Whole Genomes Reveal Evolutionary Relationships and Mechanisms Underlying Gene-Tree Discordance in *Neodiprion* Sawflies

Danielle K. Herrig[1], Ryan D. Ridenbaugh[1], Kim L. Vertacnik[1], Kathryn M. Everson[2,3], Sheina B. Sim[4], Scott M. Geib[4], David W. Weisrock[1], and Catherine R. Linnen[1*]

[1] Department of Biology, University of Kentucky, Lexington, KY, USA

[2] Department of Natural Resources and Environmental Science, University of Nevada, Reno, NV, USA

[3] Department of Integrative Biology, Oregon State University, Corvallis, OR, USA

[4] USDA-ARS Daniel K. Inouye US Pacific Basin Agricultural Research Center, Hilo, HI, USA

*Corresponding Author: catherine.linnen@uky.edu

ABSTRACT

Rapidly evolving taxa are excellent models for understanding the mechanisms that give rise to biodiversity. However, developing an accurate historical framework for comparative analysis of such lineages remains a challenge due to ubiquitous incomplete lineage sorting and introgression. Here, we use a whole-genome alignment, multiple locus-sampling strategies, and summary-tree and SNP-based species-tree methods to infer a species tree for eastern North American *Neodiprion* species, a clade of pine-feeding sawflies (Order: Hymenopteran; Family: Diprionidae). We recovered a well-supported species tree that—except for three uncertain relationships—was robust to different strategies for analyzing whole-genome data. Nevertheless, underlying gene-tree discordance was high. To understand this genealogical variation, we used multiple linear regression to model site concordance factors estimated in 50-kb windows as a function of several genomic predictor variables. We found that site concordance factors tended to be higher in regions of the genome with more parsimony-informative sites, fewer singletons, less missing data, lower GC content, more genes, lower recombination rates, and lower D-statistics (less introgression). Together, these results suggest that incomplete lineage sorting, introgression, and genotyping error all shape the genomic landscape of gene-tree discordance in *Neodiprion*. More generally, our findings demonstrate how combining phylogenomic analysis with knowledge of local genomic features can reveal mechanisms that produce topological heterogeneity across genomes.

KEYWORDS: pine sawflies, site concordance factors, introgression, incomplete lineage sorting, gene-tree discordance, phylogenomics

INTRODUCTION

It has long been recognized that gene trees need not match species trees due to stochastic sorting of ancestral polymorphism (i.e., incomplete lineage sorting), hybridization and introgression, horizontal gene transfer, and gene duplication and loss (Avise et al. 1987; Maddison 1997). An important implication of gene tree-species tree discordance is that data from many unlinked loci are necessary to accurately reconstruct divergence histories, but large molecular datasets were prohibitively expensive to generate using pre-genomic methodologies. Together, next-generation sequencing and the multi-species coalescent model have revolutionized molecular phylogenetics (Edwards 2009; McCormack et al. 2013; Edwards et al. 2016; Rannala et al. 2020). Next-generation sequencing made it possible to comprehensively sample the "cloud of gene histories" (Maddison 1997) embedded within a species tree. And by modeling the coalescent process that gives rise to these heterogeneous histories, the multi-species coalescent model (Rannala and Yang 2003) can improve species-tree inference (Ogilvie et al. 2017; Jiang et al. 2020; Rannala et al. 2020).

Despite much progress, a central challenge in species-tree analysis is that no model accounts for all possible sources of phylogenetic discordance. So-called "full-likelihood methods" (e.g., BEST and *BEAST; Liu 2008; Heled and Drummond 2010) accommodate both uncertainty in gene-tree estimation and incomplete lineage sorting (ILS) by simultaneously modelling nucleotide substitution and the coalescent process, but these approaches remain computationally burdensome (Liu et al. 2015; Rannala et al. 2020). In lieu of full-likelihood methods, many researchers use approximate approaches based on the multi-species coalescent model (e.g., MP-EST and ASTRAL, Liu et al. 2010; Zhang et al. 2018). These methods—sometimes referred to as summary-tree methods (Bryant and Hahn 2020)—use estimated gene trees for species-tree inference and assume that input gene trees are accurate and that there is no recombination within, but free recombination between loci. These assumptions create a trade-off

for sampling loci: shorter loci are more likely to satisfy the "no intralocus recombination" assumption, but less likely to yield sufficient information for accurate gene-tree inference (Chou et al. 2015). An alternative to having to define non-recombining loci is to use single nucleotide polymorphism (SNP)-based methods that assume free recombination between variable sites in the dataset (e.g., SNAPP and SVDquartets, Bryant et al. 2012; Chifman and Kubatko 2014).

Another source of phylogenetic discordance that is widespread in nature is introgression (Harrison and Larson 2014; Leaché et al. 2014; Fontaine et al. 2015; Mallet et al. 2016; Edelman et al. 2019; Hibbins and Hahn 2022). The multi-species coalescent model has been extended to include interspecific gene flow (Hey and Nielsen 2004; Yu et al. 2014), but full-likelihood implementations of these models (Jones 2018; Wen and Nakhleh 2018; Wen et al. 2018; Zhang et al. 2018a; Flouri et al. 2020) are computationally demanding (Flouri et al. 2020; Hibbins and Hahn 2022). For this reason, heuristic approaches for detecting introgression (e.g., SNaQ, ABBA-BABA tests, or HyDe; Green et al. 2010; Durand et al. 2011; Solís-Lemus et al. 2016; Blischak et al. 2018) are often used in conjunction with other species-tree methods that assume discordance is due to incomplete lineages sorting (e.g., Edelman et al. 2019; Meleshko et al. 2021). Although no species-tree method accounts for all sources of gene-tree discordance, coalescent-based methods nevertheless appear to perform reasonably well even when model assumptions are violated (Lanier and Knowles 2012; Chou et al. 2015; Adams et al. 2018; Long and Kubatko 2018; Borges et al. 2022; Yan et al. 2022).

In addition to creating new analytical challenges, genome-scale datasets also reveal that phylogenetic discordance is unevenly distributed across the genome (Pollard et al. 2006; White et al. 2009; Fontaine et al. 2015; Edelman et al. 2019; Small et al. 2020). Two potential sources of heterogeneous discordance are genotyping error (e.g., due to sequencing, alignment, and genotype-calling errors) and gene-tree estimation error, which may vary across the genome due to local base composition, repetitive sequence

content, and the density of phylogenetically informative sites (Betancur-R. et al. 2013). The genomic landscape of discordance is also likely to be influenced by the interplay between natural selection, gene flow, and recombination. For example, because selection on a locus reduces the effective population size (N_e) at linked neutral sites, regions of the genome evolving under positive or negative selection are expected to have reduced levels of ILS (Maynard Smith and Haigh 1974; Charlesworth et al. 1993; Slatkin and Pollack 2006; Stukenbrock et al. 2011; Pease and Hahn 2013; Dutheil et al. 2015). Because the effects of linked selection are most pronounced in low-recombination and gene-dense regions of the genome (Payseur and Nachman 2002; Charlesworth et al. 2009; Barton 2010; Cutter and Payseur 2013), such regions may have reduced phylogenetic discordance via ILS (Pease and Hahn 2013).

Low-recombination regions of the genome may also have reduced rates of introgression. This is because neutral or positively selected variants in low-recombination regions of the genome are more likely to be linked to deleterious alleles that prevent introgression (Nachman and Payseur 2012; Brandvain et al. 2014; Schumer et al. 2018; Li et al. 2019; Martin et al. 2019). Overall, research to date suggests that genealogical concordance (i.e., the extent to which gene trees match the underlying species tree) should correlate with genomic variables—such as the density of variable sites (parsimony-informative sites and sites for which a rare allele is present on one chromosome only [singletons]), the amount of missing data, gene density, base composition, and recombination rate (Table 1)—that influence the probability of genotyping error, gene-tree estimation error, incomplete lineage sorting, and introgression. However, the relative contribution of different genomic predictor variables to genome-wide variation in concordance with the underlying species tree remains an open question.

Here, we revisit a classic case study in messy species-tree inference (Linnen and Farrell 2007, 2008b, 2008a; Linnen 2010) armed with a high-quality reference genome, whole-genome resequencing data for 19 species, and newer species-tree methods. Specifically, we focus on the eastern North American "*Lecontei*" clade of *Neodiprion* sawflies (Order: Hymenoptera, Family: Diprionidae). Previous studies

suggest that this clade radiated with gene flow sometime within the last 2-10 million years (Linnen and Farrell 2007, 2008b; Bendall et al. 2022). Upon colonizing eastern North America, population divergence and speciation were likely driven by rapid adaptation to new pine (*Pinus*) hosts (Linnen and Farrell 2010; Bagley et al. 2017; Bendall et al. 2017; Glover et al. 2023). In addition to variation in host use, there is intra- and interspecific variation in larval and adult morphology, behavior, and overwintering strategy (Coppel and Benjamin 1965; Knerer and Atwood 1973). This phenotypic variation, coupled with a rich natural history literature and experimentally tractable species that can be reared and crossed in the lab (Knerer 1984; Bendall et al. 2017, 2023; Linnen et al. 2018), makes *Neodiprion* an excellent model for characterizing both the genetic mechanisms and evolutionary drivers of population differentiation and speciation. For accurate inferences about evolution in this group, an accurate species-tree estimate is essential.

The first informal phylogenetic hypothesis for *Neodiprion* consisted of five named species complexes (*lecontei*, *pinusrigidae, pratti*, *abbotii,* and *virginianus*) based on shared morphological and ecological traits (Ross 1955). Over fifty years later, these proposed species groups were evaluated with DNA sequence data from one mitochondrial locus and three nuclear genes. As expected under a scenario of rapid and recent divergence with gene flow, gene-tree topologies differ among the four loci (Linnen and Farrell 2007, 2008a). However, gene-tree discordance is especially pronounced between the mitochondrial locus and the three nuclear loci, likely due to extensive mitochondrial introgression (Linnen and Farrell 2007). To obtain a species-tree estimate from the remaining three nuclear loci, Linnen and Farrell (2008a) used multiple species-tree methods (Takahata 1989; Maddison 1997; Maddison and Knowles 2006; Edwards et al. 2007; Liu and Pearl 2007). Overall, these analyses yielded consistent support for two of Ross's proposed species groups (*lecontei* and *pinusrigidae*), mixed support for the *virginianus* and *pratti* species groups, and no support for the *abbotii* species group. Although different methods and subsets of individuals per species produced different topologies, Linnen and

Farrell (2008a) used areas of agreement across phylogenetic analyses to propose a provisional species tree for the *Lecontei* group (Fig. 1).

In this study, we expand the four-locus *Lecontei*-group dataset to the entire genome. Our study has three goals. First, we take advantage of a small, tractable genome (~272 Mb) to evaluate how inferred species-tree topologies are influenced by: (1) different analysis methods; specifically, concatenation vs. summary-tree (ASTRAL-III) and SNP-based (SVDquartets) species-tree approaches; and (2) subsampling the genome in ways that mimic reduced-representation approaches such as exon capture (approximated by sampling coding exons only) and restriction-associated DNA sequencing (approximated by subsampling SNPs). Second, we integrate all species-tree results to suggest an updated species tree for comparative work, highlighting areas of remaining uncertainty that are sensitive to sampling strategy and species-tree methodology. Third, we investigate genome-wide variation in concordance, measured using site concordance factors (sCFs), which describe the proportion of "decisive" sites for that are concordant with a focal branch in a reference species-tree topology (Minh et al. 2020a). Using multiple linear regression, we then evaluate several genomic variables that may predict genealogical concordance (Table 1). Based on our findings, we highlight priorities for future work on this system and make recommendations for other genome-wide phylogenomic analyses.

MATERIALS & METHODS

Assembly and Annotation of a Reference-Quality Genome for Neodiprion lecontei

DNA extraction and library preparation.—We obtained samples for DNA extraction from haploid male siblings that were the progeny of a single virgin *N. lecontei* female that had been collected in Lexington, KY (38°00'50.4"N 84°30'14.4"W) as a larva and lab-reared to adulthood. To maximize sawfly DNA yields and minimize host plant material in the gut, we flash-froze male larvae that were either in the final, non-feeding instar or dissected from freshly spun cocoons. We extracted genomic DNA from a single haploid male with a MagAttract HMW DNA Kit (Qiagen, Hilden Germany) using the fresh or frozen tissue protocol. To further improve sample purity, we performed a 2.0x bead clean-up using polyethylene glycol containing solid-phase reversible immobilization beads solution for each sample (DeAngelis et al. 1995). We quantified double-stranded DNA using a dsDNA Broad Range (BR) Qubit assay and assessed using the fluorometer feature of a DS-11 Spectrophotometer and Fluorometer (DeNovix Inc, Wilmington, DE, USA). We quantified DNA purity using the UV-Vis spectrometer feature on the DS-11, which reported OD 230/260/280 ratios.

We sheared DNA to a mean size distribution of ~20 kb using a Diagenode Megaruptor 2 according to the manufacturer's protocol, (Denville, New Jersey, USA) and sized sheared DNA on a Fragment Analyzer (Agilent Technologies, Santa Clara, California, USA) using the High Sensitivity (HS) Large Fragment kit. Sheared DNA was the starting input for PacBio SMRTBell library preparation using the SMRTbell Express Template Prep Kit 2.0 according to the manufacturer's protocol (Pacific Biosciences, Menlo Park, California, USA). The final library was shipped to the USDA-ARS Genomics and Bioinformatics Research Unit in Stoneville, Mississippi, USA where it was sequenced on one Pacific Biosciences 8M SMRT Cell on a Sequel II system (Pacific Biosciences, Menlo Park, California, USA) with a pre-extension time of 2 hours and a movie collection time of 30 hours.

In a parallel pipeline, we prepared an enriched chromosome conformation capture (HiC) library using another *N. lecontei* sample from the same haploid male family. Briefly, we crosslinked tissue using the Arima HiC low input protocol and performed proximity ligation using the Arima HiC Kit (Arima Genomics, San Diego, California, USA). After proximity ligation, we sheared the DNA using a Diagenode Bioruptor and then size-selected to enrich for DNA fragments of 200-600bp. We prepared an Illumina library from the sheared and size-selected DNA using the Swift Accel NGS 2S Plus kit (Integrated DNA Technologies, Coralville, Iowa, USA). The final Illumina HiC library was sequenced on a NovaSeq 6000 at the Hudson Alpha Genome Sequencing Center (Huntsville, Alabama, USA) with paired-end 150 bp (PE150) reads. We trimmed sequence reads for Illumina adapter artifacts using the Illumina BaseSpace software (Illumina, San Diego, California, USA).

Genome assembly.—Following sequencing, circular consensus sequence (CCS) calling was performed on the raw subreads generated by the Sequel II system using the SMRTLink v8.0 software (Pacific Biosciences, Menlo Park, California, USA). We filtered the resulting CCS reads for adapter contamination using the software HiFiAdapterFilt v2.0 (Sim et al. 2022). The filtered dataset served as input sequences for the HiFi data assembly software HiFiASM v0.16.1-r375 (Cheng et al. 2021) using default parameters. We converted the resulting contig assembly (.gfa format) to .fasta format using the software any2fasta (Seeman, 2018 https://github.com/tseemann/any2fasta).

To produce the HiC scaffolded assembly, we mapped paired Illumina HiC reads to the HiFi derived contig assembly using the mem function of the software BWA. We removed PCR duplicate artifacts from the resulting .sam file using the software samblaster (Faust and Hall 2014). We then used the resulting .bam file as the input file to the Phase Genomics Matlock suite of HiC functions (Kronenberg and Sullivan, 2018 https://github.com/phasegenomics/matlock) (Phase Genomics, Seattle, Washington, USA), which converted mapped reads into a HiC format that could be converted to a .hic and .assembly file using the

3d-dna script `run-assembly-visualizer.sh` and visualized using the software Juicebox v1.11.08 (Durand et al. 2016). Manual edits to the HiC scaffold assembly were performed using Juicebox v1.11.08 and changes were applied to the assembly using the Phase Genomics script `juicebox_assembly_converter.py` (https://github.com/phasegenomics/juicebox_scripts).

Assembly quality analysis.—To estimate genome size, we performed k-mer distribution analysis on the raw data using the k-mer counting software KMC (Deorowicz et al. 2013) and GenomeScope v.2.0 (Ranallo-Benavidez et al. 2020). To evaluate the assembly for duplicate contigs, we performed k-mer spectra analysis with the K-mer Analysis Toolkit (KAT) (Mapleson et al. 2016) using the raw data and the contig assembly. We evaluated duplicate content and genome completeness with BUSCO v5.0 in `genome` mode for the Metazoa, Arthropoda, Insecta, Endopterygota, and Hymenoptera ortholog sets (Manni et al. 2021). To characterize the read depth of each contig, we used the mapping software minimap2 (Li 2018) with the contig assembly and raw HiFi CCS read set. We used the NCBI nucleotide (nt) database (accessed 2017-0605) and the UniProt Reference Proteomes database (accessed 2020-03) to assign each contig to a taxonomic class using BLAST+ (Camacho et al. 2009) in blastn mode and Diamond (Buchfink et al. 2021) in blastx mode, respectively. Results from the hierarchical BUSCO analysis, read depth analysis, and taxon assignment were visualized using BlobTools2 (Laetsch and Blaxter 2017), and summarized using the python script blobblurb (https://github.com/sheinasim/blobblurb).

Gene annotation.—The NCBI *Neodiprion lecontei* Annotation Release 101 was completed using the NCBI Eukaryotic Genome Annotation Pipeline Software version 9.0. Briefly, BUSCO v4.1.4 was run in protein mode on the annotated gene set and the longest gene was retained. WindowMasker (Morgulis et al. 2006) was used to mask the genome (26.24% masked). Previously deposited transcriptome sequences from *Neodiprion* (including 77 specific tissues isolated from *N. lecontei* at five life stages from Herrig et al. 2021), RefSeq proteins, and GenBank Insecta proteins were then aligned to the masked genome using Splign, minimap2, or ProSplign (Kapustin et al. 2008; Li 2018). The alignments were then passed to Gnomon (Souvorov et al. 2010) for gene predictions. 14,732 genes and pseudogenes were identified in this process including 11,969 protein-coding genes.

Taxonomic Sampling, Library Preparation and Sequencing, and Reference-Anchored Alignment

We extracted fresh DNA from ethanol-preserved exemplars from 19 *Neodiprion* species, often from the same individuals or colonies as used in earlier phylogenetic studies of this genus (Linnen and Farrell 2007). Larval individuals had been collected in the United States and Canada (Table S1; this and all supplementary material can be found in the Dryad data repository: https://doi.org/10.5281/zenodo.11154212) and stored in ethanol at -20ºC. Sampling included all species in the eastern North American "*Lecontei*" clade except *N. insularis* and *N. cubensis*, both endemic to Cuba. We included the western North American species *N. autumnalis* an outgroup (Linnen and Farrell 2008b). Based on the presence of heterozygous sites in Sanger-sequence data from three nuclear loci, all individuals included in this study were diploid.

We dissected tissue from the prolegs and the ventral region of the larvae, avoiding the gut region. We then ground liquid nitrogen-frozen tissue with pestles made from 1-mL micropipette tips, and incubated the resulting powder in CTAB buffer with proteinase K and RNase A. We extracted DNA using phenol-

chloroform-isoamyl alcohol, dried the ethanol precipitate overnight, and resuspended in TE buffer. We assessed DNA integrity with a 0.7% agarose gel, DNA purity with the 260/280 ratio, and DNA concentration with a Quant-iT dsDNA High-Sensitivity fluorescence assay (Thermo Fisher Scientific). The Georgia Genomics and Bioinformatics Core (Athens, GA, USA) prepared and sequenced one small-insert DNA library for each species. Libraries had a mean fragment size of 619 bp and were sequenced on Illumina NextSeq 500 with PE150 reads. Sequencing produced 14-27 million filtered reads per individual.

To obtain a multi-genome alignment, we used a pseudo-reference-based approach, with our annotated, reference quality *N. lecontei* genome (described above) serving as the reference. This approach is appropriate for this clade because synteny is high and genetic divergence among species is low (see results). Our pipeline was based on the published Pseudo-It pipeline (Sarver et al. 2017) (Fig. S1) with some modifications. Briefly, we first used bowtie2 v2.4.1 (Langmead and Salzberg 2012) to map reads from each species to the *N. lecontei* reference genome. To allow for divergence between reads and the *N. lecontei* reference, we initially allowed a mismatch in the seed and "local" mapping options in bowtie2. New variants (excluding indels) were incorporated using samtools v1.10 (Li et al. 2009) and bcftools v1.10.2 (Li 2011). In a second round of mapping, this process was repeated using the first iteration of the genome for each species as the new reference genome. The third round of mapping removed the seed mismatch. The fourth and fifth iterations required end-to-end mapping. After the fifth iteration, we replaced any nucleotide that had a read depth less than 4 or that had excessively high mapping depth (highest 1% of depths for each species) with an "N" using a custom script. Heterozygous genotypes were indicated using the IUPAC nucleotide code. Unless otherwise noted, all bioinformatics commands and scripts can be found on Zenodo and the LinnenLab GitHub page under the Herrig_etal_NeodiprionPhylogeny repository.

To calculate genetic distance between species, we used the dist.dna command within the R v3.6.2 package ape v5.5 (Paradis and Schliep 2019) and default parameters to calculate Kimura's two-parameter (K80) distance (Kimura 1980) between every pair of species.

Dataset Preparation and Phylogenetic Inference

Our genome assembly and pseudo-reference approach produced 20 aligned genomes: the *de novo N. lecontei* genome and 19 pseudo-reference genomes (18 *Lecontei* clade species plus outgroup *N. autumnalis* from the western North American *Sertifer* group). To explore how genomic sampling strategy and analysis method affects species-tree inference, we sampled the aligned genomes in three ways. First, we used bedtools v2.30.0 (Quinlan 2014) to divide the seven *Neodiprion* chromosomes into non-overlapping windows of 50 kb. To evaluate the effect of window size on species-tree inference, we also generated non-overlapping windows of 10 kb, 100 kb, 500 kb, and 1000 kb. Because some regions of the genome had high levels of missing data, we used custom python scripts to exclude windows for which the amount of missing data (Ns in the alignment) was more than 10%. We then used IQ-TREE v2.2.2.6 (Minh et al. 2020b) to estimate a maximum-likelihood (ML) tree for the entire dataset (concatenated tree) and a gene tree for each window, in both cases allowing IQ-tree to select the best-fit substitution model for each window. We then used the estimated gene trees as input for ASTRAL-III v5.7.3 (Zhang et al. 2018b) to produce a single coalescent-based species tree for each of the windowed datasets. To visualize variation in gene tree topologies among 50-kb windows, we used DensiTree v3.0.2 (Bouckaert 2010). To prepare gene trees for visualization, the R package ape v5.7-1 (Paradis and Schliep 2019) was used to root trees using *N. autumnalis* as an outgroup and to standardize branch lengths via the "compute.brlen" command and the "Grafen" method (Grafen 1989).

Second, to approximate a dataset of protein-coding genes analogous to an RNAseq or exon-capture phylogenomic dataset, we used gffread v0.11.7 with the –w flag to write FASTA files with spliced exons

for each transcript for each species using the NCBI *Neodiprion lecontei* Annotation Release (iyNeoLeco1.1 RefSeq GCF_021901455.1). Custom scripts were used to define and keep the isoform with the most parsimony informative sites. We again used IQ-TREE to estimate a ML gene tree each gene, excluding genes containing fewer than 10 parsimony-informative sites. Finally, we used ASTRAL-III to estimate a species tree for the protein-coding dataset as described above.

Third, we called SNPs across the entire genome using SNP-sites v2.5.1 (Page et al. 2016). We then filtered the data to exclude SNPs that were absent in more than 10% of species and sites with more than two alleles. In addition to analyzing all SNPs (which likely contains tightly linked sites), we produced additional datasets with one SNP sampled every 1 kb, 5 kb, 10 kb, 50 kb, or 100 kb using SNP-sites, with more sparsely sampled SNPs on par with a dataset that might be generated via RADseq. We transformed each of the six datasets into nexus format and used SVDquartets (Chifman and Kubatko 2014), implemented in PAUP v4.0a (Swofford 2000), to produce a species tree for each dataset. Heterozygous sites were handled using the "Distribute" option.

To generate comparable measures of topological variation among sites for branches in species trees produced by the different methods and datasets, we used IQ-TREE v2.2.2.6 (Minh et al. 2020b) to generate ML estimates of sCFs (Mo et al. 2023) for each branch in a focal species tree, using the same alignment used to produce the corresponding species tree (e.g., an alignment of SNPs was used for the SVDquartets tree). For methods that use gene trees as input, we also estimated gene concordance factors to quantify topological variation among loci.

IQ-TREE analyses also produced phylogenies with branch lengths scaled to substitutions per site; however, introgression and hemiplasy cause biases in branch length estimation (Hibbins et al. 2020). To correct for these biases and convert branch lengths to coalescent units, we used the function 'subs2coal' within the python package HeIST v.0.3.1 (Hibbins et al. 2020). This tool redistributes branch lengths

based on information from gene- or site-concordance factors. We used our 50-kb window phylogeny with sCF values as input, and the resulting tree was visualized using FigTree v.1.4.4 (Rambaut 2007).

To evaluate the potential for introgression to contribute to topological variation revealed by DensiTree and concordance-factor analyses, we used the f-branch metric (Malinsky et al. 2018), which detects gene flow among tips as well as internal branches on the phylogeny. To perform this analysis, we first used the 'DtriosCombine' function in the Dsuite v0.4 package (Malinsky et al. 2021) to calculate genome-wide f4-ratio statistics for all possible trios of species, with the 50-kb ASTRAL-III species-tree as the guide tree. The resulting f4-ratio values were then used as input for the Dsuite command 'Fbranch' with default settings. The f-branch statistics were visualized using the 'dtools.py' python utility included with the Dsuite package.

Genomic Correlates of Concordance

To quantify how concordance with a focal species tree varies across the genome, we used IQ-TREE v2.2.2.6 to calculate ML site concordance factor estimates for each internal branch in the species tree for each 50-kb window, with the corresponding ASTRAL-III species tree as the focal tree. To visualize how topological concordance varies across the genome for each internal node of the species tree, we used chromPlot v1.26.0 (Oróstica and Verdugo 2016) and colorspace v2.1-0 to paint the chromosomes based on observed site concordance factors. All regression and painting analyses were performed in R v4.3.3 (R Core Team 2024).

For these same 50-kb windows, we then generated summary statistics corresponding to the predictor variables in Table 1. First, to describe variation in each 50-kb window, we used summaries of the number of parsimony-informative sites and singletons produced in IQ-TREE log files. Second, to quantify missing data for each window, we calculated the proportion of sites in each 50-kb alignment that were Ns. Third, to quantify GC content for each window, we calculated the proportion of nucleotides in each

alignment that were G or C. Fourth, to calculate gene density for each window, we used the NCBI *Neodiprion lecontei* Annotation Release 1.1 (iyNeoLeco1.1 RefSeq GCF_021901455.1) to define the number of genes with start codons within each window using custom scripts.

Fifth, to estimate local recombination rate for each window, we used data from a mapping population that was previously used to identify quantitative trait loci for larval color traits that differ between two populations of *N. lecontei* (Linnen et al. 2018). The original analysis consisted of 503 SNPs genotyped in 429 F_2 haploid males (like all hymenopterans, *Neodiprion* are haplodiploid) generated via double-digest restriction-associated DNA sequencing (Peterson et al. 2012). To increase marker density, we first mapped the raw sequencing reads (NCBI BioProject PRJNA434591, SRR6749156) to the new *N. lecontei* genome assembly (iyNeoLeco1.1). To identify fixed differences between the two parental populations, we first mapped reads from the cross parents (4 males; 4 females) and 10 F_1 females to the reference genome using BWA-MEM v0.7.17 (Li 2013) with the –M option, followed by samtools v1.10 (Li et al. 2009) to convert the sam output to a bam file. SNPs were called using mpileup in bcftools v1.9 (Li et al. 2009). We then used vcftools v0.1.16 (Danecek et al. 2011) to remove sites with indels and sites that were missing genotype information for more than 50% of the parents. To retain only putative fixed differences between the parent populations, we used vcftools to calculate the F_{ST} for each site and a custom python script to retain sites with $F_{ST} = 1$, resulting in 21,887 SNPs. To further validate that these SNPs were fixed between populations, we required that at least one of the F_1 females was heterozygous and that none of the F_1 females were homozygous at a read depth of 10. We used a custom python script to drop any sites that did not meet these criteria, reducing our dataset to 13,462 SNPs. We next mapped F_2 haploid male reads to the reference genome and called SNPs using BWA, samtools, and bcftools. For genotyping, we required a minimum read depth of 4 at each site and removed sites that had more than 50% missing information and that did not pass filtering in parents and F_1 females using

custom scripts. In total, we retained 3,104 ancestry-informative SNPs that were called in at least 50% of F_2 males.

We used R/qtl (Broman et al. 2003) to remove sites with identical genotypes to another site, sites exhibiting segregation distortion, and individuals that were missing genotypes at more than 50% of markers, producing a final dataset of 1436 markers genotyped in 402 F_2 males. Marker location was inferred based on physical location in the reference genome. To estimate a genetic map with positions of markers in centiMorgans (cM), we used the "quickEst" function in ASMap v1.0-4 (Taylor and Butler 2017), with a Kosambi mapping function. To estimate local recombination rate, we used MareyMap Online (Rezvoy et al. 2007; Siberchicot et al. 2017) with the sliding window option. We then used the recombination rate estimate for the midpoint of each window for our regression analysis.

Our final genomic variable was the D-statistic (or ABBA-BABA statistic), which we calculated for each internal branch for every 50-kb window. To calculate this statistic, we used the two site discordance factors (sDF1 and sDF2) produced by IQ-TREE. These values give the percentage of sites within a window that support the two alternative resolutions of a focal branch. Under ILS, these percentages are expected to be roughly equal; strong deviations from this expectation could be produced by gene flow (Green et al. 2010; Patterson et al. 2012). We calculated D-statistics describing the magnitude of skew in discordant topologies using the following formula: |(sDF1*sN)-(sDF2*sN)/(sDF1*sN)+(sDF2*sN)|, where sDF1, sDF2, and sN are the two site discordance factors and the number of decisive sites for a particular branch in the species tree.

To determine which genomic summary statistics predicted topological concordance, we fit a linear model for each internal branch in the window-based ASTRAL-III species tree, with all seven summary statistics (number of parsimony-informative sites, number of singletons, proportion missing data, GC content, D-statistic, gene density, and recombination rate) as predictor variables. To ensure all predictor

variables were on the same scale, we applied a normal-quantile transformation to each variable prior to fitting the model. To select a model that best explains the data without including unnecessary variables, we used the "step" function in R to perform bi-directional stepwise model selection via the Akaike information criterion (AIC). To assess correlations between variables, we used ggplot2 v3.3.5 and the GGally v2.1.2 extension to create a scatterplot matrix and estimate Pearson correlation coefficients and p-values.

RESULTS

Neodiprion De Novo *and Pseudo-Reference Genome Assemblies*

The *N. lecontei* iyNeoLeco1.1 assembly was sequenced to 93x coverage of PacBio HiFi reads and produced an assembly size of 272.074 MB in 106 scaffolds from 168 contigs with only 2% of the genome represented in gaps. Additional assembly statistics such as contig and scaffold N/L50 and N/L90 can be found in Table S2. Assessment for genome completeness using BUSCOs revealed that for all relevant databases from Eukaryota, Metazoa, Arthropoda, Insecta, Endopterygota, and Hymenoptera, genome completeness estimates ranged from 95.2% (Hymenoptera ortholog database v.10) to 99.6% (Eukaryota ortholog database v.10) (Table S3). All BUSCOs were found in assembled chromosomes, and none were in unplaced contigs.

For the remaining species, an average of 96.2% (range: 93.5-97.7%) of sequencing reads mapped to the *N. lecontei* reference genome, resulting in a final average coverage of 20.1x (range: 16x-28x) after removing regions with low or unusually high coverage (Table S4). The high mapping rates are likely attributable to low overall genetic divergence among eastern North American *Neodiprion* species: across all species, genome-wide average pairwise genetic distance (K80) was 0.0047 (range: 0.0003 - 0.0093) (Table S5).

The *N. lecontei* genome and our 19 reference-based whole-genomes produced a 50-kb window dataset, a "gene" dataset, and six SNP datasets, summarized in Table S6 (all data are available on Dryad https://doi.org/10.5061/dryad.bg79cnpf7). For the 50-kb window dataset, ML analysis of concatenated data and a summary-tree method (ASTRAL-III) produced identical topologies (Fig. 2a, 2b). We also explored the effect of window-size on ASTRAL-III species-tree estimates and found that windows of size 10-kb, 100-kb, 500-kb, and 1000-kb produced topologies identical to the 50-kb dataset. An ASTRAL-III tree estimated using 10,601 coding loci differed from the 50-kb window tree only in the placement of *N. compar* and *N. dubiosus* (Fig. 2c). Finally, a topology produced by our "all SNPs" dataset (i.e., 13,732,314 SNPs that passed quality and completeness thresholds) differed from both ASTRAL-III trees in the placement of *N. dubiosus* and the *pratti* species complex (Fig. 2d).

We also examined the effect of subsampling SNPs on species-tree inference in SVDquartets. A dataset consisting of SNPs sampled every1 kb (212,496 SNPs) produced a tree identical to the "all-SNPs" tree (Fig. S2a). SNPs sampled at 5-kb and 10-kb intervals (46,473 SNPs and 12,427 SNPs, respectively) produced a topology that differed from the all-SNPs tree in the placement of the *pratti* group (Fig. S2b). Unique topologies were also produced by the most sparsely sampled SNP datasets (1 per 50-kb = 5,131 SNPs; 1 per 100-kb = 2,611 SNPs; Fig. S2c,d).

Across all species trees (except the smallest SNP dataset), we consistently recovered five primary clades: *N. lecontei* + *N. pinetum* (*lecontei* species complex); *N. excitans* + *N. hetricki* + *N. pinusrigidae* + *N. swainei* (*pinusrigidae* species complex); *N. maurus* + *N. pratii* + *N. taedae* (*pratti* species complex), *N. abbotii* + *N. dubiosus* + *N. fabricii* + *N. nigroscutum* (*abbotii* species complex minus *N. compar* and plus *N. dubiosus*), and *N. knereri* + *N. merkeli* + *N. rugifrons* + *N. virginiana* + *N. warreni* (*virginianus* complex minus *N. dubiosus*) (Fig. 2 and Fig. S2). Most of the relationships within these clades were also highly stable across analyses and datasets. However, three relationships were sensitive to sampling approach

and analysis method: 1) the placement of *N. dubiosus* within the *abbotii* clade, 2) the placement of *N. compar* relative to remaining *Lecontei* group species, and 3) the placement of the *pratti* clade relative to the *virginianus* and *abbotii* clades. Not surprisingly, these relationships also tended to have the lowest site and gene concordance factors across the trees, regardless of dataset (Fig. 2, Fig. S2).

To visualize variation in gene-tree topologies, we plotted gene trees estimated from 50-kb windows using DensiTree (Fig. 3). Consistent with generally low to moderate concordance factors for branches in our species-tree estimates (Fig. 2, Fig. S2), the DensiTree plot revealed considerable heterogeneity among gene trees. Although heterogeneity is expected under incomplete lineage sorting, f-branch statistics reveal evidence of introgression in areas of the tree with especially pronounced site and gene-tree discordance. For example, the highest f-branch scores involve taxa from the *virginianus, abbotii,* and *pratti* species complexes (e.g., *N. rugifrons*, *N. abbotii*, *N. nigroscutum*, and *N. pratti*; Fig. 3), clades that tend to have lower concordance factors and tend to be more unstable across analyses (Fig. 2). Notably, however, there is minimal signal of introgression between *N. compar* and other species (Fig. 3). Thus, the uncertainty in the placement of this species may stem instead from factors that increase incomplete lineage sorting.

Genomic Correlates of Concordance

To quantify concordance across branches in the *Neodiprion* species tree and across the genome, we extracted site concordance factors for every 50-kb window and every branch in the corresponding species tree (Fig. 2b). To visualize variation in concordance, we painted the chromosomes according to the 50-kb site concordance factors for each branch (Fig. 4; to improve interpretability, branch lengths are scaled to coalescent units). Consistent with variable genome-wide concordance factors across branches in the 50-kb window species tree (Fig. 2b), this visualization revealed extensive variation in

concordance across clades, with some clades having much higher concordance levels on average than others (Fig. 4). For each clade, there were also variable levels of site concordance across the genome, with areas of relatively high and low concordance spread fairly evenly across the chromosomes (Fig. 4). There were two notable exceptions to this pattern. For some clades, Chromosome 1 or Chromosome 7 (or both) had areas of noticeably more uniform site concordance values compared to the rest of the genome. These areas, denoted by stars in Figure 4, could be explained by inversions that restrict recombination in these genomic locations. Notably, the putative inversion on Chromosome 1 is supported by independent evidence from synteny plots generated from recently produced chromosome-level assemblies for four *Neodiprion* species (Fig. S3). Comparable assemblies from additional species are needed to evaluate the putative inversion on Chromosome 7.

To explore factors that give rise to heterogeneous site concordance factors across the genome and species tree, we quantified several genomic predictor variables for each 50-kb window. To estimate local recombination rate, we constructed a new genetic map using sequencing reads from a previous cross between *N. lecontei* populations (Linnen et al. 2018). By mapping reads to a new reference genome, we nearly tripled the number of markers (from 503 to 1436) and decreased the average spacing between markers from 2.4 cM (maximum spacing = 24.3 cM) to 0.9 cM (maximum spacing = 13.2 cM). The total map length was 1271.6 cM. With an assembled genome size of ~272 Mb, this corresponds to a genome-wide recombination rate of 4.68 cM/Mb, a slightly higher rate than was previously reported (3.43 cM/Mb; Linnen et al. 2018). Plotting genetic distance as a function of physical distance revealed an even coverage of markers across chromosomes; gaps in the new genetic map corresponded to low-recombination centromeric regions (Figs. S4, S5). Using these data, local recombination rates were estimated via sliding windows, revealing heterogeneous recombination rates across the genome (Figs. S4, S5).

We also investigated pairwise correlations between all genomic predictor variables (Fig. S6). The strongest correlation we observed was between the number of parsimony informative sites and the number of singletons. These statistics also correlated positively with the amount of missing data and negatively with GC content and gene count. Recombination rate was positively correlated with missing data, GC content, and gene count (Fig. S6).

To determine which genomic variables best predict variation in concordance across the genome for each of the 17 branches in our species tree, we performed stepwise regression for each clade. These regression results are summarized in Figure 5, Table S7, and Table 2. To improve interpretability, branch lengths in Figure 5 are scaled to coalescent units. Overall, the amount of variation in site concordance factors explained by our regression models ranged from 3% to a 62%, with the highest value observed for the clade containing *N. pinetum + N. lecontei* (notably, the source species for recombination rate and gene density estimates). In general, site concordance factors were highest in 50-kb windows with: more parsimony-informative sites, fewer singletons, less missing data, lower GC content, higher D-statistics, more genes, and lower recombination rates (Table 2). Overall, these findings fit our predictions outlined in Table 1.

Discussion

Whole-genome datasets provide a comprehensive sample of heterogeneous histories, and the multi-species coalescent provides a framework for modeling sources of heterogeneity. Here, we combine a high-quality reference genome with whole-genome alignments for 20 pine sawfly species to achieve three goals: 1) determine the effect of sampling strategy and analysis method on species-tree estimation from whole-genome data, 2) estimate a species tree for eastern North American *Neodiprion* species, and 3) identify genomic predictors of gene-tree heterogeneity and phylogenetic concordance.

Below, we discuss progress on each of our three goals and possible explanations for regression results that deviate from predictions outlined in Table 1, while also highlighting both the broader implications and limitations of our analyses.

Species-Tree Estimates Mostly Robust to Locus-Sampling Strategy and Analysis Method

Although there is growing body of research investigating how marker-sampling strategies and species-tree analysis methods impact the accuracy of species-tree estimates, many outstanding questions remain (Chou et al. 2015; Chen et al. 2017; Reddy et al. 2017; Huang et al. 2020; Karin et al. 2020; Alda et al. 2021; Literman and Schwartz 2021; Mongiardino Koch 2021). In the absence of a consensus for phylogenomic best practices, whole-genome alignments offer an opportunity to investigate how partitioning and analyzing the data in different ways affect species-tree estimates. Based on previous work (Linnen and Farrell 2007; Bendall et al. 2022), f-branch statistics (Fig. 3), and the presence of multiple short internal branches in our species tree (Figs. 4,5), both introgression and high levels of incomplete lineage sorting complicate species-tree inference in *Neodiprion*. Nevertheless, our analysis of whole-genome alignments from closely related pine sawfly species revealed that, with a few exceptions, topologies were remarkably robust to marker sampling strategy and species-tree method (Fig. 2, S2).

We found that a concatenated ML analysis and an ASTRAL-III analysis of ML gene trees for 50-kb windows yielded identical topologies (Fig. 2a, 2b). This is perhaps surprising because concatenation is expected to perform particularly poorly when there are many conflicting histories in the dataset (Kubatko and Degnan 2007; Roch and Steel 2015; Mendes and Hahn 2018). Of course, with 50-kb windows, we are almost certainly violating the multi-species coalescent model assumption that gene trees were produced from non-recombining loci. Thus, our 50-kb loci are likely to contain many discordant histories (Mendes et al. 2019), an assumption that is supported by relatively high recombination rates (Fig. S4, S5) and many windows with low site concordance factors across the

genome (Fig. 4). However, we also found that our ASTRAL-III results were insensitive to window sizes ranging from 10 kb to 1 Mb, demonstrating that varying how badly we violated the "no intra-locus recombination" assumption had little effect on the inferred species tree.

We also found that subsampling coding sequences from the whole-genome dataset—a dataset analogous to one that might be produced from transcriptome sequencing or exome capture—had only a modest effect on the inferred species tree (Fig. 2c vs. Fig. 2b). Compared to the 50-kb window dataset, a dataset of 10,601 genes (coding sequences only) averaging 3,335 base pairs in length differed from the 50-kb window species tree in the placement of two species (*N. compar* and *N. dubiosus*). Although our protein-coding genes are contained within the genomic windows, most of the windowed data consists of intergenic or intronic sequence. Thus, our finding that chromosomal windows (largely non-coding) and genes (coding sequence only) produced slightly different topologies is consistent with previous work that indicates that data type (coding vs. non-coding sequence) can influence species-tree estimates (Chen et al. 2017; Reddy et al. 2017; Alda et al. 2021; Literman and Schwartz 2021). That said, topological conflicts between the 50-kb window dataset and the protein-coding dataset for this recently diverged clade were modest in comparison to more striking differences that have been documented in more distantly related taxa (e.g., hundreds of millions of years; Reddy et al. 2017; Literman and Schwartz 2021). Thus, the impact of data type on species-tree inference may be dependent on the divergence times of the taxa under study.

Turning to our SNP-based analyses, we found that the largest SNP datasets (all SNPs or 1 SNP sampled every kb) produced topologies that differed from the concatenation (Fig. 2a) and summary-tree analyses (Fig. 2b, 2c) only in the placement of *N. dubiosus* and the *pratti* complex. For *N. compar*, the SNP-based analysis recovered the same placement as the summary-tree analysis of coding loci. Although SVDquartets assumes that SNPs are independent—an assumption that is almost certainly violated in our all-SNPS and possibly our 1 kb-SNPs datasets—previous work demonstrates that this method performs

well even for datasets that include linked sites (Chifman and Kubatko 2014; Chou et al. 2015). Also, the similarity of our SNP tree (assumes free recombination among SNPs) and ASTRAL-III trees (assumes no recombination within loci) further suggests that violating model assumptions about recombination did not have much of an effect on species tree inference for this group of organisms.

We did find, however, that reducing the number of SNPs in our SVDquartets analyses affected our inferred species-tree topology (Fig. S2). For the more densely sampled datasets (1 per 1 kb, 5 kb, and 10 kb), topologies differed only in the same three relationships that were already unstable across datasets and methods (placements of *N. dubiosus*, *N. compar*, and the *pratti* group). But as the SNP datasets got sparser, we started to recover relationships that differed from all other analyses. For example, the 1 SNP per 50kb dataset recovered a different placement for *N. knereri* in the *virginianus* complex. Additional novel (and likely erroneous) relationships were recovered for the 1 SNP per 100kb dataset. Our observation that otherwise robust relationships became increasingly unstable with reduced SNP number is consistent with simulation work demonstrating that under some parameter combinations, large numbers of SNPs may be needed for accurate species-tree inference using this method (Long and Kubatko 2018; Wascher and Kubatko 2021). Notably, our smaller SNP datasets (~2,000-5,000 SNPs) are on par with those that might be generated via RADseq coupled with choosing a single site per RAD locus. Based on our results and previous studies of the behavior of SVDquartets, phylogenomic studies using RADseq should choose library preparation protocols and filtering strategies that maximize the number of loci and SNPs available for analysis.

While we have explored two types of species-tree methods (locus-based and SNP-based) and several different ways of partitioning whole-genome alignments, we have not exhaustively compared all species-tree methods or locus-sampling strategies. For example, data could be further subsampled to mimic other types of phylogenomic datasets that use different types of markers (e.g., UCEs). Although our whole-genome alignment would be prohibitive for full-likelihood methods, subsampled data could

be analyzed using full-likelihood methods that account for uncertainty in gene-tree estimation, as well as sources of discordance other than incomplete lineage sorting. A key question, however, is whether the size of datasets that could be analyzed in a reasonable time frame would be sufficient for the complex models under consideration. Consideration of divergence models that include hybrid species formation—which has been hypothesized for at least one *Neodiprion* species (Ross 1961)—may also be informative (e.g., Blischak et al. 2018). Finally, we have analyzed only a single exemplar per species, and previous work on the *Lecontei* clade (albeit with far fewer loci) suggests that taxon sampling can have a large effect on the inferred species tree (Linnen and Farrell 2008a). Also, when multiple individuals are sampled per species, a suite of complementary methods can be used to estimate demographic parameters, introgression rates, and population structure that might impact species-tree inference. For these reasons, examining the impact of taxon sampling is also a high priority for future work on the *Lecontei* species group.

An Updated Lecontei *Group Species Tree*

Aside from three lineages with uncertain placement (*N. compar*, *N. dubiosus,* and the *pratti* species complex), remaining relationships among *Lecontei* group species were consistently recovered across methods and datasets (Fig. 2). Based on these findings, we propose an updated species tree for the *Lecontei* group, with the three uncertain placements represented as polytomies (Fig. 6). Our updated species tree resembles the previous three-locus species tree (Fig. 1) but recovered an additional Ross (1955) species complex (*pratti* group) and resolved most relationships within species complexes. Both old and new species trees strongly reject the placement of *N. compar* within the *abbotii* complex, a relationship that was proposed by Ross (1955) based on similar larval coloration and behavior. Based on our results, we propose an updated *N. abbotii* complex that excludes *N. compar*.

Despite some uncertainty in the placement of *N. dubiosus*, our analyses consistently placed this species within the *abbotii* species group, rendering the *virginianus* complex polyphyletic and the *abbotii*

complex paraphyletic (Fig. 6). Curiously, the previous three-gene species tree—which contained multiple *N. dubiosus* individuals—consistently placed *N. dubiosus* within the *virginianus* species complex and recovered the *abbotii* complex (minus *N. compar*) as monophyletic (Fig. 1; Linnen and Farrell 2007, 2008a). This discrepancy could simply mean that the whole-genome data enabled us to recover a more accurate species-tree estimate. However, *N. dubiosus* adults are nearly indistinguishable from *N. rugifrons* adults (Becker et al. 1966), raising the possibility that placement within the *abbotii* complex in the whole-genome tree is incorrect. For example, it is possible that the *N. dubiosus* exemplar we chose, which had typical *N. dubiosus* larval morphology, had an admixed genome due to recent hybridization. We therefore refrain from reassigning *N. dubiosus* to the *abbotii* complex until additional samples can be analyzed.

Our updated species-tree also provides the necessary historical framework for investigating the evolution of many different types of life history and morphological traits, such as overwintering strategy, dietary specialization, and larval coloration (Fig. 6). However, how this tree should be used depends on whether the goal is to assess phenotypic convergence or genetic convergence (Manceau et al. 2010; Rosenblum et al. 2014). If the goal of an analysis is to assess phenotypic convergence—i.e., whether a trait spread to high frequency or fixation independently in two or more lineages following their divergence from a common ancestor, regardless of the genetic underpinnings of those trait changes—then the extensive underlying gene-tree discordance (Fig. 3) should not impact overall conclusions so long as any resulting uncertainty in the species-tree estimate is taken into account. This can be done, for example, by considering how alternative resolutions of the polytomies in Fig. 6 affect conclusions. But if the goal is to evaluate genetic convergence—i.e., whether phenotypic convergence is due to independent genetic mutations rather than the same genetic mutation shared via ILS or introgression (hemiplasy)—then gene-tree heterogeneity must be considered (Hahn and Nakhleh 2016; Guerrero and Hahn 2018). One approach for evaluating the probability of genetic convergence vs. hemiplasy in the

presence of ILS and introgression makes use of species trees with branch lengths scaled to coalescent units (e.g., Fig. 4, 5) (Hibbins et al. 2020). The ultimate test of genetic convergence, however, is to identify the genes and mutations responsible for convergent phenotypes (e.g., Wessinger and Rausher 2015).

Multiple Genomic Variables Predict Site Concordance Factors

Although most clades recovered in our species-tree analyses were robust to sampling and analysis method, there was nevertheless substantial variation in gene-tree topologies (Fig. 3) and site concordance factors (Fig. 4) across the genome, a pattern observed in many other taxa that diverged rapidly, often with gene flow (Pease et al. 2016; Edelman et al. 2019; Alda et al. 2021; Kozak et al. 2021; Meleshko et al. 2021b; Zhang et al. 2021). For the most part, variation in site concordance was highly heterogeneous across clades and chromosomes. However, for some clades, chromosome painting revealed unusually homogenous stretches of site concordance factors on Chromosomes 1 and 7 (Fig. 4). Such patterns could be generated by inversions segregating within and between species that yield more homogeneous phylogenetic histories due to restricted recombination within those regions. In support of this hypothesis, synteny plots for four *Neodiprion* species revealed independent evidence of the putative Chromosome 1 inversion (Fig. S3). Further evaluation of putative inversions and their contribution to observed variation in site concordance factors will require long-read data from additional populations and species.

Our high-quality chromosome-level assembly also provided us with an opportunity to investigate mechanisms that produce gene-tree heterogeneity across the genome (Table 1). Because several genomic predictors of concordance are correlated (Fig. S6), we used a stepwise regression approach to disentangle the contributions of individual variables to local site concordance factors while controlling

for other variables. We found that all seven genomic variables predicted windowed site concordance factors, although some of their effects varied depending on the branch under consideration (Fig. 5, Table 2, Table S7).

We found that both the amount and type of variation—as well as the amount of missing data—contributed to variation in site concordance factors across the genome. For example, we found that site concordance factors tended to increase with the number of parsimony-informative sites (10/14 branch-specific regression models that retained parsimony informative sites revealed a positive relationship; Fig. 5, Table 2). Although we had mixed predictions about the number of parsimony informative sites (Table 1), our regression models suggest that loci with more informative sites tend to yield higher levels of site concordance. In contrast to regression results for parsimony informative sites—and despite positive genome-wide correlations between parsimony informative sites, singletons, and missing data (Fig. S6)—we found that site concordance factors tended to *decrease* as the number of singletons and the amount of missing data increased (Fig. 5, Table 2). We consider two non-mutually exclusive explanations for these patterns (Table 1). First, decreased site concordance, increased singleton numbers, and increased missing data in some windows may simply reflect increased genotyping error, which could be caused by genome-wide heterogeneity in alignment error, sequencing error, and read counts. Second, windows with increased levels of variation—i.e., more singletons and parsimony informative sites and more indels that would be coded as missing data—may also be explained by reduced positive and negative selection relative to the rest of the genome. Because selection reduces local *Ne*, such relaxed-selection windows would be expected to have a higher density of discordant sites due to ILS.

We also found that local base composition (GC content) predicted site concordance in 50-kb windows. Out of our 17 branch-specific stepwise regression analyses, 14 retained GC-content in the final model, with 10 of these exhibiting the predicted negative relationship between GC content and site

concordance factors (Table 2). These findings are consistent with other phylogenomic datasets that have reported high levels of phylogenetic discordance in GC-rich regions of the genome (Jarvis et al. 2014; Bossert et al. 2017; Reddy et al. 2017; Romiguier and Roux 2017). One potential explanation for these findings is that repeated bouts of GC-biased gene conversion and compensatory mutations produce high levels of homoplasy (Jarvis et al. 2014; Romiguier and Roux 2017). GC-biased gene conversion is expected to be most pronounced in high-recombination regions of the genome (Eyre-Walker 1993; Galtier et al. 2001; Duret and Galtier 2009; Pessia et al. 2012; Figuet et al. 2015). Consistent with patterns in numerous taxa (e.g., Fullerton et al. 2001; Backström et al. 2010; Stevison and Noor 2010; Roesti et al. 2013), we found a positive correlation between GC content and recombination rate (Fig. S6).

Our regression analyses also support the hypothesis that linked selection, through its effects on ILS, contribute to genome-wide heterogeneity in site concordance factors. Because linked selection reduces local N_e (Maynard Smith and Haigh 1974; Charlesworth et al. 1993) and these effects are most pronounced in low-recombination regions of the genome (Kaplan et al. 1989; Charlesworth 2012; Bryant and Hahn 2020), ILS should correlate positively—and site concordance factors negatively—with recombination rate (Pease and Hahn 2013). Additionally, because the density of selected sites will be higher in gene-dense regions of the genome (Payseur and Nachman 2002; Pease and Hahn 2013), gene density should correlate negatively with ILS and positively with site concordance factors (Pease and Hahn 2013). However, correlations between gene density and recombination rate can potentially obscure these patterns (Flowers et al. 2012; Cutter and Payseur 2013).

Consistent with patterns in several other taxa (Freudenberg et al. 2009; Gore et al. 2009; Flowers et al. 2012; but see Wright et al. 2003; Cutter and Payseur 2013), we find that gene density and recombination rates—both inferred from *N. lecontei*—are positively correlated (Fig. S6). Nevertheless, the correlation is relatively weak, and our multiple regression models were able to tease apart their

independent effects (Fig. 5; Table S7). Out of our 17 branch-specific stepwise regression analyses, 14 retained gene density in the final model, with 12 of these exhibiting the predicted positive relationship between gene density and site concordance factors (Table 2). For recombination rate, 8 of 12 branch-specific models that included recombination rate had the predicted negative relationship with site concordance factors (Table 2). When both terms were retained in the final model, gene density tended to have an effect size that was equal or larger in magnitude compared to that of recombination rate, implying that the local density of selected sites might be a better predictor of concordance than local recombination rates. Alternatively, these differences may reflect more accurate estimates of gene density than of local recombination rate (see below). Either way, our regression results are consistent with work in other taxa demonstrating a negative relationship between the intensity of linked selection and discordance via incomplete lineage sorting (Hobolth et al. 2011; Prüfer et al. 2012; Pease and Hahn 2013).

Like ILS, levels of introgression are likely to vary across the genome. For example, selection against locally maladaptive alleles or genetic incompatibility alleles will reduce introgression in some parts of the genome (Nachman and Payseur 2012; Brandvain et al. 2014; Schumer et al. 2018; Li et al. 2019; Martin et al. 2019). To evaluate the contribution of variable introgression across the genome to variation in site concordance factors, we calculated D-statistics for each window for each branch in the species tree. Although these D-statistics were calculated using the same sites that were used to calculate site concordance factors, they are not inherently correlated since D-statistics reflect the magnitude of imbalance between the proportion of sites that support the two discordant topologies. In support of this, observed correlation coefficients between windowed site concordance factors and D-statistics for the 17 branches in the species tree were highly variable, ranging between r = -0.029 and r = -0.684. We found that windowed D-statistics were retained in all 17 regression models, with all 17

models exhibiting the predicted negative relationship between introgression (D-statistic) and site concordance factors (Fig. 5; Table 2).

Overall, our regression results provide strong evidence that ILS, introgression, genotyping error, and base composition all contribute to gene-tree heterogeneity and variable concordance across the genome. Although previous work and our own results point to biological causes such as ILS and introgression as major sources of gene-tree heterogeneity in recently diverged taxa (Bryant and Hahn 2020), our findings suggest that technological error—especially due to variable genotyping error across the genome—should not be discounted. Also, effect sizes in our regression models suggest that the relative importance of these different sources of heterogeneous concordance differ across branches in the species tree (Table S7).

While our results were mostly consistent with the predictions outlined in Table 1, we did observe some deviations from expected patterns in some of our branch-specific models (Fig. 5, Table S7, Table 2). Two possible explanations for deviations from predicted relationships for some model terms for some branches are: (1) some branches in the reference species tree used to calculate site concordance factors are incorrect, and (2) *N. lecontei*-based genomic variable estimates are not applicable to all species in this group. Regarding the second explanation, our estimates for genomic predictor variables came from four main sources: alignments for our 19 focal species (parsimony-informative sites, singletons, missing data, and GC content), site patterns for each branch (D-statistics), annotated genes in the *N. lecontei* reference genome (gene density), and recombination frequencies from a mapping cross between diverged *N. lecontei* populations (recombination rate). Of these sources, we expect our multi-species alignments to provide a reasonably accurate gauge of potential genotyping error, overall levels of variation, and base composition for each window. Site patterns were computed for each branch separately. Also, based on synteny plots from available species (Fig. S3), we expect gene density estimates from *N. lecontei* to provide a good approximation of gene densities in the other species. We

do not know, however, whether recombination rates estimated from a *N. lecontei* cross accurately predict recombination rates in other *Neodiprion* species. Although recombination rate is conserved among some closely related species, it is rapidly evolving in others (Smukowski and Noor 2011). It would therefore be worthwhile to estimate recombination rates in additional populations and species to further evaluate the robustness of our conclusions.

Conclusions

Over 15 years ago, we documented pervasive mitochondrial introgression in *Neodiprion* involving many members of the eastern *Lecontei* clade, rendering mitochondrial data unreliable for species-tree inference (Linnen and Farrell 2007). A "multi-locus" dataset of three nuclear genes was, unsurprisingly, insufficient for generating a robust species-tree estimate (Linnen and Farrell 2008a). Here, a whole-genome alignment analyzed with contemporary methods has produced a well-resolved species-tree that—except for three uncertain relationships—is robust to locus-sampling and analysis strategy. Using multiple regression, we found that genotyping error, ILS, and introgression all contribute to heterogeneous phylogenetic signal across the genome. This approach also revealed multiple genomic summary statistics that could be useful for identifying loci that are especially likely to recapitulate or depart from the underlying species tree. Overall, our results demonstrate how combining phylogenomic analysis with high-quality reference genomes, complementary genome feature data, and multiple analysis strategies can not only improve species-tree inference in difficult groups, but also reveal the sources of gene-tree heterogeneity that complicate phylogenetic inference. Similar analyses in additional taxa are needed to determine whether correlates of concordance vary predictably among taxa with different divergence histories. Essential ingredients for these analyses are high-quality, annotated reference genomes, which are increasingly available for non-model organisms (Hotaling et al. 2021).

Acknowledgements

We thank members of the Linnen and Weisrock labs for helpful discussion and Matt Hahn for constructive comments and helpful advice that improved our analyses and interpretations. We also thank an anonymous reviewer for helpful comments.

Funding

This work was supported by the University of Kentucky Center for Computational Sciences and the Lipscomb High Performance Computing Cluster, the SCINet project and the AI Center of Excellence of the United States Department of Agriculture (USDA) Agricultural Research Service (ARS), ARS project numbers 0201-88888-003-000D and 0201-88888-002-000D, the USDA National Institute of Food and Agriculture (2016-67014-2475; CRL), the National Science Foundation (DEB-CAREER-1750946; CRL and DEB-1355000; DWW), and a University of Kentucky Lyman T. Johnson Fellowship (KV). The USDA ARS is an equal opportunity/affirmative action employer and all agency services are available without discrimination.

Data Availability Statement

The *Neodiprion lecontei* reference genome assembly and annotation are available on NCBI (NCBI RefSeq Assembly GCF_021901455.1 and NCBI *Neodiprion lecontei* Annotation Release 101). All *Neodiprion* sequencing reads are available on NCBI (NCBI BioProject NCBI BioProject PRJNA854171, accession numbers SRR19909288-SRR19909306). *Neodiprion* pseudo-reference genomes (genomes for each species in *Neodiprion lecontei* genome coordinates), locus alignments (50-kb windows and genes), SNP datasets, and input files for regression analyses are available on Dryad

(https://doi.org/10.5061/dryad.bg79cnpf7). Custom scripts for generating *Neodiprion* pseudo-reference genomes, converting reference and pseudo-reference genomes into datasets for analysis (windows, genes, and SNPs), and scripts for downstream analyses are available on Zenodo (https://doi.org/10.5281/zenodo.11154207) and the LinnenLab GitHub page (https://github.com/LinnenLab) under the Herrig_etal_NeodiprionPhylogeny repository.

References

Adams R.H., Schield D.R., Card D.C., Castoe T.A. 2018. Assessing the impacts of positive selection on coalescent-based species tree estimation and species delimitation. Syst Biol. 67:1076–1090.

Alda F., Ludt W.B., Elías D.J., McMahan C.D., Chakrabarty P. 2021. Comparing ultraconserved elements and exons for phylogenomic analyses of middle American cichlids: When data agree to disagree. Genome Biol Evol. 13:evab161.

Avise J.C., Arnold J., Ball R.M., Bermingham E., Lamb T., Neigel J.E., Reeb C.A., Saunders N.C. 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu Rev Ecol Syst. 18:489–522.

Backström N., Forstmeier W., Schielzeth H., Mellenius H., Nam K., Bolund E., Webster M.T., Öst T., Schneider M., Kempenaers B., Ellegren H. 2010. The recombination landscape of the zebra finch *Taeniopygia guttata* genome. Genome Res. 20:485–495.

Bagley R.K., Sousa V.C., Niemiller M.L., Linnen C.R. 2017. History, geography and host use shape genomewide patterns of genetic variation in the redheaded pine sawfly (*Neodiprion lecontei*). Mol Ecol. 26:1022–1044.

Barton N.H. 2010. Genetic linkage and natural selection. Philos Trans R Soc Lond B Biol Sci. 365:2559–69.

Becker G.C., Wilkinson R.C., Benjamin D.M. 1966. Taxonomy of *Neodiprion rugifrons* and *N. dubiosus* (Hymenoptera: Tenthredinoidea: Diprionidae). Ann Entomol Soc Am. 59:173–178.

Bendall E.E., Bagley R.K., Sousa V.C., Linnen C.R. 2022. Faster-haplodiploid evolution under divergence-with-gene-flow: Simulations and empirical data from pine-feeding hymenopterans. Mol Ecol. 31:2348–2366.

Bendall E.E., Mattingly K.M., Moehring A.J., Linnen C.R. 2023. A test of Haldane's rule in *Neodiprion* sawflies and implications for the evolution of postzygotic isolation in haplodiploids. Am Nat. 202:40–54.

Bendall E.E., Vertacnik K.L., Linnen C.R. 2017. Oviposition traits generate extrinsic postzygotic isolation between two pine sawfly species. BMC Evol Biol. 17:26.

Betancur-R. R., Li C., Munroe T.A., Ballesteros J.A., Ortí G. 2013. Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). Syst Biol. 62:763–785.

Blischak P.D., Chifman J., Wolfe A.D., Kubatko L.S. 2018. HyDe: A Python package for genome-scale hybridization detection. Syst Biol. 67:821–829.

Borges R., Boussau B., Szöllősi G.J., Kosiol C. 2022. Nucleotide usage biases distort inferences of the species tree. Genome Biol Evol. 14.

Bossert S., Murray E.A., Blaimer B.B., Danforth B.N. 2017. The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data. Mol Phylogenet Evol. 111:149–157.

Bouckaert R.R. 2010. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics. 26:1372–3.

Brandvain Y., Kenney A.M., Flagel L., Coop G., Sweigart A.L. 2014. Speciation and Introgression between *Mimulus nasutus* and *Mimulus guttatus*. PLoS Genet. 10:e1004410.

Broman K.W., Wu H., Sen S., Churchill G.A. 2003. R/qtl: QTL mapping in experimental crosses. Bioinformatics. 19:889–890.

Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., RoyChoudhury A. 2012. Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. Mol Biol Evol. 29:1917–1932.

Bryant D., Hahn M.W. 2020. The Concatenation Question. In: Scornavacca C., Delsuc F., Galtier N., editors. Phylogenetics in the Genomic Era. No commercial publisher. p. 3.4:1-3.4:23.

Buchfink B., Reuter K., Drost H.-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 18:366–368.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421.

Charlesworth B. 2012. The Effects of Deleterious Mutations on Evolution at Linked Sites. Genetics. 190:5–22.

Charlesworth B., Betancourt A.J., Kaiser V.B., Gordo I. 2009. Genetic recombination and molecular evolution. Cold Spring Harb Symp Quant Biol. 74:177–86.

Charlesworth B., Morgan M.T., Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics. 134:1289–1303.

Chen M.-Y., Liang D., Zhang P. 2017. Phylogenomic resolution of the phylogeny of Laurasiatherian mammals: exploring phylogenetic signals within coding and noncoding sequences. Genome Biol Evol. 9:1998–2012.

Cheng H., Concepcion G.T., Feng X., Zhang H., Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 18:170–175.

Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics. 30:3317–3324.

Chou J., Gupta A., Yaduvanshi S., Davidson R., Nute M., Mirarab S., Warnow T. 2015. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. BMC Genomics. 16:S2.

Coppel H.C., Benjamin D.M. 1965. Bionomics of Nearctic pine-feeding diprionids. Annu Rev Entomol. 10:69–96.

Cutter A.D., Payseur B.A. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat Rev Genet. 14:262–274.

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R. 2011. The variant call format and VCFtools. Bioinformatics. 27:2156–2158.

DeAngelis M.M., Wang D.G., Hawkins T.L. 1995. Solid-phase reversible immobilization for the isolation of PCR products. Nucleic Acids Res. 23:4742–4743.

Deorowicz S., Debudaj-Grabysz A., Grabowski S. 2013. Disk-based k-mer counting on a PC. BMC Bioinformatics. 14:160.

Durand E.Y., Patterson N., Reich D., Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol Biol Evol. 28:2239–2252.

Durand N.C., Robinson J.T., Shamim M.S., Machol I., Mesirov J.P., Lander E.S., Aiden E.L. 2016. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 3:99–101.

Duret L., Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 10:285–311.

Dutheil J.Y., Munch K., Nam K., Mailund T., Schierup M.H. 2015. Strong selective sweeps on the X chromosome in the human-chimpanzee ancestor explain its low divergence. PLoS Genet. 11:e1005451.

Edelman N.B., Frandsen P.B., Miyagi M., Clavijo B., Davey J., Dikow R.B., García-Accinelli G., Van Belleghem S.M., Patterson N., Neafsey D.E., Challis R., Kumar S., Moreira G.R.P., Salazar C., Chouteau M., Counterman B.A., Papa R., Blaxter M., Reed R.D., Dasmahapatra K.K., Kronforst M., Joron M., Jiggins C.D., McMillan W.O., Di Palma F., Blumberg A.J., Wakeley J., Jaffe D., Mallet J. 2019. Genomic architecture and introgression shape a butterfly radiation. Science (1979). 366:594–599.

Edwards S. v. 2009. Is a new and general theory of molecular systematics emerging? Evolution (N Y). 63:1–19.

Edwards S. v, Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. PNAS. 104:5936–5941.

Edwards S. v., Xi Z., Janke A., Faircloth B.C., McCormack J.E., Glenn T.C., Zhong B., Wu S., Lemmon E.M., Lemmon A.R., Leaché A.D., Liu L., Davis C.C. 2016. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. Mol Phylogenet Evol. 94:447–462.

Eyre-Walker A. 1993. Recombination and mammalian genome evolution. Proc Biol Sci. 252:237–43.

Faust G.G., Hall I.M. 2014. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 30:2503–5.

Figuet E., Ballenghien M., Romiguier J., Galtier N. 2015. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. Genome Biol Evol. 7:240–250.

Flouri T., Jiao X., Rannala B., Yang Z. 2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. Mol Biol Evol. 37:1211–1223.

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. Mol Biol Evol. 29:675–687.

Fontaine M.C., Pease J.B., Steele A., Waterhouse R.M., Neafsey D.E., Sharakhov I. v., Jiang X., Hall A.B., Catteruccia F., Kakani E., Mitchell S.N., Wu Y.-C., Smith H.A., Love R.R., Lawniczak M.K., Slotman M.A., Emrich S.J., Hahn M.W., Besansky N.J. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science (1979). 347.

Freudenberg J., Wang M., Yang Y., Li W. 2009. Partial correlation analysis indicates causal relationships between GC-content, exon density and recombination rate in the human genome. BMC Bioinformatics. 10:S66.

Fullerton S.M., Bernardo Carvalho A., Clark A.G. 2001. Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol. 18:1139–1142.

Galtier N., Piganeau G., Mouchiroud D., Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics. 159:907–11.

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 (N Y). 77:437–453.

Gore M.A., Chia J.-M., Elshire R.J., Sun Q., Ersoz E.S., Hurwitz B.L., Peiffer J.A., McMullen M.D., Grills G.S., Ross-Ibarra J., Ware D.H., Buckler E.S. 2009. A First-Generation Haplotype Map of Maize. Science (1979). 326:1115–1117.

Grafen, A. 1989. The phylogenetic regression. Philos Trans R Soc Lond B Biol Sci. 326:119-157.

Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai W., Fritz M.H.-Y., Hansen N.F., Durand E.Y., Malaspinas A.-S., Jensen J.D., Marques-Bonet T., Alkan C., Prüfer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-Petri A., Butthof A., Höber B., Höffner B., Siegemund M., Weihmann A., Nusbaum C., Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic D., Kucan Ž., Gušic I., Doronichev V.B., Golovanova L. V., Lalueza-Fox C., de la Rasilla M., Fortea J., Rosas A., Schmitz R.W., Johnson P.L.F., Eichler E.E., Falush D., Birney E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D., Pääbo S. 2010. A draft sequence of the Neandertal genome. Science (1979). 328:710–722.

Guerrero R.F., Hahn M.W. 2018. Quantifying the risk of hemiplasy in phylogenetic inference. Proceedings of the National Academy of Sciences. 115:12787–12792.

Hahn M.W., Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evolution (N Y). 70:7–17.

Harrison R.G., Larson E.L. 2014. Hybridization, introgression, and the nature of species boundaries. Journal of Heredity. 105:795–809.

Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol Biol Evol. 27:570–580.

Herrig D.K., Vertacnik K.L., Kohrs A.R., Linnen C.R. 2021. Support for the adaptive decoupling hypothesis from whole-transcriptome profiles of a hypermetamorphic and sexually dimorphic insect, *Neodiprion lecontei*. Mol Ecol. 30:4551–4566.

Hey J., Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics. 167:747–760.

Hibbins M.S., Gibson M.J., Hahn M.W. 2020. Determining the probability of hemiplasy in the presence of incomplete lineage sorting and introgression. Elife. 9.

Hibbins M.S., Hahn M.W. 2022. Phylogenomic approaches to detecting and characterizing introgression. Genetics. 220:iyab173.

Hobolth A., Dutheil J.Y., Hawks J., Schierup M.H., Mailund T. 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Res. 21:349–56.

Hotaling S., Sproul J.S., Heckenhauer J., Powell A., Larracuente A.M., Pauls S.U., Kelley J.L., Frandsen P.B. 2021. Long reads are revolutionizing 20 years of insect genome sequencing. Genome Biol Evol. 13.

Huang J., Flouri T., Yang Z. 2020. A simulation study to examine the information content in phylogenomic data sets under the multispecies coalescent model. Mol Biol Evol. 37:3211–3224.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C. v, Lovell P. v, Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Vargas Velazquez A.M., Alfaro-Núñez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jønsson K.A., Johnson W., Koepfli K.-P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alström P., Edwards S. v, Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 346:1320–31.

Jiang X., Edwards S. v, Liu L. 2020. The multispecies coalescent model outperforms concatenation across diverse phylogenomic data sets. Syst Biol. 69:795–812.

Jones G.R. 2018. Divergence estimation in the presence of incomplete lineage sorting and migration. Syst Biol. 68:19–31.

Kaplan N.L., Hudson R.R., Langley C.H. 1989. The ``hitchhiking effect'' revisited. Genetics. 123:887–899.

Kapustin Y., Souvorov A., Tatusova T., Lipman D. 2008. Splign: algorithms for computing spliced alignments with identification of paralogs. Biol Direct. 3:20.

Karin B.R., Gamble T., Jackman T.R. 2020. Optimizing phylogenomics with rapidly evolving long exons: comparison with anchored hybrid enrichment and ultraconserved elements. Mol Biol Evol. 37:904–922.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 16:111–20.

Knerer G. 1984. Diprionid Sawflies : Biological Topics and Rearing Techniques (Hymenoptera : Symphyta). Bulletin of the ESA. 30:53–57.

Knerer G., Atwood C.E. 1973. Diprionid sawflies: polymorphism and speciation. Science (1979). 179:1090–1099.

Kozak K.M., Joron M., McMillan W.O., Jiggins C.D. 2021. Rampant genome-wide admixture across the *Heliconius* radiation. Genome Biol Evol. 13:evab099.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol. 56:17–24.

Laetsch D.R., Blaxter M.L. 2017. BlobTools: Interrogation of genome assemblies. F1000Res. 6:1287.

Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9:357–9.

Lanier H.C., Knowles L.L. 2012. Is recombination a problem for species-tree analyses? Syst Biol. 61:691–701.

Leaché A.D., Harris R.B., Rannala B., Yang Z. 2014. The influence of gene flow on species tree estimation: a simulation study. Syst Biol. 63:17–30.

Li G., Figueiró H. V, Eizirik E., Murphy W.J. 2019. Recombination-aware phylogenomics reveals the structured genomic landscape of hybridizing cat species. Mol Biol Evol. 36:2111–2126.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 27:2987–93.

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

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34:3094–3100.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–9.

Linnen C.R. 2010. Species-tree estimation for complex divergence histories: a case study in *Neodiprion* sawflies. In: Knowles L.L., Kubatko L.S., editors. Estimating Species Trees: In Practice and Theory. Hoboken: Wiley-Blackwell. p. 145–162.

Linnen C.R., Farrell B.D. 2007. Mitonuclear discordance is caused by rampant mitochondrial introgression in *Neodiprion* (Hymenoptera: Diprionidae) sawflies. Evolution (N Y). 61:1417–1438.

Linnen C.R., Farrell B.D. 2008a. Comparison of methods for species-tree inference in the sawfly genus *Neodiprion* (Hymenoptera: Diprionidae). Syst Biol. 57:876–90.

Linnen C.R., Farrell B.D. 2008b. Phylogenetic analysis of nuclear and mitochondrial genes reveals evolutionary relationships and mitochondrial introgression in the *sertifer* species group of the genus *Neodiprion* (Hymenoptera: Diprionidae). Mol Phylogenet Evol. 48:240–257.

Linnen C.R., Farrell B.D. 2010. A test of the sympatric host race formation hypothesis in *Neodiprion* (Hymenoptera: Diprionidae). Proceedings of the Royal Society B: Biological Sciences. 277:3131–3138.

Linnen C.R., O'Quin C.T., Shackleford T., Sears C.R., Lindstedt C. 2018. Genetic basis of body color and spotting pattern in redheaded pine sawfly larvae (*Neodiprion lecontei*). Genetics. 209:291–305.

Literman R., Schwartz R. 2021. Genome-scale profiling reveals noncoding loci carry higher proportions of concordant data. Mol Biol Evol. 38:2306–2318.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics. 24:2542–2543.

Liu L., Pearl D.K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst Biol. 56:504–14.

Liu L., Xi Z., Wu S., Davis C.C., Edwards S. v. 2015. Estimating phylogenetic trees from genome-scale data. Ann N Y Acad Sci. 1360:36–53.

Liu L., Yu L., Pearl D.K. 2010. Maximum tree: a consistent estimator of the species tree. J Math Biol. 60:95–106.

Long C., Kubatko L. 2018. The effect of gene flow on coalescent-based species-tree inference. Syst Biol. 67:770–785.

Maddison W.P. 1997. Gene trees in species trees. Syst Biol. 46:523–536.

Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst Biol. 55:21–30.

Malinsky M., Matschiner M., Svardal H. 2021. Dsuite - Fast D-statistics and related admixture evidence from VCF files. Mol Ecol Resour. 21:584–595.

Malinsky M., Svardal H., Tyers A.M., Miska E.A., Genner M.J., Turner G.F., Durbin R. 2018. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nature Ecology & Evolution 2018 2:12. 2:1940–1955.

Mallet J., Besansky N., Hahn M.W. 2016. How reticulated are species? BioEssays. 38:140–149.

Manceau M., Domingues V.S., Linnen C.R., Rosenblum E.B., Hoekstra H.E. 2010. Convergence in pigmentation at multiple levels: mutations, genes and function. Philos Trans R Soc Lond B Biol Sci. 365:2439–50.

Manni M., Berkeley M.R., Seppey M., Zdobnov E.M. 2021. BUSCO: assessing genomic data quality and beyond. Curr Protoc. 1:1–41.

Mapleson D., Garcia Accinelli G., Kettleborough G., Wright J., Clavijo B.J. 2016. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics.:btw663.

Martin S.H., Davey J.W., Salazar C., Jiggins C.D. 2019. Recombination rate variation shapes barriers to introgression across butterfly genomes. PLoS Biol. 17:e2006288.

Maynard Smith J.M., Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res. 23:23–35.

McCormack J.E., Hird S.M., Zellmer A.J., Carstens B.C., Brumfield R.T. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol Phylogenet Evol. 66:526–538.

Meleshko O., Martin M.D., Korneliussen T.S., Schröck C., Lamkowski P., Schmutz J., Healey A., Piatkowski B.T., Shaw A.J., Weston D.J., Flatberg K.I., Szövényi P., Hassel K., Stenøien H.K. 2021a. Extensive genome-wide phylogenetic discordance is due to incomplete lineage sorting and not ongoing introgression in a rapidly radiated bryophyte genus. Mol Biol Evol. 38:2750–2766.

Meleshko O., Martin M.D., Korneliussen T.S., Schröck C., Lamkowski P., Schmutz J., Healey A., Piatkowski B.T., Shaw A.J., Weston D.J., Flatberg K.I., Szövényi P., Hassel K., Stenøien H.K. 2021b. Extensive genome-wide phylogenetic discordance Is due to incomplete lineage sorting and not ongoing introgression in a rapidly radiated bryophyte genus. Mol Biol Evol. 38:2750–2766.

Mendes F.K., Hahn M.W. 2018. Why Concatenation Fails Near the Anomaly Zone. Syst Biol. 67:158–169.

Mendes F.K., Livera A.P., Hahn M.W. 2019. The perils of intralocus recombination for inferences of molecular convergence. Philosophical Transactions of the Royal Society B: Biological Sciences. 374:20180244.

Minh B.Q., Hahn M.W., Lanfear R. 2020a. New Methods to Calculate Concordance Factors for Phylogenomic Datasets. Mol Biol Evol. 37:2727–2733.

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A., Lanfear R. 2020b. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 37:1530–1534.

Mo Y.K., Lanfear R., Hahn M.W., Minh B.Q. 2023. Updated site concordance factors minimize effects of homoplasy and taxon sampling. Bioinformatics. 39.

Mongiardino Koch N. 2021. Phylogenomic Subsampling and the Search for Phylogenetically Reliable Loci. Mol Biol Evol. 38:4025–4038.

Morgulis A., Gertz E.M., Schaffer A.A., Agarwala R. 2006. WindowMasker: window-based masker for sequenced genomes. Bioinformatics. 22:134–141.

Nachman M.W., Payseur B.A. 2012. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Philosophical Transactions of the Royal Society B: Biological Sciences. 367:409–421.

Ogilvie H.A., Bouckaert R.R., Drummond A.J. 2017. StarBEAST2 brings faster species Ttee inference and accurate estimates of substitution rates. Mol Biol Evol. 34:2101–2114.

Oróstica K.Y., Verdugo R.A. 2016. chromPlot: visualization of genomic data in chromosomal context. Bioinformatics. 32:2366–2368.

Page A.J., Taylor B., Delaney A.J., Soares J., Seemann T., Keane J.A., Harris S.R. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom. 2:e000056.

Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528.

Patterson N., Moorjani P., Luo Y., Mallick S., Rohland N., Zhan Y., Genschoreck T., Webster T., Reich D. 2012. Ancient admixture in human history. Genetics. 192:1065–1093.

Payseur B.A., Nachman M.W. 2002. Gene density and human nucleotide polymorphism. Mol Biol Evol. 19:336–40.

Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol. 14:e1002379.

Pease J.B., Hahn M.W. 2013. More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. Evolution (N Y). 67:2376–84.

Pessia E., Popa A., Mousset S., Rezvoy C., Duret L., Marais G.A.B. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. Genome Biol Evol. 4:675–682.

Peterson B.K., Weber J.N., Kay E.H., Fisher H.S., Hoekstra H.E. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One. 7:e37135.

Pollard D.A., Iyer V.N., Moses A.M., Eisen M.B. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genet. 2:e173.

Prüfer K., Munch K., Hellmann I., Akagi K., Miller J.R., Walenz B., Koren S., Sutton G., Kodira C., Winer R., Knight J.R., Mullikin J.C., Meader S.J., Ponting C.P., Lunter G., Higashino S., Hobolth A., Dutheil J., Karakoç E., Alkan C., Sajjadian S., Catacchio C.R., Ventura M., Marques-Bonet T., Eichler E.E., André C., Atencia R., Mugisha L., Junhold J., Patterson N., Siebauer M., Good J.M., Fischer A., Ptak S.E., Lachmann M., Symer D.E., Mailund T., Schierup M.H., Andrés A.M., Kelso J., Pääbo S. 2012. The bonobo genome compared with the chimpanzee and human genomes. Nature. 486:527–531.

Quinlan A.R. 2014. BEDTools: the Swiss-army tool for genome feature analysis. Curr Protoc Bioinformatics. 47.

R Core Team. 2024. R: A Language and Environment for Statistical Computing. .

Rambaut A. 2007. FigTree, a graphical viewer of phylogenetic trees. Available from http://tree.bio.ed.ac.uk/software/figtree.

Ranallo-Benavidez T.R., Jaron K.S., Schatz M.C. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 11:1432.

Rannala B., Edwards S. v., Leaché A., Yang Z. 2020. The multispecies coalescent model and species-tree inference. In: Scornavacca C., Delsuc F., Galtier N., editors. Phylogenetics in the genomic era. No commercial publisher. p. 3.3:1-3.3:21.

Rannala B., Yang Z. 2003. Bayes estimation of species divergence times and ancestral population Sizes using DNA sequences from multiple loci. Genetics. 164:1645–1656.

Reddy S., Kimball R.T., Pandey A., Hosner P.A., Braun M.J., Hackett S.J., Han K.-L., Harshman J., Huddleston C.J., Kingston S., Marks B.D., Miglia K.J., Moore W.S., Sheldon F.H., Witt C.C., Yuri T., Braun E.L. 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the Avian Tree of Life more than taxon sampling. Syst Biol. 66:857–879.

Rezvoy C., Charif D., Gueguen L., Marais G.A.B. 2007. MareyMap: an R-based tool with graphical interface for estimating recombination rates. Bioinformatics. 23:2188–2189.

Roch S., Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor Popul Biol. 100:56–62.

Roesti M., Moser D., Berner D. 2013. Recombination in the threespine stickleback genome-patterns and consequences. Mol Ecol. 22:3014–3027.

Romiguier J., Roux C. 2017. Analytical biases associated with GC-content in molecular evolution. Front Genet. 8:16.

Rosenblum E.B., Parent C.E., Brandt E.E. 2014. The Molecular Basis of Phenotypic Convergence. :203–226.

Ross H.H. 1955. The taxonomy and evolution of the sawfly genus *Neodiprion*. Forest Science. 1:196–209.

Ross H.H. 1961. Two new species of *Neodiprion* from southeastern North America (Hymenoptera: Diprionidae). Ann Entomol Soc Am. 54:451–453.

Sarver B.A.J., Keeble S., Cosart T., Tucker P.K., Dean M.D., Good J.M. 2017. Phylogenomic insights into mouse evolution using a pseudoreference approach. Genome Biol Evol. 9:726–739.

Schumer M., Xu C., Powell D.L., Durvasula A., Skov L., Holland C., Blazier J.C., Sankararaman S., Andolfatto P., Rosenthal G.G., Przeworski M. 2018. Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science (1979). 360:656–660.

Siberchicot A., Bessy A., Guéguen L., Marais G.A.B. 2017. MareyMap online: a user-friendly web application and database service for estimating recombination rates using physical and genetic maps. Genome Biol Evol. 9:2506–2509.

Sim S.B., Corpuz R.L., Simmonds T.J., Geib S.M. 2022. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. BMC Genomics. 23:157.

Slatkin M., Pollack J.L. 2006. The concordance of gene trees and species trees at two linked loci. Genetics. 172:1979–1984.

Small S.T., Labbé F., Lobo N.F., Koekemoer L.L., Sikaala C.H., Neafsey D.E., Hahn M.W., Fontaine M.C., Besansky N.J. 2020. Radiation with reticulation marks the origin of a major malaria vector. Proceedings of the National Academy of Sciences. 117:31583–31590.

Smukowski C.S., Noor M.A.F. 2011. Recombination rate variation in closely related species. Heredity (Edinb). 107:496–508.

Solís-Lemus C., Yang M., Ané C. 2016. Inconsistency of species tree methods under gene flow. Syst Biol. 65:843–851.

Souvorov A., Kapustin Y., Kiryutin B., Chetvernin V., Tatusova T., Lipman D. 2010. Gnomon-NCBI eukaryotic gene prediction tool. Natinoal Center for Biotechnology Information.:1–24.

Stevison L.S., Noor M.A.F. 2010. Genetic and evolutionary correlates of fine-scale recombination rate variation in *Drosophila persimilis*. J Mol Evol. 71:332–45.

Stukenbrock E.H., Bataillon T., Dutheil J.Y., Hansen T.T., Li R., Zala M., McDonald B.A., Wang J., Schierup M.H. 2011. The making of a new pathogen: Insights from comparative population genomics of the domesticated wheat pathogen *Mycosphaerella graminicola* and its wild sister species. Genome Res. 21:2157–2166.

Swofford D.L. 2000. PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods). Sunderland, MA: Sinauer.

Takahata N. 1989. Gene genealogy in three related populations: Consistency probability between gene and population trees. Genetics. 122:957–966.

Taylor J., Butler D. 2017. *R* Package **ASMap** : efficient genetic linkage map construction and diagnosis. J Stat Softw. 79.

Wascher M., Kubatko L. 2021. Consistency of SVDQuartets and maximum likelihood for coalescent-based species tree estimation. Syst Biol. 70:33–48.

Wen D., Nakhleh L. 2018. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. Syst Biol. 67:439–457.

Wen D., Yu Y., Zhu J., Nakhleh L. 2018. Inferring phylogenetic networks using PhyloNet. Syst Biol. 67:735–740.

Wessinger C.A., Rausher M.D. 2015. Ecological Transition Predictably Associated with Gene Degeneration. Mol Biol Evol. 32:347–354.

White M.A., Ané C., Dewey C.N., Larget B.R., Payseur B.A. 2009. Fine-scale phylogenetic discordance across the house mouse genome. PLoS Genet. 5:e1000729.

Wright S.I., Agrawal N., Bureau T.E. 2003. Effects of Recombination Rate and Gene Density on Transposable Element Distributions in *Arabidopsis thaliana*. Genome Res. 13:1897–1903.

Yan Z., Smith M.L., Du P., Hahn M.W., Nakhleh L. 2022. Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. Syst Biol. 71:367–381.

Yu Y., Dong J., Liu K.J., Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proceedings of the National Academy of Sciences. 111:16448–16453.

Zhang C., Ogilvie H.A., Drummond A.J., Stadler T. 2018a. Bayesian inference of species networks from multilocus sequence data. Mol Biol Evol. 35:504–517.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018b. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:153.

Zhang D., Rheindt F.E., She H., Cheng Y., Song G., Jia C., Qu Y., Alström P., Lei F. 2021. Most genomic loci misrepresent the phylogeny of an avian radiation because of ancient gene flow. Syst Biol. 70:961–975.

FIGURE CAPTIONS

Figure 1: Proposed *Neodiprion* species complexes and species tree estimated from three nuclear loci. Shading corresponds to proposed species complexes (Ross 1955) and tree topology corresponds to proposed topology in Linnen and Farrell 2008a.

Figure 2: *Neodiprion* species trees estimated via concatenation, ASTRAL-III, and SVDquartets. a) ML topology estimated via concatenating all 50-kb windows, **b)** ASTRAL-III species tree estimated from ML gene trees estimated from 50-kb windows, **c)** ASTRAL-III species tree estimated from ML gene trees estimated from coding loci only, **d)** SVDquartets species tree estimated from all SNPs (after quality filtering). Shading corresponds to the named species complexes in Fig. 1. Numbers above nodes are site concordance factors; numbers below nodes are gene concordance factors.

Figure 3: Genome-wide patterns of gene-tree discordance with f-branch statistics. The tree on the left was generated using DensiTree and depicts topologies for gene-trees estimated from 50-kb windows (gray/thin lines) and the corresponding ASTRAL-III species tree (blue/thick line). The matrix to the right shows f-branch statistics generated using the Dsuite package. Since only a single exemplar was included per species, f-branch statistics could not be obtained for sister taxa.

Figure 4: The genomic landscape of concordance. The tree on the left is the ASTRAL-III species tree estimated from 50-kb windows, with branch lengths scaled to coalescent units. Labeled clades correspond to painted chromosomes in the 17 panels, each depicting how site concordance factors

estimated in 50-kb windows change along each chromosome. Gray areas on each chromosome are windows that were removed prior to analysis due to excessive missing data (mostly in centromeric and telomeric regions). Stars highlight stretches of relatively uniform sites concordance factors on Chromosome 1 and 7 that could be explained by inversions (see Fig. S3).

Figure 5: Summary of multiple regression results for site concordance in 50-kb windows. The tree is the ASTRAL-III species tree from 50-kb windows used as the reference tree to calculate site concordance factors for each 50-kb window. Each node has the seven predictor variables included in the multiple regression models. Gray variables were not retained after stepwise model selection; blue and red variables were retained in the final model with positive and negative effects, respectively. An asterisk indicates that estimated effects from the regression model were in the opposite direction predicted in Table 1. R^2 is reported for the final models; additional model results are in Table S7.

Figure 6: An updated species tree for eastern North American *Neodiprion*. The topology reflects relationships robust to species-tree method and locus-sampling strategy (Fig. 2, Fig. S2), with three uncertain relationships depicted as polytomies. Evolutionary analyses of variable characters such as overwintering mode, host preferences (modified from Linnen and Farrell 2010), and larval color should use alternative resolutions of these branches.

Table 1. Predicted effects of genomic variables on site concordance factors (sCFs), a measure of genealogical concordance

Genomic predictor variable	Predicted effect on sCFs	Rationale
Parsimony-informative sites	**+ or -**	**+**: increased # of sites increases information content; **-**: increased # of sites may reflect relaxed selection and increased ILS
Singletons	**-**	increased # of sites may reflect relaxed selection (increased N_e and ILS) or increased genotyping error
Missing data	**-**	higher levels of missing data may reflect higher genotyping error
GC content	**-**	GC-biased gene conversion and compensatory mutations can lead to high levels of homoplasy in GC-rich regions of the genome
D-statistic	**-**	Genomic windows with evidence of introgression are more likely to produce topologies that depart from the species tree
Gene density	**+**	Gene-rich regions have a higher density of selected sites, decreasing N_e, and reducing ILS
Recombination rate	**-**	Linked selection is most pronounced in low-recombination regions, decreasing N_e, and reducing ILS

Table 2. Summary of regression coefficient signs for genomic predictor variables obtained from stepwise regression analysis of site concordance factors (sCFs) for 17 clades in the *Neodiprion* species tree.[1]

Genomic predictor variable	Predicted effect on sCFs[2]	# Clades **+** effect on sCFs	# Clades **-** effect on sCFs	# Clades no effect on sCFs[3]
Parsimony-informative sites	**+ or -**	10	4	3
Singletons	**-**	4	11	2
Missing data	**-**	2	14	1
GC content	**-**	4	10	3
D-statistic	**-**	0	17	0
Gene density	**+**	12	2	3
Recombination rate	**-**	4	8	5

[1] See Table S7 for full regression results.
[2] See Table 1 for the rationale underlying each prediction.
[3] "No effect" indicates that predictor variable was not retained after stepwise model selection.

Figure 1

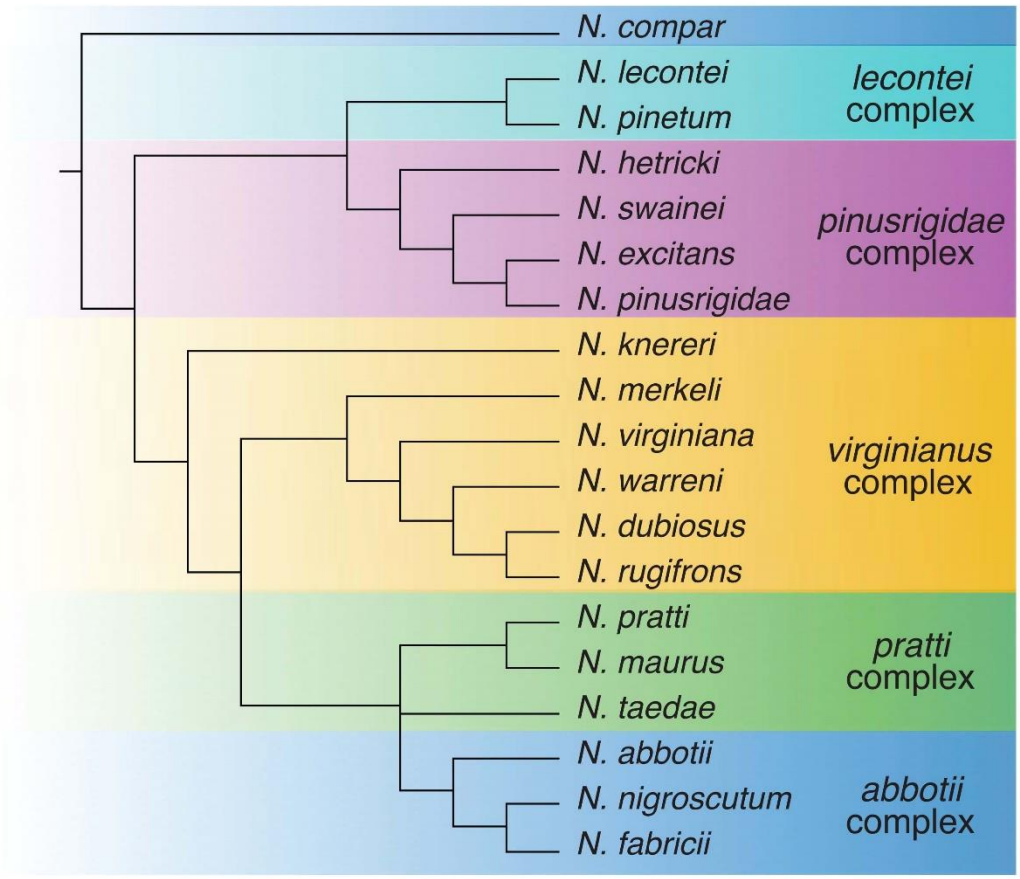

Figure 2

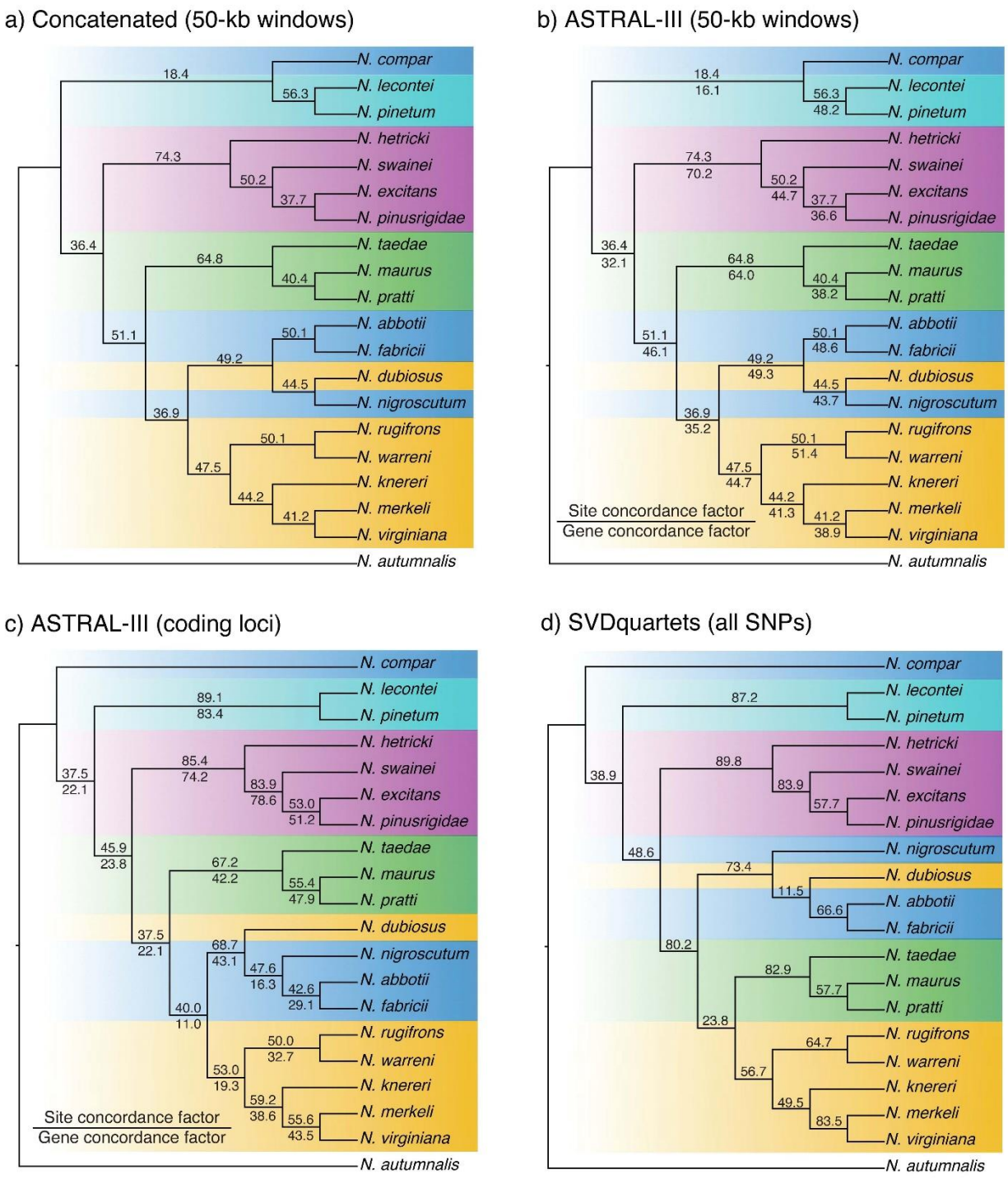

Figure 3

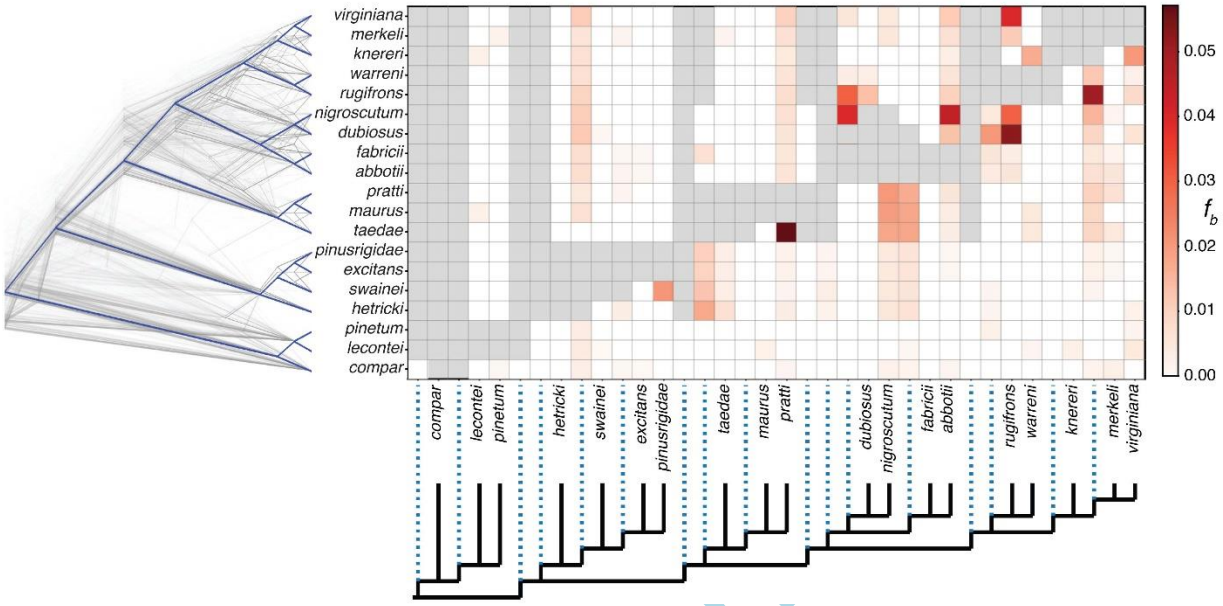

Figure 4

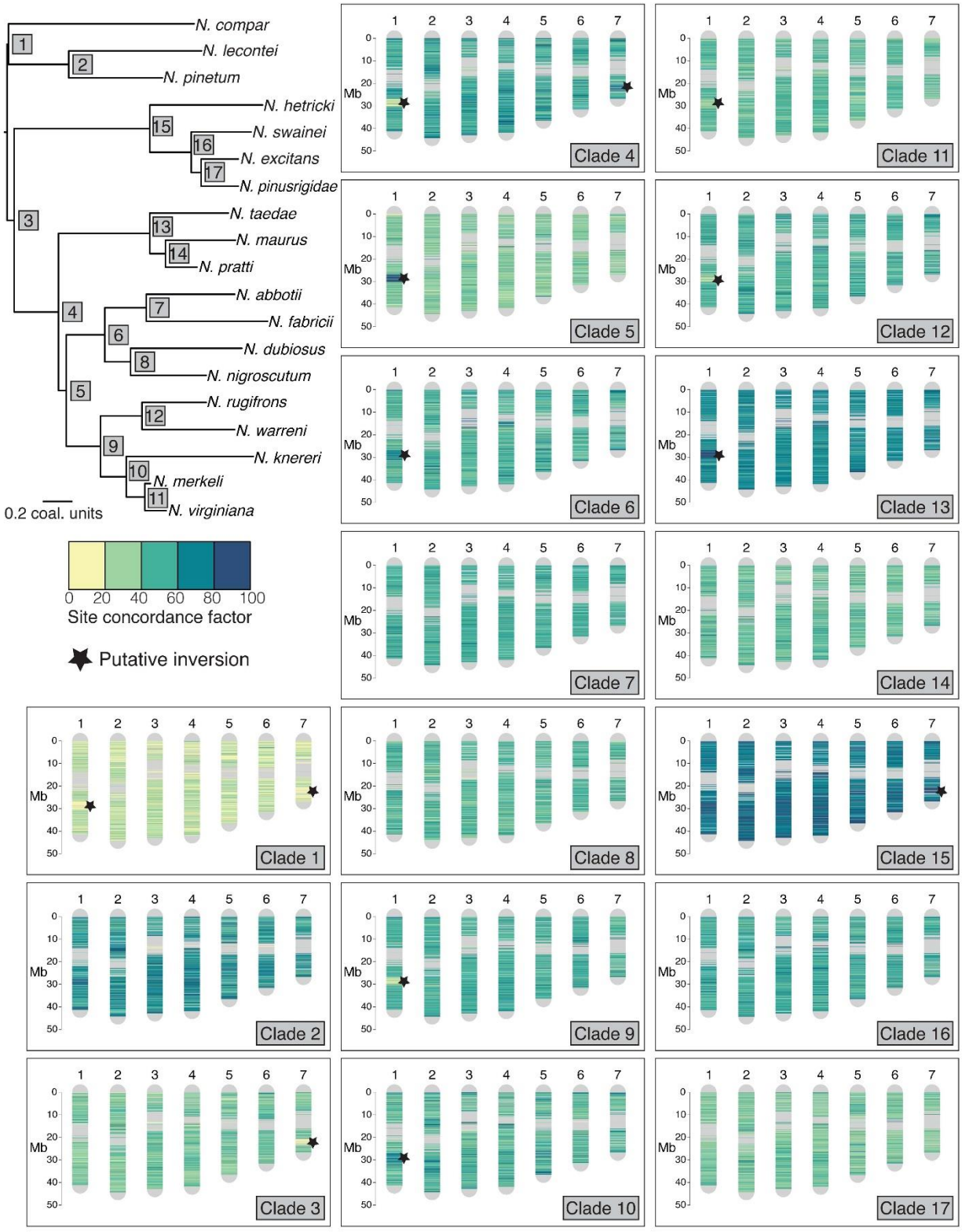

Figure 5

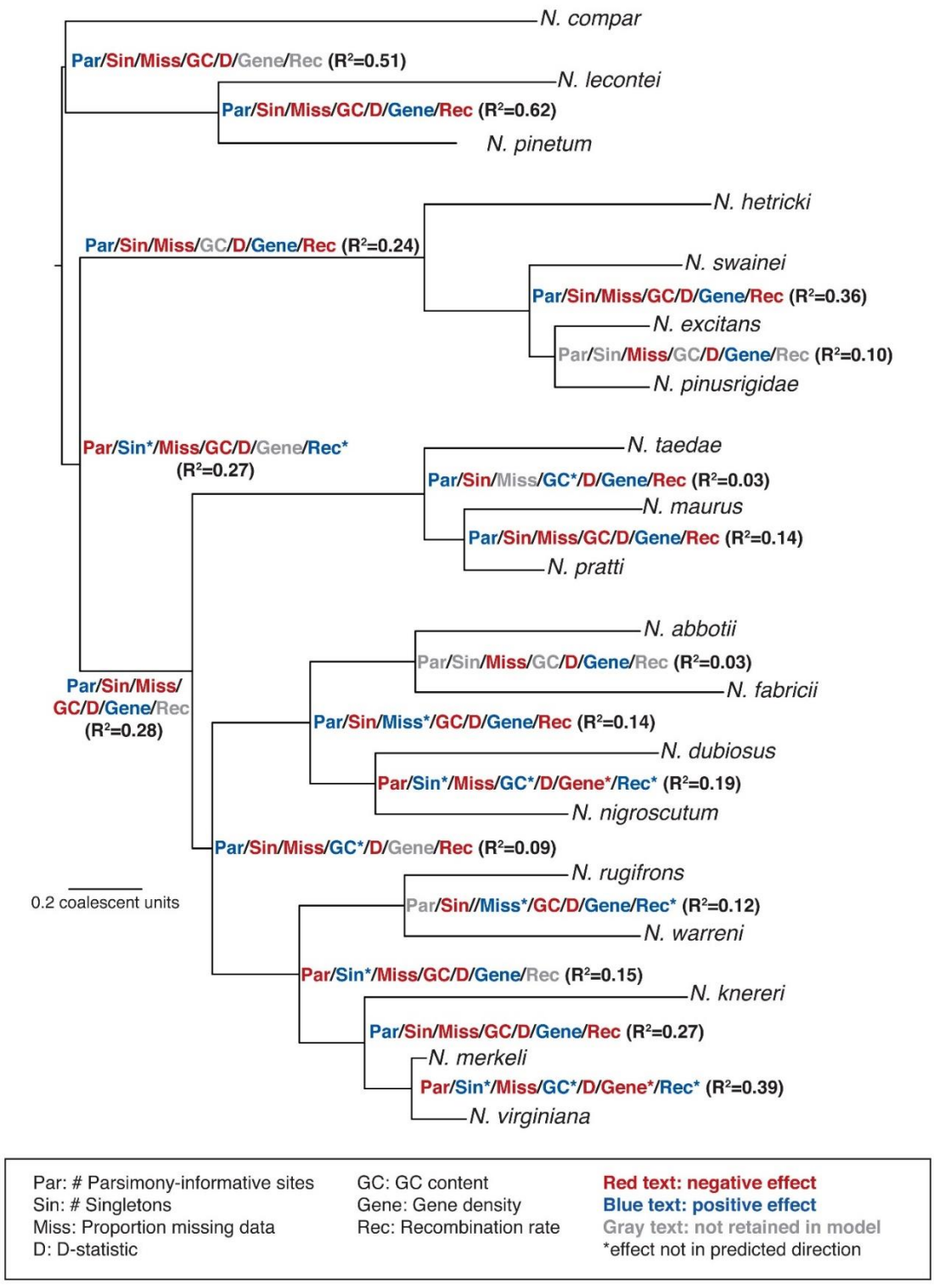

Figure 6