi.org/10.1111/jeb.13108

Burri, R. (2017b). Interpreting differentiation landscapes in the light of long-term linked selection. *Evolution Letters*, 1, 118-131. https://doi.org/10.1002/evl3.14

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., Suh, A., Dutoit, L., Bureš, S., Garamszegi, L. Z., Hogner, S., Moreno, J., Qvarnström, A., Ružić, M., Sæther, S.-A., Sætre, G.-P., Török, J., & Ellegren, H. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome*

Research, 25(11), 1656-1665. https://doi.org/10.1101/gr.196485.115

Castellano, D., Eyre-Walker, A., & Munch, K. (2019). Impact of mutation rate and selection at linked sites on DNA variation across the genomes of humans and other Homininae. *Genome Biology and Evolution*, 12(1), 3550-3561. https://doi.org/10.1093/gbe/evz215

Charlesworth, B. (1998). Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, 15(5), 538-543. https://doi.org/10.1093/oxfordjournals.molbev.a025953

Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, 2(4), e64. https://doi.org/10.1371/journ al.pgen.0020064

Charlesworth, B., Coyne, J. A., & Barton, N. H. (1987). The relative rates of evolution of sex chromosomes and autosomes. *American Naturalist*, 130(1), 113-146. https://doi.org/10.1086/284701

Charlesworth, B., Morgan, M. T., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4), 1289-1303. https://doi.org/10.1093/genetics/134.4.1289

Charlesworth, D., Charlesworth, B., & Morgan, M. T. (1995). The pattern of neutral molecular variation under the background selection model. *Genetics*, 141(4), 1619-1632. https://doi.org/10.1093/genetics/141.4.1619

Chase, M. A., Ellegren, H., & Mugal, C. F. (2021). Positive selection plays a major role in shaping signatures of differentiation across the genomic landscape of two independent *Ficedula* flycatcher species pairs. *Evolution*, 75(9), 2179-2196. https://doi.org/10.1111/evo.14234

Christmas, M. J., Jones, J. C., Olsson, A., Wallerman, O., Bunikis, I., Kierczak, M., ... & Webster, M. T. (2021). Genetic barriers to historical gene flow between cryptic species of alpine bumblebees revealed by comparative population genomics. *Molecular Biology and Evolution*, 38(8), 3126-3143. https://doi.org/10.1093/molbev/msab086

Clark, R. M., Schweikert, T., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., ... & Weigel, D. (2007). Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science*, 317(5836), 338-342. https://doi.org/10.1126/science.1138632

Comeron, J. M. (2017). Background selection as null hypothesis in population genomics: insights and challenges from *Drosophila* studies. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 372, 20160471. http://dx.doi.org/10.1098/rstb.2016.0471

Comeron, J. M., Williford, A., & Kilman, R. M. (2008). The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. *Heredity*, 100(1), 19-31. https://doi.org/10.1038/sj.hdy.6801059

Coppel, H., & Benjamin, D. (1965). Bionomics of the Nearctic pine-feeding diprionids. *Annual Review of Entomology*, 10, 69-96. https://doi.org/10.1146/annur ev.en.10.010165.000441

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23(13), 3133-3157. https://doi.org/10.1111/mec.12796

Cutter, A. D., & Payseur, B. A. (2013). Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics*, 14(4), 262-274. https://doi.org/10.1038/nrg3425

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A.,

884 Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools
885 and BCFtools. *GigaScience*, 10(2), giab008. https://doi.org/10.1093/gigascience/giab008
886 de la Filia, A. G., Bain, S. A., & Ross, L. (2015). Haplodiploidy and the reproductive ecology
887 of arthropods. *Current Opinion in Insect Science*, 9, 36-43.
888 https://doi.org/10.1016/j.cois.2015.04.018
889 Derelle, R., Philippe, H., & Colbourne, J. K. (2020). Broccoli: combining phylogenetic and
890 network analyses for orthology assignment. *Molecular Biology and Evolution*, 37(11),
891 3389-3396. https://doi.org/10.1093/molbev/msaa159
892 Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., &
893 Aiden, E. L. (2016). Juicebox provides a visualization system for Hi-C contact maps with
894 unlimited zoom. *Cell Systems*, 3(1), 99-101. http://dx.doi.org/10.1016/j.cels.2015.07.012
895 Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., Künstner, A.,
896 Mäkinen, H., Nadachowska-Brzyska, K., Qvarnström, A., Uebbing, S., & Wolf, J. B.
897 (2012). The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*,
898 491(7426), 756-760. https://doi.org/10.1038/natur e11584
899 Everitt, T., Wallberg, A., Christmas, M. J., Olsson, A., Hoffmann, W., Neumann, P., & Webster,
900 M. T. (2023). The genomic basis of adaptation to high elevations in Africanized honey
901 bees. *Genome Biology and Evolution*, 15(9), evad157.
902 https://doi.org/10.1093/gbe/evad157
903 Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations.
904 *Nature Reviews Genetics*, 8(8), 610-618. https://doi.org/10.1038/nrg2146
905 Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, 155,
906 1405-1413. https://doi.org/10.1093/genetics/155.3.1405

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics*, 78(2), 737-756. https://doi.org/10.1093/genetics/78.2.737

Fishman, L., & Saunders, A. (2008). Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. *Science*, 322(5907), 1559-1562. https://doi.org/10.1126/science.1161406

Fiteni, E., Durand, K., Gimenez, S., Meagher Jr, R. L., Legeai, F., Kergoat, G. J., ... & Nam, K. (2022). Host-plant adaptation as a driver of incipient speciation in the fall armyworm (*Spodoptera frugiperda*). *BMC Ecology and Evolution*, 22(1), 133. https://doi.org/10.1186/s12862-022-02090-x

Flowers, J. M., Molina, J., Rubinstein, S., Huang, P., Schaal, B. A., & Purugganan, M. D. (2012). Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. *Molecular Biology and Evolution*, 29(2), 675-687. https://doi.org/10.1093/molbev/msr225

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage.

Fumagalli, M., Vieira, F. G., Korneliussen, T. S., Linderoth, T., Huerta-Sanchez, E., Albrechtsen, A., & Nielsen, R. (2013). Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, 195, 979-992. https://doi.org/10.1534/genetics.113.154740

Gauthier, J., Gayral, P., Le Ru, B. P., Jancek, S., Dupas, S., Kaiser, L., ... & Herniou, E. A. (2018). Genetic footprints of adaptive divergence in the bracovirus of *Cotesia sesamiae* identified by targeted resequencing. *Molecular Ecology*, 27(8), 2109-2123.

https://doi.org/10.1111/mec.14574

Glover, A. N., Bendall, E. E., Terbot II, J. W., Payne, N., Webb, A., Filbeck, A., Norman, G., & Linnen, C. R. (2023). Body size as a magic trait in two plant-feeding insect species. *Evolution*, 77(2), 437-453. https://doi.org/10.1093/evolut/qpac053

Glover, A. N., Sousa, V. C., Ridenbaugh, R. D., Sim, S. B., Geib, S. M., & Linnen, C. R. (2024). Data from: Recurrent selection shapes the genomic landscape of differentiation between a pair of host-specialized haplodiploids that diverged with gene flow. *Dryad*. doi: https://doi.org/10.5061/dryad.fxpnvx128

Gore, M. A., Chia, J. M., Elshire, R. J., Sun, Q., Ersoz, E. S., Hurwitz, B. L., ... & Bucker, E. S. (2009). A first-generation haplotype map of maize. *Science*, 326(5956), 1115-1117. https://doi.org/10.1126/science.1177837

Guerrero, R. F,, & Hahn, M. W. (2017). Speciation as a sieve for ancestral polymorphism. *Molecular Ecology*, 26, 5362-5368. https://doi.org/10.1111/mec.14290

Hahn, M. W. (2008). Toward a selection theory of molecular evolution. *Evolution*, 62(2), 255-265. https://doi.org/10.1111/j.1558-5646.2007.00308.x

Han, F., Lamichhaney, S., Grant, B. R., Grant, P. R., Andersson, L., & Webster, M. T. (2017). Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. *Genome Research*, 27, 1004-1015. https://doi:10.1101/gr.212522.116

Harper, K. E., Bagley, R. K., Thompson, K. L., & Linnen, C. R. (2016). Complementary sex determination, inbreeding depression and inbreeding avoidance in a gregarious sawfly. *Heredity*, 117(5), 326-335. https://doi.org/10.1038/hdy.2016.46

Henikoff, S., Ahmad, K., & Malik, H. (2001). The centromere paradox: stable inheritance with

rapidly evolving DNA. *Science*, 293(5532), 1098-1102. https://doi.org/10.1126/science.1062939

Herrig, D. K., Ridenbaugh, R. D., Vertacnik, K. L., Everson, K. M., Sim, S. B., Geib, S. M., Weisrock, D. W., & Linnen, C. R. (2024). Whole genomes reveal evolutionary relationships and mechanisms underlying gene-tree discordance in *Neodiprion* sawflies. *Systematic Biology*, syae036. https://doi.org/10.1093/sysbio/syae036

Hill, W. G., & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetics Research*, 8(3), 269-294. https://doi.org/10.1017/S0016672300010156

Hofstatter, P. G., Thangavel, G., Castellani, M., & Marques, A. (2021). Meiosis progression and recombination in holocentric plants: what is known? *Frontiers in Plant Science*, 12, 658296. https://doi.org/10.3389/fpls.2021.658296

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. http://www.jstor.org/stable/4615733

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic rcombination. *Theoretical Population Biology*, 23(2), 183-201. https://doi.org/10.1016/0040-5809(83)90013-8

Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, 132(2), 583-589. https://doi.org/10.1093/genetics/132.2.583

Irwin, D. E., Alcaide, M., Delmore, K. E., Irwin, J. H., & Owens, G. L. (2016). Recurrent selection explains parallel evolution of genomic regions of high relative but low absolute differentiation in a ring species. *Molecular Ecology*, 25(18), 4488-4507. https://doi.org/10.1111/mec.13792

976 Irwin, D. E., Milá, B., Toews, D. P. L., Brelsford, A., Kenyon, H. L., Porter, A. N., Grossen, C.,
977 Delmore, K. E., Alcaide, M., & Irwin, J. H. (2018). A comparison of genomic islands of
978 differentiation across three young avian species pairs. *Molecular Ecology*, 27(23),
979 4839-4855. https://doi.org/10.1111/mec.14858
980 Jensen, J. D., Payseur, B. A., Stephan, W., Aquadro, C. F., Lynch, M., Charlesworth, D., &
981 Charlesworth, B. (2018). The importance of the Neutral Theory in 1968 and 50 years on:
982 A response to Kern and Hahn 2018. *Evolution*, 73(1), 111-114.
983 https://doi.org/10.1111/evo.13650
984 Jiang, Z., Song, G., Luo, X., Zhang, D., Lei, F., & Qu, Y. (2023). Recurrent selection and
985 reduction in recombination shape the genomic landscape of divergence across multiple
986 population pairs of Green-backed Tit. *Evolution Letters*, 7(2), 99-111.
987 https://doi.org/10.1093/evlett/qrad005
988 Kartje, M. E., Jing, P., & Payseur, B. A. (2020). Weak correlation between nucleotide variation
989 and recombination rate across the house mouse genome. *Genome Biology and Evolution*,
990 12(4), 293-299. https://doi.org/10.1093/gbe/evaa045
991 Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7:
992 improvements in performance and usability. *Molecular Biology and Evolution*, 30(4),
993 772-780. https://doi.org/10.1093/molbev/mst010
994 Kern, A. D., & Hahn, M. W. (2018). The neutral theory in light of natural selection. *Molecular*
995 *Biology and Evolution*, 35(6), 1366-1371. https://doi.org/10.1093/molbev/msy092
996 Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217, 624-626.
997 https://doi.org/10.1038/217624a0
998 Knerer, G., & Atwood, C. E. (1973). Diprionid sawflies: Polymorphism and speciation. *Science*,

179(4078), 1090-1099. https://doi.org/10.1126/scien ce.179.4078.1090

Korneliussen, T. S., Moltke, I., Albrechtsen, A., & Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, 14, 289. https://doi.org/10.1186/1471-2105-14-289

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*, 15, 1-13. https://doi.org/10.1186/s12859-014-0356-4

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Lindtke, D., Lucek, K., Soria-Carrasco, V., Villoutreix, R., Farkas, T. E., Riesch, R., Dennis, S. R., & Gompert, Z. (2017). Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect. *Molecular Ecology*, 26, 6189-6205. https://doi.org/10.1111/mec.14280

Linnen, C. R., & Farrell, B. D. (2007). Mitonuclear discordance is caused by rampant mitochondrial introgression in *Neodiprion* (Hymenoptera: Diprionidae) sawflies. *Evolution*, 61(6), 1417-1438. https://doi.org/10.1111/j.1558-5646.2007.00114.x

Linnen, C. R., & Farrell, B. D. (2008a). Comparison of methods for species-tree inference in the sawfly genus *Neodiprion* (Hymenoptera: Diprionidae). *Systematic Biology*, 57(6), 876-890. https://doi.org/10.1080/10635150802580949

1022 Linnen, C. R., & Farrell, B. D. (2008b). Phylogenetic analysis of nuclear and mitochondrial
1023 genes reveals evolutionary relationships and mitochondrial introgression in the sertifer
1024 species group of the genus *Neodiprion* (Hymenoptera: Diprionidae). *Molecular*
1025 *Phylogenetics and Evolution*, 48(1), 240-257.
1026 https://doi.org/10.1016/j.ympev.2008.03.021
1027 Linnen, C. R., & Farrell, B. D. (2010). A test of the sympatric host race formation hypothesis in
1028 *Neodiprion* (Hymenoptera: Diprionidae). *Proceedings of the Royal Society B: Biological*
1029 *Sciences*, 277(1697), 3131-3138. https://doi.org/10.1098/rspb.2010.0577
1030 Linnen, C. R., O'Quin, C. T., Shackleford, T., Sears, C. R., & Lindstedt, C. (2018). Genetic basis
1031 of body color and spotting pattern in redheaded pine sawfly larvae (*Neodiprion lecontei*).
1032 *Genetics*, 209(1), 291-305. https://doi.org/10.1534/genetics.118.300793
1033 Llaurens, V., Whibley, A., & Joron, M. (2017). Genetic architecture and balancing selection: the
1034 life and death of differentiated variants. *Molecular Ecology*, 26(9), 2430-2448.
1035 https://doi.org/10.1111/mec.14051
1036 Ma, T., Wang, K., Hu, Q., Xi, Z., Wan, D., Wang, Q., Feng, J., Jiang, D., Ahani, H., Abbott, R.
1037 J.,
1038 Lascoux, M., Nevo, E., & Liu, J. (2018). Ancient polymorphisms and divergence
1039 hitchhiking contribute to genomic islands of divergence within a poplar species complex.
1040 *Proceedings of the National Academy of Sciences*, 115(2), E236-E243.
1041 https://doi.org/10.1073/pnas.1713288114
1042 Makowski, D., Ben-Shachar, M. S., Patil, I., & Lüdecke, D. (2020). Methods and algorithms for
1043 correlation analysis in R. *Journal of Open Source Software*, 5(51), 2306.
1044 https://doi.org/10.21105/joss.02306

1045 Martin, S. H., Dasmahapaptra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F.,
1046 Blaxter, M., Manica, A., Mallet, J., & Jiggins, C. D. (2013). Genome-wide evidence for
1047 speciation with gene flow in *Heliconius* butterflies. *Genome Research*, 23(11), 1817-
1048 1828. https://doi.org/10.1101/gr.159426.113
1049 Matthey-Doret, R., & Whitlock, M. C. (2019). Background selection and F_{ST}: Consequences for
1050 detecting local adaptation. *Molecular Ecology*, 28, 3902-3914.
1051 https://doi.org/10.1111/mec.15197
1052 Maynard Smith, J., & Haigh, J. (1974). The hitch-hiking effect of a favorable gene. *Genetical*
1053 *Research*, 23(1), 23-35. https://doi.org/10.1017/S0016672300014634
1054 McGaugh, S., Smukowski, C., Manzano-Winkler, B., Himmel, T., & Noor, M. (2012).
1055 Recombination modulates how selection affects linked sites in *Drosophila*. *Nature*
1056 *Precedings*, 1-1. https://doi.org/10.1038/npre.2012.7005.1
1057 Miller, S. E., & Sheehan, M. J. (2023). Sex differences in deleterious genetic variants in a
1058 haplodiploid social insect. *Molecular Ecology*, 00, 1-11.
1059 https://doi.org/10.1111/mec.17057
1060 Moreira, L. R., Klicka, J., & Smith, B. T. (2023). Demography and linked selection interact to
1061 shape the genomic landscape of codistributed woodpeckers during the Ice Age.
1062 *Molecular Ecology*, 32, 1739-1759. https://doi.org/10.1111/mec.16841
1063 Mozhaitseva, K., Tourrain, Z., & Branca, A. (2023). Population genomics of the mostly
1064 thelytokous *Diplolepis rosae* (Linnaeus, 1758)(Hymenoptera: Cynipidae) reveals
1065 population-specific selection for sex. *Genome Biology and Evolution*, 15(10), evad185.
1066 https://doi.org/10.1093/gbe/evad185
1067 Muller, H. J. (1950). Our load of mutations. *American journal of human genetics*, 2(2), 111.

1068 Nachman, M. W., & Payseur, B. A. (2012). Recombination rate variation and speciation:
1069 Theoretical predictions and empirical results from rabbits and mice. *Philosophical*
1070 *Transactions of the Royal Society B: Biological Sciences*, 367(1587), 409-421.
1071 https://doi.org/10.1098/rstb.2011.0249
1072 Nosil, P., & Feder, J. L. (2012). Genomic divergence during speciation: Causes and
1073 consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*,
1074 367(1587), 332-342. https://doi.org/10.1098/rstb.2011.0263
1075 Nouhaud, P., Blanckaert, A., Bank, C., & Kulmuni, J. (2020). Understanding admixture:
1076 haplodiploidy to the rescue. *Trends in Ecology & Evolution*, 35(1), 34-42.
1077 https://doi.org/10.1016/j.tree.2019.08.013
1078 Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual Review of Ecology,*
1079 *Evolution and Systematics*, 23(1), 263-286.
1080 https://doi.org/10.1146/annurev.es.23.110192.001403
1081 Oyler-McCance, S. J., Cornman, R. S., Jones, K. L., & Fike, J. A. (2015). Z chromosome
1082 divergence, polymorphism and relative effective population size in a genus of lekking
1083 birds. *Heredity*, 115(5), 452-459. https://doi.org/10.1038/hdy.2015.46
1084 Padmanabhan, S., Thakur, J., Siddharthan, R., & Sanyal, K. (2008). Rapid evolution of Cse4p-
1085 rich centromeric DNA sequences in closely related pathogenic yeasts, *Candida albicans*
1086 and *Candida dubliniensis*. *Proceedings of the National Academy of Sciences*, 105(50),
1087 19797-19802. https://doi.org/10.1073/pnas.0809770105
1088 Payseur, B. A., & Nachman, M. W. (2002). Natural selection at linked sites in humans. *Gene*,
1089 300(1-2), 31-42. https://doi.org/10.1016/S0378-1119(02)00849-1

Pease, J. B., & Hahn, M. W. (2013). More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. *Evolution*, 67(8), 2376-2384. https://doi.org/10.1111/evo.12118

Piatkowski, B., Weston, D. J., Aguero, B., Duffy, A., Imwattana, K., Healey, A. L., ... & Shaw, A. J. (2023). Divergent selection and climate adaptation fuel genomic differentiation between sister species of *Sphagnum* (peat moss). *Annals of Botany*, 132(3), 499-512. https://doi.org/10.1093/aob/mcad104

Pouyet, F., Aeschbacher, S., Thiéry, A., & Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife*, 7, e36317. https://doi.org/10.7554/eLife.36317

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., & Camerini-Otero, R. D. (2014). DNA recombination. Recombination initiation maps of individual human genomes. *Science*, 346, 1256442. https://doi.org/10.1126/science.1256442

Presgraves, D. C. (2018). Evaluating genomic signatures of "the large X-effect" during complex speciation. *Molecular Ecology*, 27(19), 3822-3830. https://doi.org/10.1111/mec.14777

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M. A. F., Mehlig, B., & Westram, A. M. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, 30, 1450-1477. https:// doi: 10.1111/jeb.13047

1113 Rettelbach, A., Nater, A., & Ellegren, H. (2019). How linked selection shapes the diversity
1114 landscape in *Ficedula* flycatchers. *Genetics*, 212(1), 277-285.
1115 https://doi.org/10.1534/genetics.119.301991
1116 Rodrigues, M. F., Kern, A. D., & Ralph, P. L. (2023). Shared evolutionary processes shape
1117 landscapes of genomic variation in the great apes. bioRxiv.
1118 https://doi.org/10.1101/2023.02.07.527547
1119 Rougemont, Q., Moore, J.-S., Leroy, T., Normandeau, E., Rondeau, E. B., Withler, R. E., Van
1120 Doornik, D. M., Crane, P. A., Naish, K. A., Garza, J. C., Beacham, T. D., Koop, B. F., &
1121 Bernatchez, L. (2019). Demographic history, linked selection, and recombination shape
1122 the genomic landscape of a broadly distributed Pacific salmon. bioRxiv.
1123 https://doi.org/10.1101/732750
1124 Samuk, K., Owens, G. L., Delmore, K. E., Miller, S. E., Rennison, D. J., & Schluter, D. (2017).
1125 Gene flow and selection interact to promote adaptive divergence in regions of low
1126 recombination. *Molecular Ecology*, 26(17), 4378-4390.
1127 https://doi.org/10.1111/mec.14226
1128 Schrider, D. R. (2020). Background selection does not mimic the patterns of genetic diversity
1129 produced by selective sweeps. *Genetics*, 216(2), 499-519.
1130 https://doi.org/10.1534/genetics.120.303469
1131 Shang, H., Field, D. L., Paun, O., Rendón-Anaya, M., Hess, J., Vogl, C., Liu, J., Ingvarsson,
1132 P. K., Lexer, C., & Leroy, T. (2023). Drivers of genomic landscapes of differentiation
1133 across a *Populus* divergence gradient. *Molecular Ecology*, 00, 1-14.
1134 https://doi.org/10.1111/mec.17034
1135 Stankowski, S., Chase, M. A., Fuiten, A. M., Rodrigues, M. F., Ralph, P. L., & Streisfeld, M. A.

(2019). Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. *PLoS Biology*, 17(7), e3000391. https://doi.org/10.1371/journal.pbio.3000391

Sun, N., Yang, L., Tian, F., Zeng, H., He, Z., Zhao, K., Wang, C., Meng, M., Feng, C., Fang, C., Lv, W., Bo, J., Tang, Y., Gan, X., Peng, Z., Chen, Y., & He, S. (2022). Sympatric or micro-allopatric speciation in a glacial lake? Genomic islands support neither. *National Science Review*, 9, nwac291. https://doi.org/10.1093/nsr/nwac291

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), 585-595. https://doi.org/10.1093/genetics/123.3.585

Talla, V., Johansson, A., Dinca, V., Vila, R., Friberg, M., Wilklund, C., & Backström, N. (2019). Lack of gene flow: Narrow and dispersed differentiation islands in a triplet of *Leptidea* butterfly species. *Molecular Ecology*, 28, 3756-3770. https ://doi.org/10.1111/mec.15188

Turner, T. L., Hahn, M. W., & Nuzhdin, S. V. (2005). Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, 3(9), e285. https://doi.org/10.1371/journal.pbio.0030285

Vertacnik, K. L., Herrig, D. K., Godfrey, R. K., Hill, T., Geib, S. M., Unckless, R. L., ... & Linnen, C. R. (2023). Evolution of five environmentally responsive gene families in a pine-feeding sawfly, *Neodiprion lecontei* (Hymenoptera: Diprionidae). *Ecology and Evolution*, 13(10), e10506. https://doi.org/10.1002/ece3.10506

Via, S. (2012). Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1587), 451-460. https://doi.org/10.1098/rstb.2011.0260

Via, S., & West, J. (2008). The genetic mosaic suggests a new role for hitchhiking in ecological

speciation. *Molecular Ecology*, 17(19), 4334-4345. https://doi.org/10.1111/J.1365-294X.2008.03921.X

Wallberg, A., Glémin, S., & Webster, M. T. (2015). Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera*. *PLoS Genetics*, 11(4), e1005189. https://doi.org/10.1371/journal.pgen.1005189

Wang, L., Liu, S., Yang, Y., Meng, Z., & Zhuang, Z. (2022). Linked selection, differential introgression and recombination rate variation promote heterogeneous divergence in a pair of yellow croakers. *Molecular Ecology*, 31(22), 5729-5744. https://doi.org/10.1111/mec.16693

Wei, T., & Simko, V. (2021). R package 'corrplot': Visualization of a Correlation Matrix (Version 0.90). https://github.com/taiyun/corrplot

Wilson, L. F., Wilkinson, R. C., & Averill, R. D. (1992). *Redheaded pine sawfly: its ecology and management*. U.S. Dept. of Agriculture, Forest Service.

Wiuf, C., Zhao, K., Innan, H., & Nordborg, M. (2004). The probability and chromosomal extent of trans-specific polymorphism. *Genetics*, 168(4), 2363-2372. https://doi.org/10.1534/genetics.104.029488

Wolf, J. B., & Ellegren, H. (2016). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, 18, 87-100. https://doi:10.1038/nrg.2016.133

Wong Miller, K. M., Bracewell, R. R., Eisen, M. B., & Bachtrong, D. (2017). Patterns of genome-wide diversity and population structure in the *Drosophila athabasca* species complex. *Molecular Biology and Evolution*, 34(8), 1912-1923. https://doi.org/10.1093/molbev/msx134

Wu, C. I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, 14(6), 851-865. https://doi.org/10.1046/j.1420-9101.2001.00335.x

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8), 1586-1591. https://doi.org/10.1093/molbev/msm088

DATA ACCESSIBILITY AND BENEFIT SHARING STATEMENT

Trimmed (with trimmomatic) *Neodiprion* sequencing reads are published in the NCBI SRA database (BioProject accession number PRJNA1107580). All input files and scripts required for reproducing the analyses within the manuscript are published on DRYAD (doi: https://doi.org/10.5061/dryad.fxpnvx128).

Benefits Generated: Benefits from this study accrue from the sharing of our data and results on public databases as described above.

AUTHOR CONTRIBUTIONS

ANG prepared the resequencing dataset. ANG and CRL conceived of and designed the study, performed data analysis, and drafted the manuscript. VCS, RDR, SBS, and SMG contributed to data analysis and writing the manuscript. All authors have read and approved the final manuscript.

TABLES

Table 1. Predicted patterns for genomic statistics under different evolutionary scenarios. (A) Theory predicts that widespread linked selection across the genome will produce specific genome-wide correlations between gene density and recombination rate (factors that affect the intensity of linked selection) and genetic summary statistics. (B) Differences in the timing and nature of selection are expected to produce distinct genome-wide correlations between F_{ST}, d_{XY}, and mean π. Mean π refers to the average π for the two species included in the comparison. (C) Expected local patterns of F_{ST}, d_{XY}, and π compared to the genomic background for each of the four primary evolutionary scenarios considered in (B).

(A) Predicted genome-wide correlations between genomic summary statistics and genome features under linked selection

Statistic	Recombination rate	Gene density
π	positive	negative
Tajima's D	positive	negative
Fay & Wu's H	positive	negative
F_{ST}	negative	positive
d_{XY}[†]	positive	negative

(B) Predicted genome-wide correlations among summary statistics under four evolutionary scenarios

Scenario	F_{ST} vs. d_{XY}	Mean π vs. F_{ST}	Mean π vs. d_{XY}
Divergence-with-gene-flow	positive	negative	negative
Allopatric selection	none	negative	none
Recurrent selection	negative	negative	positive
Balancing selection	negative	negative	positive

(C) Predicted local patterns of summary statistics under four evolutionary scenarios

Scenario	Expected Patterns
Divergence-with-gene-flow	F_{ST}: high
	d_{XY}: high
	π: low
Allopatric selection	F_{ST}: high
	d_{XY}: average
	π: low
Recurrent selection	F_{ST}: high
	d_{XY}: low
	π: low
Balancing selection	F_{ST}: low
	d_{XY}: high
	π: high

[†]Predictions expected when linked selection occurred in the ancestral population.

Table 2. Effect size estimates and type II ANOVA tables for multiple linear regression models for genetic summary statistics. All predictor variables were normal-quantile transformed prior to running each model. Significant *p*-values ($p < 0.05$) are indicated in bold.

Response variable	Genomic predictor variable	Estimate	Sum Sq	df	F value	*p*-value
π - *N. lecontei*	recombination rate	4.97×10^{-5}	8.80×10^{-6}	1	21.43	**3.79×10^{-6}**
	exon density	-2.74×10^{-4}	2.51×10^{-4}	1	610.43	**$< 2.2 \times 10^{-16}$**
	mutation rate	2.42×10^{-5}	2.37×10^{-6}	1	5.78	**0.016**
	distance from centromere	-2.25×10^{-7}	0	1	0.0004	0.99
	site count	-1.69×10^{-4}	8.51×10^{-5}	1	207.08	**$< 2.2 \times 10^{-16}$**
	residuals		1.68×10^{-4}	4083		
π - *N. pinetum*	recombination rate	-8.20×10^{-5}	2.38×10^{-5}	1	40.61	**2.07×10^{-10}**
	exon density	-1.71×10^{-4}	1.00×10^{-4}	1	170.75	**$< 2.2 \times 10^{-16}$**
	mutation rate	-2.51×10^{-6}	3.00×10^{-8}	1	0.044	0.83
	distance from centromere	-1.64×10^{-4}	7.99×10^{-5}	1	136.40	**$< 2.2 \times 10^{-16}$**
	site count	-1.42×10^{-4}	6.49×10^{-5}	1	110.67	**$< 2.2 \times 10^{-16}$**
	residuals		0.0024	4083		
Tajima's D - *N. lecontei*	recombination rate	5.36×10^{-2}	10.26	1	47.34	**6.89×10^{-12}**
	exon density	-2.16×10^{-1}	155.91	1	719.67	**$< 2.2 \times 10^{-16}$**
	mutation rate	1.87×10^{-2}	1.42	1	6.56	**0.010**
	distance from centromere	2.30×10^{-2}	1.50	1	6.93	**8.52×10^{-3}**
	site count	-1.23×10^{-1}	44.83	1	206.92	**$< 2.2 \times 10^{-16}$**
	residuals		884.56	4083		
Tajima's D - *N. pinetum*	recombination rate	-1.58×10^{-2}	0.88	1	2.55	0.11
	exon density	-1.57×10^{-1}	83.56	1	241.57	**$< 2.2 \times 10^{-16}$**
	mutation rate	-4.13×10^{-4}	0	1	0.0020	0.96
	distance from centromere	-8.77×10^{-2}	22.93	1	66.28	**5.16×10^{-16}**
	site count	-1.27×10^{-1}	51.83	1	149.84	**$< 2.2 \times 10^{-16}$**
	residuals		1412.40	4083		
Fay & Wu's H - *N. lecontei*	recombination rate	-8.13×10^{-3}	0.24	1	4.79	**0.029**
	exon density	2.78×10^{-2}	2.59	1	52.56	**4.97×10^{-13}**
	mutation rate	-7.12×10^{-4}	0.002	1	0.042	0.84
	distance from centromere	-8.62×10^{-3}	0.21	1	4.29	**0.038**
	site count	4.85×10^{-2}	6.98	1	141.92	**$< 2.2 \times 10^{-16}$**
	residuals		200.83	4083		

Fay & Wu's H - *N. pinetum*	recombination rate	-5.67×10^{-3}	0.11	1	1.20	0.27
	exon density	5.52×10^{-2}	10.40	1	110.16	**$< 2.2 \times 10^{-16}$**
	mutation rate	-6.40×10^{-3}	0.17	1	1.76	0.19
	distance from centromere	-1.58×10^{-2}	0.75	1	7.91	**4.95×10^{-3}**
	site count	3.46×10^{-2}	3.85	1	40.81	**1.87×10^{-10}**
	residuals		385.31	4083		
F_{ST}	recombination rate	-8.94×10^{-3}	0.29	1	19.70	**9.31×10^{-6}**
	exon density	2.28×10^{-2}	1.76	1	121.94	**$< 2.2 \times 10^{-16}$**
	mutation rate	1.79×10^{-4}	0	1	0.0090	0.92
	distance from centromere	-1.07×10^{-3}	0.003	1	0.23	0.63
	site count	4.14×10^{-3}	0.053	1	3.68	0.055
	residuals		59.08	4083		
d_{XY}	recombination rate	-5.75×10^{-5}	1.18×10^{-5}	1	12.37	**4.41×10^{-4}**
	exon density	-2.96×10^{-4}	2.98×10^{-4}	1	312.71	**$< 2.2 \times 10^{-16}$**
	mutation rate	1.49×10^{-5}	9.00×10^{-7}	1	0.94	0.33
	distance from centromere	-1.67×10^{-4}	8.35×10^{-5}	1	87.70	**$< 2.2 \times 10^{-16}$**
	site count	-1.77×10^{-4}	1.02×10^{-4}	1	107.19	**$< 2.2 \times 10^{-16}$**
	residuals		3.89×10^{-3}	4083		

FIGURE LEGENDS

Figure 1. *Neodiprion lecontei* and *N. pinetum* as a model speciation genomics system. *Neodiprion lecontei* and *N. pinetum* have adapted to pine hosts with very different needle morphology, exhibiting differences in larval and adult traits that enhance fitness on their respective hosts (left panels show feeding larvae and ovipositing females of each species). *Neodiprion lecontei* and *N. pinetum* also exhibit strong but incomplete reproductive isolation and a history of divergence with gene flow (middle panel shows the best-fit demographic model estimated in Bendall et al. (2022), with the sizes of boxes and arrows proportional to effective population size and migration rates). Finally, *Neodiprion* pine sawflies are haplodiploid: females develop from fertilized eggs and are diploid; males develop from unfertilized eggs and are haploid (ploidy and morphology of adult females and males are shown in the last panel). Thus, in addition to excellent genomic resources, their ecology, demographic history, and haplodiploidy make *N. lecontei* and *N. pinetum* a good system for testing how these factors affect the genomic landscape of differentiation. Photos by Robin Bagley and Ryan Ridenbaugh.

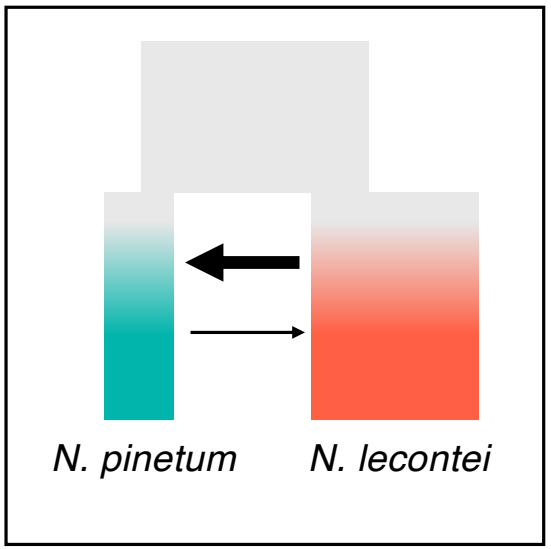
Figure 2. Patterns of genetic variation within and between *Neodiprion lecontei* and *N. pinetum*. All measures of divergence (F_{ST}, d_{XY}), diversity (π), selection (Tajima's D (*D*) and Fay & Wu's H (*H*)), recombination rate (cM/Mb), and exon density are highly heterogeneous across the genome. Alternating white and gray boxes separate the seven chromosomes. The green triangles and dotted purple lines indicate the estimated centromere midpoints.

Figure 3. Genome-wide correlations between pairs of statistics. Pearson's correlation coefficients between pairs of statistics describing genetic variation within and between *Neodiprion lecontei* and *N. pinetum* as well as genome features. Abbreviations: π (L) = *N. lecontei* π; π (P) = *N. pinetum* π; dS = synonymous substitution rate (proxy for the neutral mutation rate); cM/Mb = recombination rate; dist from cent = distance from the centromere. Significant correlations are indicated with asterisks (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$).

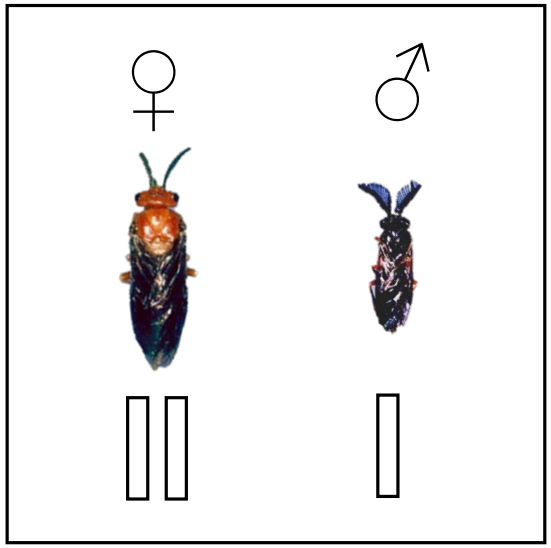
Figure 4. F_{ST}, d_{XY}, and π for *Neodiprion lecontei* and *N. pinetum* on chromosome 4. Windows that exhibit local patterns for all three summary statistics matching the expected pattern for one of the four evolutionary scenarios (Table 1C) are colored with bars. In each plot, the dotted lines represent regions where the windows were excluded from analysis due to low site count.

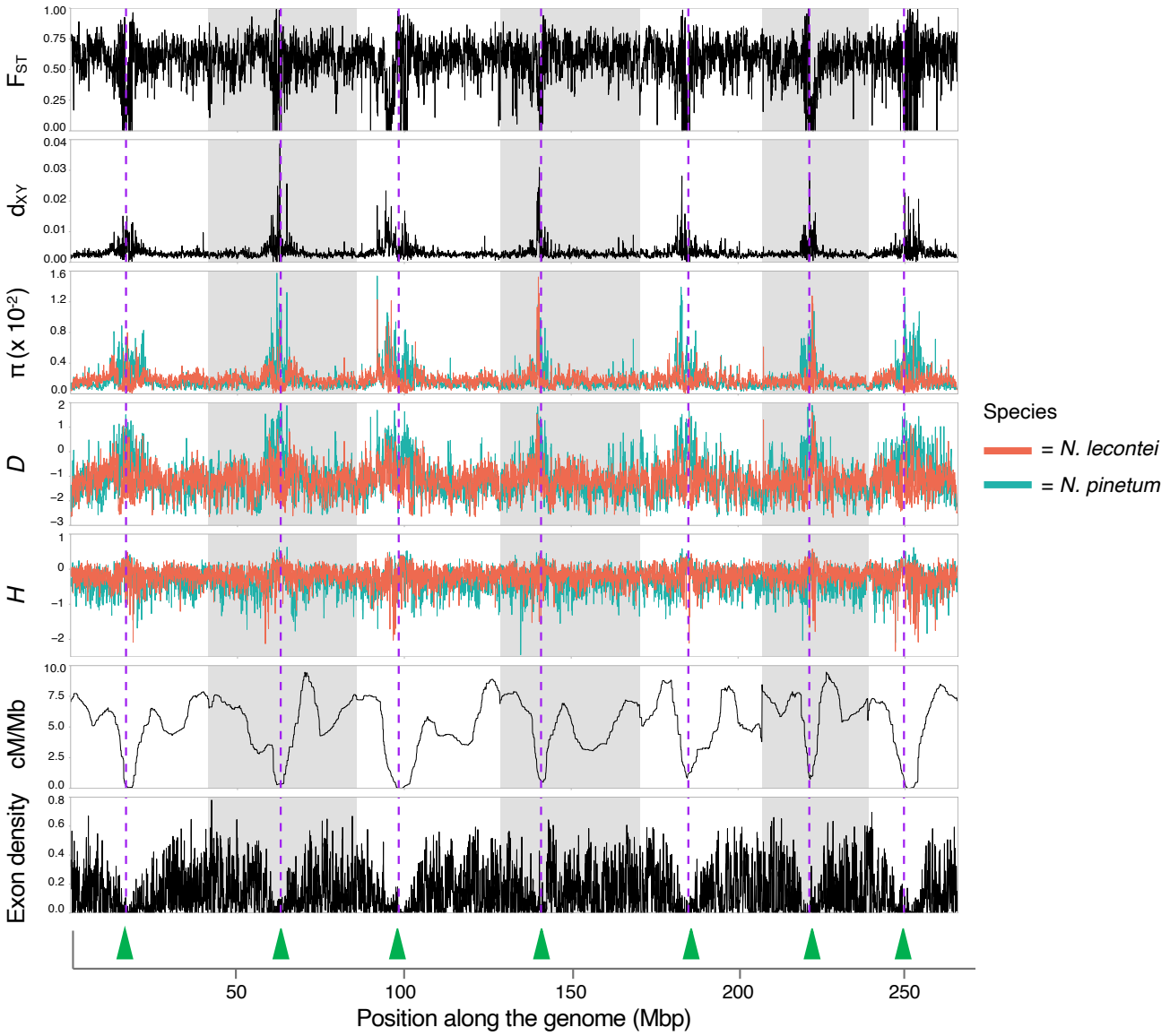
Divergent Natural Selection

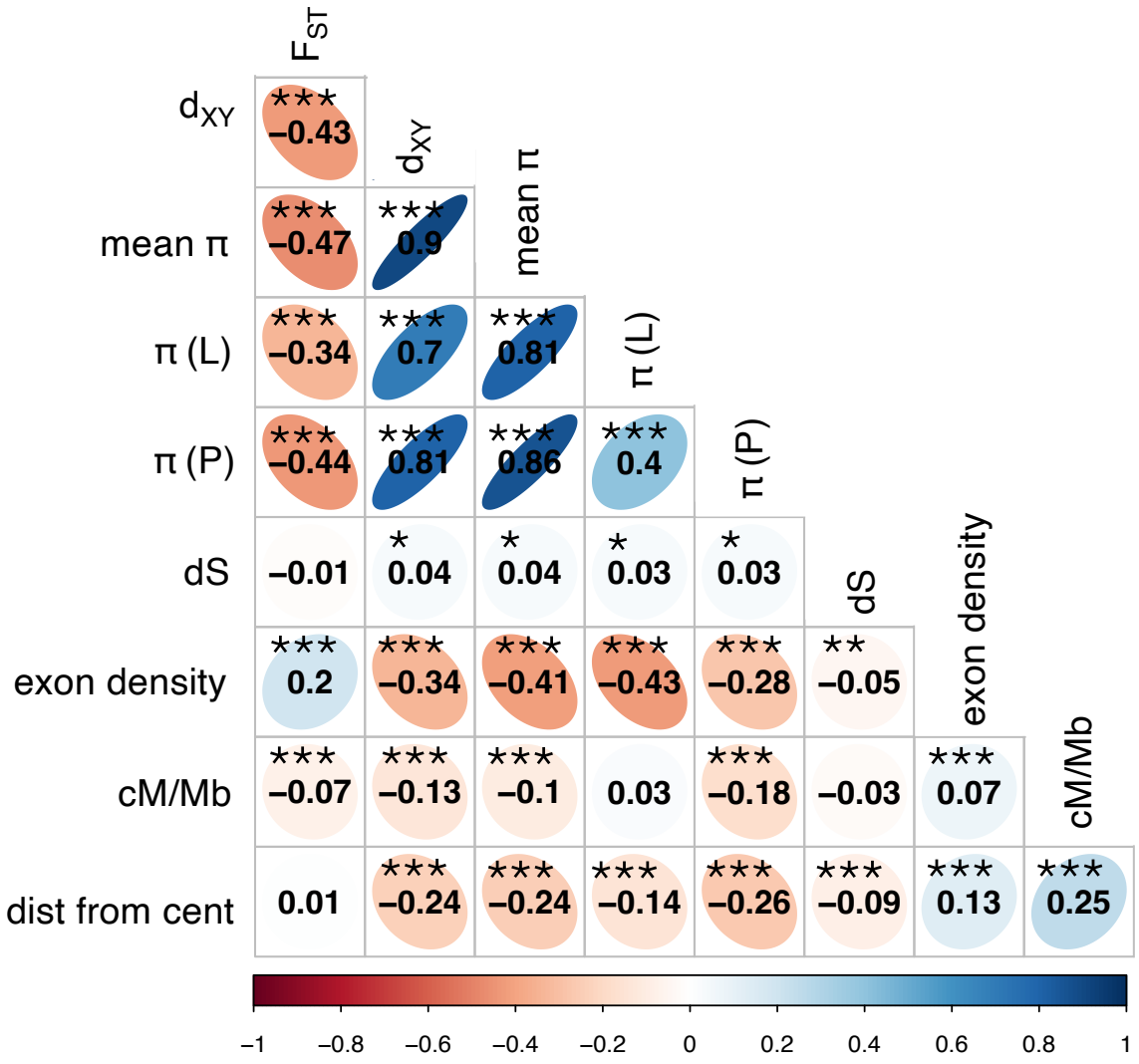
History of Divergence

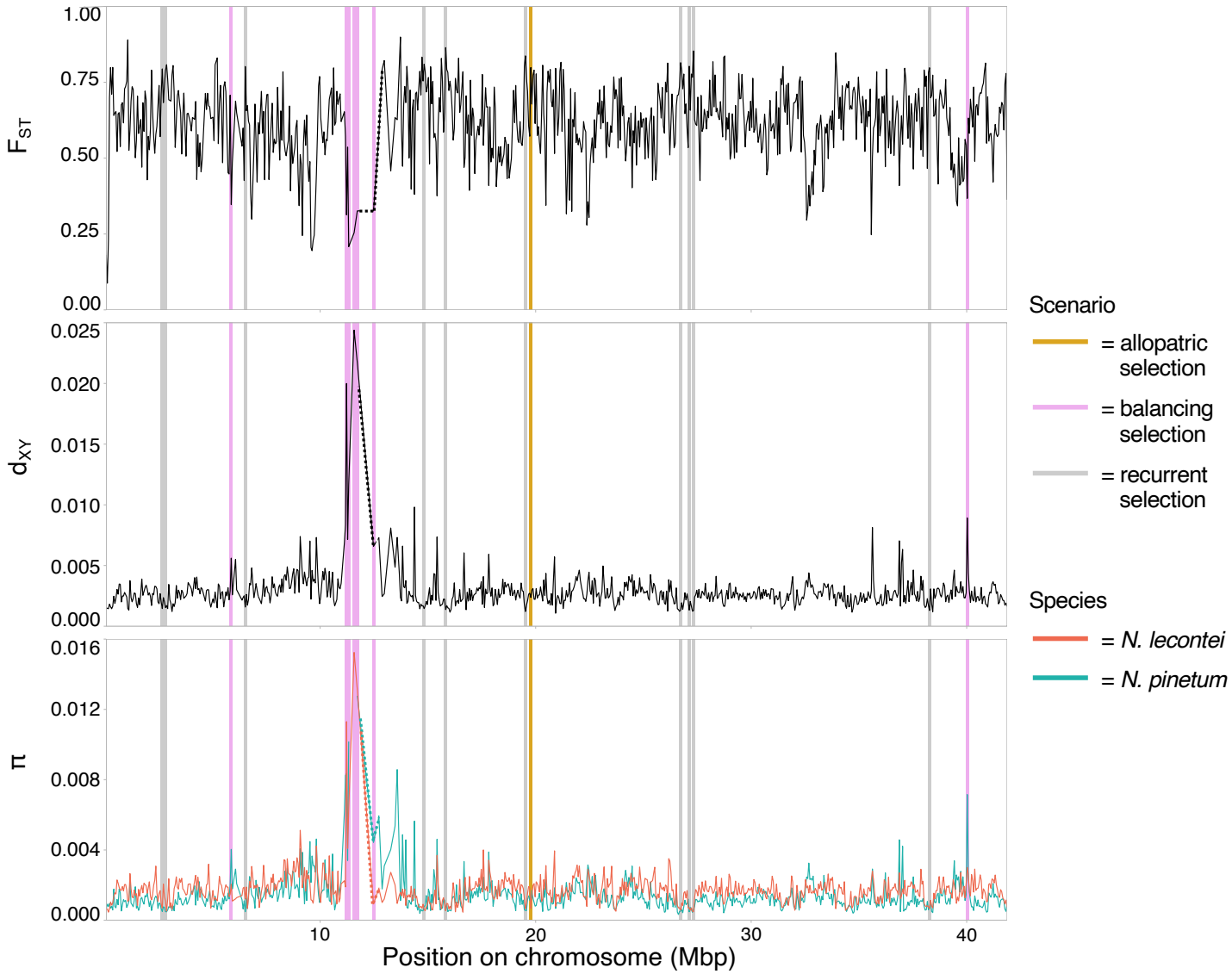


Haplodiploidy




F_{ST}
1.00
0.75
0.50
0.25
0.00
d_{XY}
0.04
0.03
0.02
0.01
0.00
π (x 10^{-2})
1.6
1.2
0.8
0.4
0.0
D
2
1
0
−1
−2
−3
H
1
0
−1
−2
cM/Mb
10.0
7.5
5.0
2.5
0.0
Exon density
0.8
0.6
0.4
0.2
0.0
50
100
150
200
250
Position along the genome (Mbp)
Species
= N. lecontei
= N. pinetum


F_{ST}
d_{XY}
mean π
π (L)
π (P)
dS
exon density
cM/Mb
dist from cent
*** −0.43
*** −0.47 *** 0.9
*** −0.34 *** 0.7 *** 0.81
*** −0.44 *** 0.81 *** 0.86 *** 0.4
−0.01 * 0.04 * 0.04 * 0.03 * 0.03
*** 0.2 *** −0.34 *** −0.41 *** −0.43 *** −0.28 ** −0.05
*** −0.07 *** −0.13 *** −0.1 0.03 *** −0.18 −0.03 *** 0.07
0.01 *** −0.24 *** −0.24 *** −0.14 *** −0.26 *** −0.09 *** 0.13 *** 0.25
−1 −0.8 −0.6 −0.4 −0.2 0 0.2 0.4 0.6 0.8 1


1.00
0.75
0.50
0.25
0.00
F_{ST}
0.025
0.020
0.015
0.010
0.005
0.000
d_{XY}
0.016
0.012
0.008
0.004
0.000
π
10
20
30
40
Position on chromosome (Mbp)
Scenario
= allopatric selection
= balancing selection
= recurrent selection
Species
= N. lecontei
= N. pinetum