PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Meyers PJ, Doellman MM, Ragland GJ, Hood GR, Egan SP, Powell THQ, Nosil P, Feder JL. 2020 Can the genomics of ecological speciation be predicted across the divergence continuum from host races to species? A case study in *Rhagoletis*. *Phil. Trans. R. Soc. B* **375**: 20190534.
http://dx.doi.org/10.1098/rstb.2019.0534

Accepted: 25 February 2020

One contribution of 19 to a theme issue 'Towards the completion of speciation: the evolution of reproductive isolation beyond the first barriers'.

Subject Areas:
evolution, genomics

Keywords:
genomics of diapause, *Rhagoletis pomonella*, *Rhagoletis mendax*, host races, sibling species, sympatric

Author for correspondence:
Meredith M. Doellman
e-mail: mdoellma@nd.edu

†Co-first authors.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5018477.

THE ROYAL SOCIETY PUBLISHING

Can the genomics of ecological speciation be predicted across the divergence continuum from host races to species? A case study in *Rhagoletis*

Peter J. Meyers[1,†], Meredith M. Doellman[1,†], Gregory J. Ragland[1,2,3], Glen R. Hood[1,4], Scott P. Egan[1,5,6], Thomas H. Q. Powell[1,7], Patrik Nosil[8,9] and Jeffrey L. Feder[1,2,6]

[1]Department of Biological Sciences, and [2]Environmental Change Initiative, University of Notre Dame, Notre Dame, IN 46556, USA
[3]Department of Integrative Biology, University of Colorado Denver, Denver, CO 80217, USA
[4]Department of Biological Sciences, Wayne State University, Detroit, MI 48202, USA
[5]Department of Biosciences, Rice University, Houston, TX 77005, USA
[6]Advanced Diagnostics and Therapeutics Initiative, University of Notre Dame, Notre Dame, IN 46556, USA
[7]Department Biological Sciences, Binghamton University, Binghamton, NY 13902, USA
[8]Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK
[9]Centre d'Ecologie Fonctionnelle and Evolutive, Centre National de la Recherche Scientifique, Montpellier 34293, France

MMD, 0000-0002-0903-6590; GRH, 0000-0002-9301-6153; PN, 0000-0002-8271-9005

Studies assessing the predictability of evolution typically focus on short-term adaptation within populations or the repeatability of change among lineages. A missing consideration in speciation research is to determine whether natural selection predictably transforms standing genetic variation within populations into differences between species. Here, we test whether and how host-related selection on diapause timing associates with genome-wide differentiation during ecological speciation by comparing ancestral hawthorn and newly formed apple-infesting host races of *Rhagoletis pomonella* to their sibling species *Rhagoletis mendax* that attacks blueberries. The associations of 57 857 single nucleotide polymorphisms in a diapause genome-wide-association study (GWAS) on the hawthorn race strongly predicted the direction and magnitude of genomic divergence among the three fly populations at a field site in Fennville, MI, USA. The apple race and *R. mendax* show parallel changes in the frequencies of putative inversions on three chromosomes associated with the earlier fruiting times of apples and blueberries compared to hawthorns. A diapause GWAS on *R. mendax* revealed compensatory changes throughout the genome accounting for the earlier eclosion of blueberry, but not apple flies. Thus, a degree of predictability, although not complete, exists in the genomics of diapause across the ecological speciation continuum in *Rhagoletis*. The generality of this result is placed in the context of other similar systems.

This article is part of the theme issue 'Towards the completion of speciation: the evolution of reproductive isolation beyond the first barriers'.

1. Introduction

The degree to which evolution is predictable is a long-standing issue in biology [1–4]. The question is important because if population genetics theory cannot forecast the phenotypic and genetic composition of populations for more than a few generations, then it has been contended that adaptation by natural selection may mainly be a local phenomenon in time and space [1,2]. In this regard, uncertainty in deterministic projections from microevolutionary process to macroevolutionary pattern can result from many sources, including unpredictable changes in environmental conditions through time [5], lack of relevant standing variation

[6], pleiotropy and epistasis [7,8], and contingency in the order that new mutations occur [9]. Much of evolution may also be influenced by stochastic processes, such as genetic drift and population bottlenecks associated with founding events [10], as well as catastrophic incidents and mass extinctions that randomly prune the tree of life [5]. Nevertheless, some successes in predicting, or at least in understanding the outcome(s) of evolutionary change, have come from laboratory studies of organisms with short generation times and manipulative or observational field studies (reviewed in [3,4]). Additional support for predictability has come from comparative studies which document repeated convergent or parallel changes among different evolutionary lineages in response to similar environmental challenges [11,12].

Speciation is a key consideration largely missing from studies assessing predictability (although see [13]), which is critical for connecting microevolutionary process with macroevolutionary pattern. To date, much of the experimental and comparative work on the predictability of evolution has focused on either relatively short-term adaptation within populations prior to speciation, or longer-term convergent/parallel phenotypic responses among phylogenetic lineages, mainly after speciation [3]. In this regard, ecological speciation may be particularly relevant for connecting and assessing the predictability of microevolutionary process in generating macroevolutionary pattern [13]. Under the ecological theory, speciation is initiated and phylogenetic diversity created by divergent natural selection associated with populations inhabiting and differentially adapting to novel environments or competing for shared resources [14]. If ecological speciation is common, then this would imply a significant role for divergent natural selection in the formation of new biodiversity and potentially a degree of predictability in the translation of intraspecific variation into interspecific differences.

The existence of extensive ancestral standing variation increases the prospects that ecological speciation is predictable and testable [6,15–18]. Specifically, genome-wide-association studies (GWAS) and manipulative selection experiments may be performed in the field or laboratory for the key phenotype(s) under divergent natural selection and underlying ecologically based reproductive isolation (RI) between populations. Evolutionary predictability can then be tested as the degree to which genomic associations or differences in allele frequencies measured in the GWAS or selection experiments are reflected in patterns of differentiation between taxa in nature. This strategy would have increased relevance when performed for multiple taxa at varying stages of divergence along the 'speciation continuum', from newly formed and partially reproductively isolated ecological races to later stage and more strongly isolated species [19]. The central issue is whether standing variation affecting traits contributing to ecologically based RI is retained for extended periods of time in ancestral populations to contribute in a parallel manner to the adaptive radiation of a series of derived taxa. In this regard, chromosomal inversions can increase the likelihood of observing parallel changes, as they may package suites of genes together into alternate co-adapted haplotypes displaying reduced recombination and elevated linkage disequilibrium (LD) [15,16]. Such a genomic architecture could facilitate: (i) the retention of standing adaptive variation through time in ancestral populations; (ii) its use in ecological speciation; and (iii) its statistical detection in GWAS, selection experiments and population surveys.

Here, we employ a combined GWAS and comparative population survey strategy to test for concordant patterns of genomic divergence along the speciation continuum for fruit flies in the *Rhagoletis pomonella* (Diptera: Tephritidae) sibling species group. Phylogenetic studies have implied that the hawthorn (*Crataegus* spp.)-infesting host race of the species *R. pomonella* sequentially gave rise to the majority of taxa in the group via a series of sympatric host plant shifts in North America [20–22]. One of these derived taxa is the sibling species, *Rhagoletis mendax*, that attacks blueberries (*Vaccinium corymbosum*) in the northeastern USA and Canada, and deerberries (*Vaccinium stamineum*) in the southeastern USA [21]. Most recently (circa 1860), the ancestral hawthorn race gave rise via a sympatric shift to the domesticated apple (*Malus domestica*)-infesting host race of *R. pomonella* [20,23,24].

To test for predictability in *Rhagoletis* divergence, we focus on diapause life-history timing, as diapause has been shown to be a major axis of host plant-related adaptation contributing to the ecological RI and sympatric differentiation of these flies (see below for details) [25]. Specifically, we investigate whether and how a GWAS of diapause timing in the ancestral hawthorn host race [26] predicts the pattern of population divergence moving from the recently formed apple race to the more distantly derived *R. mendax* at a field site in Fennville, MI, USA, where all three taxa co-occur (table 1, predicted patterns). We also conduct a GWAS of blueberry-infesting flies at the Fennville site to assess how well standing variation affecting eclosion time in the derived *R. mendax* predicts population divergence and is correlated with standing variation for diapause in the ancestral hawthorn race. In this regard, *R. mendax* is not the immediate sister taxon to *R. pomonella*, as this distinction belongs to the unnamed flowering dogwood fly (host: *Cornus florida*) [21]. However, *R. mendax* is the next most closely related sibling species to *R. pomonella* [21], with little evidence for ongoing gene flow between blueberry and apple or hawthorn flies in nature [28,29]. Thus, comparisons among these flies allow for an analysis of genomic divergence across the speciation continuum from recently formed and partially reproductively isolated sympatric *R. pomonella* host races [23] to the more strongly isolated and described sympatric sibling species, *R. mendax* [29].

The timing of the overwintering pupal diapause is a critical host-related adaptation because the various plants attacked by *R. pomonella* group flies tend to fruit at different times of the summer and autumn (figure 1) [21]. Adult *Rhagoletis* are univoltine (have one generation a year) and generally live for less than a month in nature [30]. Thus, flies must adapt the timing of their life cycle to eclose as adults when ripe host fruit are available to mate on and oviposit into (figure 1*a*) [31–34]. As a result, adults are sexually active at offset times during the season, generating prezygotic allochronic RI between taxa [35]. In addition, hybrids can have reduced fitness because they eclose at suboptimal times for using the host fruit of either parent, generating postzygotic RI [31]. Thus, eclosion time is a 'magic trait', in that divergent phenological selection to adapt to differences in host fruiting time generates non-random mating, reducing gene flow and facilitating speciation [36,37].

One aspect of the timing of diapause critical for host adaptation is diapause termination (figure 1*a*), which must be timed to occur in late winter and early spring to ensure that flies complete development and eclose from the soil as adults when ripe host fruit are present [38]. At Fennville, MI,

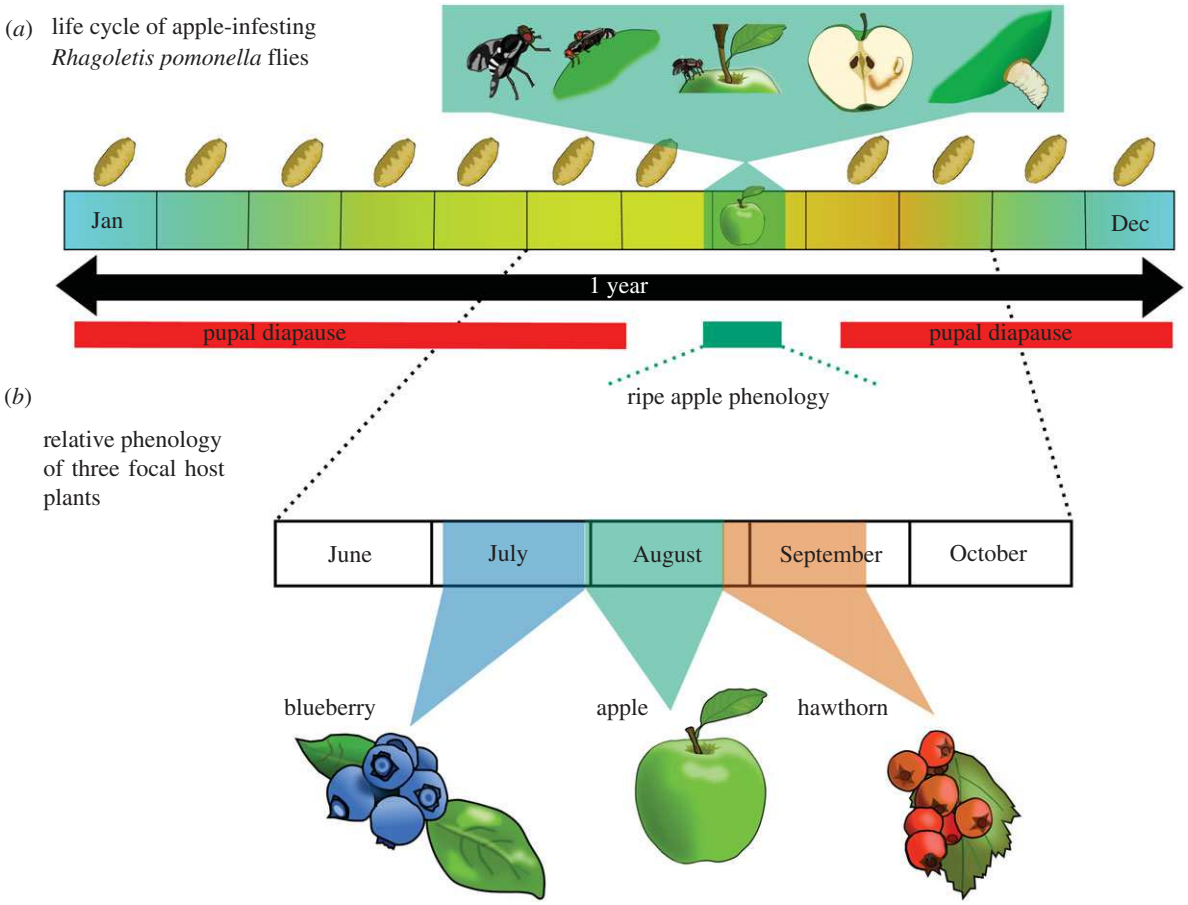

Figure 1. Natural history of *R. pomonella* apple and hawthorn host races and *R. mendax*. (*a*) The univoltine life cycle of *Rhagoletis* flies. Flies eclose as adults to coincide with the availability of their ripe host fruit, mate on or near the fruit, and (females) oviposit eggs into the fruit. Larvae develop inside the fruit, emerge, burrow into the soil, pupate and enter diapause. Note that flies infesting earlier fruiting plants must maintain diapause for a longer warm period (indicated in orange) prior to the onset of winter (indicated in blue). (*b*) The relative phenology of the three focal host plants, in order of availability of ripe host fruit for *R. mendax* (blueberry, *Vaccinium corymbosum*), *R. pomonella* apple race (apple, *Malus domestica*) and *R. pomonella* hawthorn race flies (downy hawthorn, *Crataegus mollis*).

Table 1. Predictions for relationships between eclosion variation and genomic divergence.

1	apple flies eclose before hawthorn flies and blueberry flies eclose before apple flies. Therefore, the order of frequency divergence for eclosion-associated loci should also be arrayed blueberry > apple > hawthorn
2	the magnitude of allele frequency differences should mirror this order, with allele frequency differences for eclosion-associated loci being greater between the *R. pomonella* hawthorn race and its sibling species *R. mendax* than between the hawthorn and apple host races
3	previous work has established that loci associated with eclosion timing in *R. pomonella* are concentrated in the putative inversions on chromosomes 1–3 [26]. Therefore, the frequencies of these inversions should vary among the host races and *R. mendax* according to predictions 1 and 2
4	recent work has suggested that alleles for later eclosion time and deeper initial diapause are linked within inversion haplotypes on chromosomes 1–3 [27]. Selection for survival during a longer prewinter period and hence on deeper initial diapause in apple flies than hawthorn flies results in paradoxically higher frequencies of 'late eclosion' alleles in apple flies compared to hawthorn flies. Because blueberry flies are active earlier in the year than apple flies, they should experience even stronger selection for deep initial diapause depth and thus the highest frequencies of 'late eclosion' alleles
5	owing to the linkage or pleiotropy between alleles for deep initial diapause and late eclosion, there are probably compensatory changes allowing apple flies to eclose earlier than hawthorn flies and blueberry flies to eclose earlier than both of the host races
6	given that predictions 1–5 hold, these compensatory changes are not expected to be associated with the putative inversions on chromosomes 1–3

the site of the current study, fruit on apple varieties favourable for *R. pomonella* survival tend to ripen two to four weeks earlier than those of downy hawthorn (*C. mollis*), while blueberries, the host of *R. mendax*, ripen two to three weeks earlier than apples (figure 1*b*). Consequently, blueberry flies terminate diapause and eclose earlier than apple flies, while apple flies eclose earlier than hawthorn flies, both in the field and under controlled rearing conditions [31].

A second component of diapause under divergent selection is the intensity or depth at which flies initially enter diapause following pupariation (figure 1*a*) [26,27,39,40]. If exposed to warm temperature for too long a period prior to winter, a proportion of pupae (as high as 50–60% in *R. pomonella*) will not maintain diapause and initiate adult development [39]. Such 'non-diapause' flies will suffer reduced fitness either because they will complete adult development prior to winter and have no suitable host fruit resources available, or will not be in an optimal state for cold conditions and suffer increased mortality while overwintering. Thus, it has been hypothesized that the longer prewinter pupal exposure to warm temperatures experienced by flies infesting earlier fruiting hosts (e.g. blueberries and apples) exerts stronger selection pressures to enter a more intense initial diapause and avoid non-diapause development, compared to flies attacking later fruiting hosts (e.g. hawthorns) [39].

Genomic analyses of adult eclosion time have been performed for the *R. pomonella* host races at Fennville [26]. Four important results emerged from these studies. First, adult eclosion time is a highly polygenic trait with significant allele frequency differences found for many single nucleotide polymorphisms (SNPs) between early and late eclosing apple and hawthorn flies. Second, the genetics of adult eclosion time were highly correlated between the apple and hawthorn host races [26]. Third, SNPs showing the largest associations with eclosion time were concentrated on chromosomes 1–3 (haploid $n = 6$ in *Rhagoletis*) and, in particular, were greatest for loci displaying high levels of LD with one another, putatively associated with inversion polymorphism, on these three chromosomes. We use the term 'putative' to describe these inversions because the cytology of *R. pomonella* is not of high enough quality to allow us to adequately resolve polytene chromosomes in the fly to physically (visually) confirm the presence of chromosomal rearrangements [20]. However, the high LD displayed among particular sets of loci [26,40,41], combined with the observations of significant heterogeneity in estimated recombination rates and variation in locus mapping order [42], allow inversions to be inferred based on classic genetic evidence [43]. Fourth and finally, as a consequence of the inversions, adult eclosion time was more highly correlated between the apple and hawthorn host races for the high LD class of SNPs mapping to chromosomes 1–3 than for the remainder of the genome.

Genomic studies have also been conducted on initial diapause intensity for the apple and hawthorn host races and the phenotype shown to be genetically correlated with adult eclosion time [27,39,44]. Specifically, alleles associated with greater initial pupal diapause intensity prior to chilling were correlated genome-wide with later adult eclosion following winter ($r = 0.525$, $p < 0.0001$, $n = 5990$ SNPs; [27]). The correlation was strongest for loci displaying high levels of LD with one another on chromosomes 1–3, associated with the putative inversion polymorphisms ($r = 0.808$, $p < 0.0001$, $n = 284$ SNPs; [27]).

The genetic correlation between greater initial diapause intensity and later eclosion time is aligned with historical and current geographical selection pressures on the ancestral hawthorn race [27,44,45]. Across eastern North America, hawthorns attacked by *R. pomonella* tend to fruit later in the season with decreasing latitude and increasing temperatures [45]. As a result, hawthorn flies from southern latitudes must both withstand longer warm prewinter periods as pupae and eclose later as adults than those in northern latitudes [31,32]. Consequently, loci affecting diapause display pronounced latitudinal allele frequency clines across hawthorn fly populations, with alleles related to greater initial diapause intensity and later adult eclosion found at higher frequencies in southern than in northern populations (electronic supplementary material, figure S1; [27,44,45]). Clinal variation is especially strong for loci associated with the putative inversions on chromosomes 1–3 [45]. Thus, collective evidence suggests that at least two alternate inversion haplotypes segregate within the *R. pomonella* hawthorn race, on each of chromosomes 1–3. The inversion haplotype(s) harbouring alleles associated with later adult eclosion and with more intense initial pupal diapause is more common in southern hawthorn fly populations. The other haplotype(s) is more common in the North, correlating with earlier adult eclosion and relaxed selection for intense initial pupal diapause. At this time, the degree to which the genetic correlation between diapause termination and initial depth reflects pleiotropy versus linkage of causal variants is not known [27].

The historical and geographical selection pressures on the ancestral hawthorn race and packaging of adaptive variation in inversion polymorphisms are antagonistic in direction, however, to those experienced by the derived apple race at local sympatric sites. Despite fruiting time varying latitudinally, the two to four week earlier phenology of apples and hawthorns is maintained at sympatric sites [31]. Thus, at local sites, the earlier fruiting time of apple than hawthorn is expected to exert stronger selection pressures on the apple than hawthorn race for increased initial diapause depth and earlier adult eclosion. Consistent with these expectations, controlled rearing experiments have demonstrated that apple fly populations have a lower incidence of direct development and eclose earlier as adults than co-occurring hawthorn fly populations, including at the sympatric Fennville site [27,31]. However, as discussed above, the putative inversion haplotypes on chromosomes 1–3 show an association of deeper initial diapause depth with later adult eclosion. While such a relationship is aligned with the latitudinal selection pressures on hawthorn flies, it is not for apple flies at local sites, as the optimal fitness combination for the apple compared to hawthorn race is greater initial diapause intensity and earlier, not later, adult eclosion. Indeed, GWAS, laboratory selection experiments, and population surveys have shown that apple flies do, in fact, possess higher frequencies of alleles associated with increased initial diapause intensity (and, thus, later eclosion) than hawthorn flies at sympatric sites, especially for the putative inversions on chromosomes 1–3 [27,40,44,45]. Thus, the correlation between greater diapause intensity and later adult ecolsion within these inversion haplotypes does not differ between the host races. This suggests that gene-by-environment interactions [46] and/or other yet to be identified compensatory variants, probably not associated with the inversions on chromosomes 1–3, are responsible for the earlier eclosion of apple than hawthorn flies at sympatric sites (table 1).

Given that fruits on blueberries ripen even earlier than those on apples, we predict that the same genomic pattern, only more pronounced, should be observed for *R. mendax* (table 1). Specifically, selection for increased initial diapause intensity should be manifested primarily by loci associated with the putative inversions on chromosomes 1–3 displaying high frequencies of 'southern hawthorn' race alleles in *R. mendax*. While blueberry-origin pupae have not been tested under the same controlled conditions as the *R. pomonella*

host races, it is known that *R. mendax* is generally much more recalcitrant to non-diapause development than apple or hawthorn flies (J.L.F. 1989, personal observation). As these diapause intensity alleles are also associated with later eclosion, we predict compensatory changes for earlier ecolsion time not associated with the inversions. However, it would remain to be determined whether these compensatory changes are the same or differ between *R. mendax* and the apple race, for the genomic architecture underlying adaptive phenotypes is subject to change with increasing time of separation between derived and ancestral populations.

We test these predictions concerning the genomics of diapause adaptation across the early and late stages of speciation in *Rhagoletis* through a three-pronged approach. First, we conduct a double digest restriction site-associated DNA sequencing (ddRADseq) survey quantifying genome-wide differentiation among host races and sibling species at the sympatric Fennville site where blueberry-, apple- and hawthorn-infesting populations of flies co-occur. Second, we assess the degree to which the eclosion time study for the hawthorn race is predictive of patterns of genomic differentiation among fly taxa. For these analyses, we focused on the eclosion time GWAS for the hawthorn rather than the apple race because (i) hawthorn-infesting *R. pomonella* are the presumed ancestor of *R. mendax* and then apple flies [21,22], and (ii) the genomics of eclosion time, as discussed above, are strongly correlated between the host races. Finally, we perform a GWAS analysing early versus late eclosing samples of *R. mendax* flies at the Fennville site. The GWAS allowed us to compare results for *R. mendax* to previously performed studies at the same site in the same year for apple and hawthorn flies [26] to determine if the genomic architecture of current standing variation for adult eclosion time is similar between the sibling species.

2. Methods

(a) Fly collection, population survey and eclosion time genome-wide-association studies

Standard *Rhagoletis* collection and rearing methods were used, as described in Egan *et al*. [40]. Genomic DNA sequencing data comparing apple and hawthorn fly populations at the Fennville, MI, site (42°35054.5892″ N, 86°0905.0220″ W) come from Doellman *et al*. [45,47] for flies sampled in 2009 ($n = 93$ apple and $n = 96$ hawthorn flies). These results were augmented here with new sequencing data produced for *R. mendax* at Fennville ($n = 48$ flies), also collected in 2009 [48]. We consider the fly populations infesting the three host plants at Fennville to be sympatric, as they all reside within 1 km of one another.

The sequencing data in the eclosion time GWAS for the *R. pomonella* host races come from Ragland *et al*. [26,49]. These results were supplemented by new sequencing data for an eclosion time study performed on *R. mendax* flies also collected at Fennville in 2009 [48]. Methods were the same for the *R. mendax* GWAS as those used for the host races (see the electronic supplementary material). However, owing to a lower overall number of adults eclosing in the *R. mendax* study, we expanded the earliest and latest eclosing samples of flies from 3% and 97% for hawthorn and apple flies to 9.4% and 90.6%, respectively, for blueberry flies to match the $n = 48$ individuals sequenced in each sample for the host races. We focus on eclosion time owing to the recalcitrance of *R. mendax* to direct, non-diapause development, making it difficult to accrue adequate sample sizes for a GWAS on initial diapause intensity for this species.

(b) Genotyping

ddRADseq libraries were constructed following Egan *et al*. [40] and sequenced (100 bp paired-end reads) on an Illumina HiSeq 2000 platform (Beijing Genomics Institute). Methods for demultiplexing, trimming and aligning reads, and calling and filtering SNP genotypes are detailed in the electronic supplementary material. Previous studies of diapause phenotypes and host and geographical variation in *R. pomonella* were based on 10 241 variable SNPs [26,44,45]; here, this number is increased to 57 857 based on reanalysis of the combined dataset comprising prior *R. pomonella* and new *R. mendax* sequences.

(c) Analysis of host race and species divergence

Population-level allele frequencies were calculated by averaging across the genotype likelihoods for each individual, with loci having a minimum sample size of $n = 8$ in each of the three taxa (apple and hawthorn host races, and *R. mendax*) retained for frequency analysis. PGDSPIDER was used to convert genotype data from vcf format to STRUCTURE format for additional analyses of the level divergence among taxa [50]. Principal component analysis (PCA) was conducted in the R package *adegenet* [51,52]. Population structure analysis was performed using the variational Bayes implementation in FASTSTRUCTURE, with the simple prior and convergence criterion of 10^{-6} [53,54]. We ran models from $K = 1$ to $K = 6$, used the *chooseK.py* script to select the optimum K value representing the number of genetically differentiated subpopulations for each comparison, and the R package *pophelper* to visualize the results [55].

(d) Genome structure analysis

Previous studies based on sequencing data for apple and hawthorn flies from test crosses and natural populations at a sympatric field site in Grant, MI, identified regions of high LD putatively associated with inversion polymorphism on each of the five major chromosomes comprising the *R. pomonella* genome [26,40,42]. The fly also has a small, heterochromatic dot 6th chromosome [20] for which no genetic marker is currently known. Note that these loci have been assigned to chromosomes, but not explicitly ordered within the chromosome, nor are they aligned to a contiguous genome assembly providing a linear order, precluding examination of LD in a linear order along each chromosome. We expanded upon these earlier results here by characterizing LD patterns in the hawthorn and blueberry fly populations at Fennville for a subset of 4421 SNPs mapped to the five major chromosomes [26,40]. Specifically, we used the R package *GUS-LD* [56] to calculate the r^2 measure of LD between all pairs of SNPs with a minor allele frequency of at least 0.10. Analyses were conducted separately for each chromosome and for populations of *R. mendax* and the *R. pomonella* hawthorn race. For each species and chromosome, we then used the *igraph* package in R [57] to generate Fruchterman–Reingold network plots as in Kemppainen *et al*. [58], which represented SNPs as nodes connected to one another by edges, when the pairwise LD (r^2) between SNPs was 0.80 or greater. This allowed us to identify clusters of SNPs in high LD with one another that we associate with the putative inversion polymorphism on each of the five major chromosomes constituting the *R. pomonella* and *R. mendax* genomes. To be conservative, we restricted further analysis of SNPs included in the set of high LD loci for the hawthorn race to those residing in $r^2 \geq 0.80$ clusters at both the Grant and Fennville sites.

(e) Genomic architecture of eclosion time

To determine what proportion of the phenotypic variation in eclosion time in the hawthorn race of *R. pomonella* and *R. mendax* is explained by heritable genetic variation, and how much of that genetic variation is accounted for by loci of major effect, we ran a

Bayesian sparse linear mixed model as implemented in GEMMA [59]. This method was specifically developed to 'learn' the genomic architecture of a trait from the data in order allow simultaneous estimation of a limited number of large effects, via variable selection, while also including the combination of many small effects. This model also allows the estimation of hyperparameters including the proportion of the variance in the phenotype explained by the genotype (PVE) and the proportion of the genetic variance explained by the 'measurable' genetic effects (PGE). The categorical response variable was eclosion time, coded as early (0) or late (1), and the predictor variables were genotype likelihoods. For each species, 10 independent chains were run, with 10^6 burn-in iterations, followed by 5×10^6 sampling iterations, with a thinning period of 10^3.

We used a non-parametric, Monte Carlo simulation approach to test for SNPs displaying significant allele frequency differences between both early and late eclosing hawthorn flies and early and late eclosing blueberry flies. Two populations of $n = 48$ each were generated by permuting population labels across whole fly genotypes from the combined pool ($n = 96$) of early and late eclosing hawthorn flies or early and late eclosing blueberry flies. A null distribution was derived by taking the difference between these two permuted samples for $n = 10\,000$ replicates. An SNP was considered significant if it displayed an observed difference in the lower 2.5% or upper 97.5% of the null distribution. By randomly permuting whole fly genotypes in this and subsequent analyses, we accounted for genetic linkage and LD between SNPs in the tests.

To assess the genomic distributions of significant eclosion time SNPs, we examined the locations of mapped loci on the LD plots generated for hawthorn and blueberry flies at the Fennville site. We tested for a relationship in the genomics of eclosion time between the hawthorn host race and *R. mendax* by calculating Pearson's correlation coefficients (r) between SNP allele frequency differences between early and late eclosing samples of flies for these two taxa. For these correlations and those discussed below, we included only SNPs for which the rare allele was present at a frequency of greater than or equal to 0.05 in the combined early and late eclosing samples of flies being compared. Again, we used a permutation approach to test for significance by generating a null distribution of correlation estimates. We conducted 10 000 permutations of population labels across whole fly genotypes of pooled early and late eclosing hawthorn flies and pooled early and late eclosing *R. mendax*. At each permutation, the correlation coefficient was calculated, as above.

(f) Correlations of eclosion time genome-wide-association studies with allele frequency differences in nature

We tested for associations between diapause life-history timing and genomic patterns of differentiation by calculating correlations between SNP allele frequency differences between early and late eclosing samples of hawthorn or blueberry flies in the eclosion time GWAS versus those between the host races or the hawthorn race and *R. mendax* in the population survey. We used a permutation approach, as described above, to test for significance.

Discriminant analysis of principal components (DAPC) was also used to visualize the associations of genomic (SNP) variation underlying eclosion time and population divergence among taxa in nature using the R package *adegenet* [60]. The model was first trained to discriminate between the genotypic composition of early and late eclosing samples of hawthorn or *R. mendax* flies in the GWAS. The *ascore* function was used to choose the number of principal components (PCs) to retain in each model. The full population samples, as well as the early and late eclosing samples, were then projected onto the discriminant function to graphically assess whether the genomic differences among host race and species could be predicted by the eclosion time studies.

3. Results

(a) Genomic divergence across the speciation continuum

The population survey showed that *R. mendax* is genomically distinct from the apple and hawthorn-infesting host races of *R. pomonella*. PCA and fastSTRUCTURE (best supported model, $K = 2$), indicated strong differentiation between *R. mendax* and *R. pomonella* (figure 2*a*,*b*). No fly was identified as a putative F_1 hybrid between *R. mendax* and *R. pomonella* in either analysis. A total of 11 SNPs displayed fixed frequency differences for alternative alleles (i.e. 1.0 versus 0.0) between *R. mendax* and all *R. pomonella*, with many loci showing pronounced differences of greater than or equal to 0.60 between *R. mendax* and the apple race ($n = 1872$ SNPs), and *R. mendax* and the hawthorn race ($n = 2066$ SNPs; figure 2*c*).

In comparison to *R. mendax*, PCA and fastSTRUCTURE did not separate the apple and hawthorn host races into diagnostically discrete non-overlapping genotypic clusters (figure 2*a*,*b*; electronic supplementary material, figure S2). In addition, only one locus displayed an allele frequency difference greater than or equal to 0.60 between the apple and hawthorn populations, although several showed differences greater than or equal to 0.30 ($n = 578$ SNPs; figure 2*c*). As a result, the host races, while not forming discrete clusters in the PCA, did not overlap completely (figure 2*a*), consistent with their characterization as partially differentiated ecological and genetic entities [44,45].

(b) Hawthorn race eclosion time genomics predict population divergence of host races and species

The putatively ancestral *R. pomonella* hawthorn race harboured substantial variation underlying the eclosion time phenotype. For the hawthorn race GWAS, the 95% credible interval (CI) of the PVE between early versus late eclosing flies including all 57 857 SNPs genotyped was 0.776–0.999, centred on 0.97. The 95% CI for the per cent of genetic variance (PGE) explained by loci with measurable effects was wider, ranging from 0.002 to 0.998 and centred on 0.617. The median estimated number of loci of measurable effect contributing to diapause termination time was 39 (95% CI 1–266). Given the relatively uncertain estimates of PGE and the number of loci of measurable effect, yet a large estimate for PVE, polygenic (infinitesimal effect) variation appears to account for a sizeable amount of the variance in eclosion time for hawthorn flies, consistent with previous results [26].

Genetic associations of SNPs in the hawthorn race GWAS (early–late) significantly predicted genome-wide allele frequency differentiation between the apple and hawthorn host races (apple–hawthorn; $r = -0.463$, $p < 0.0001$, $n = 44\,749$ SNPs variable in the hawthorn race GWAS), as well as between *R. mendax* and the hawthorn race (*R. mendax*–hawthorn; $r = -0.227$, $p < 0.0001$, $n = 44\,749$ SNPs; figure 3, grey circles; electronic supplementary material, table S2). These genetic correlations were negative in sign; thus, compared to the hawthorn fly population, both apple and blueberry fly populations tended to have higher frequencies of alleles associated with

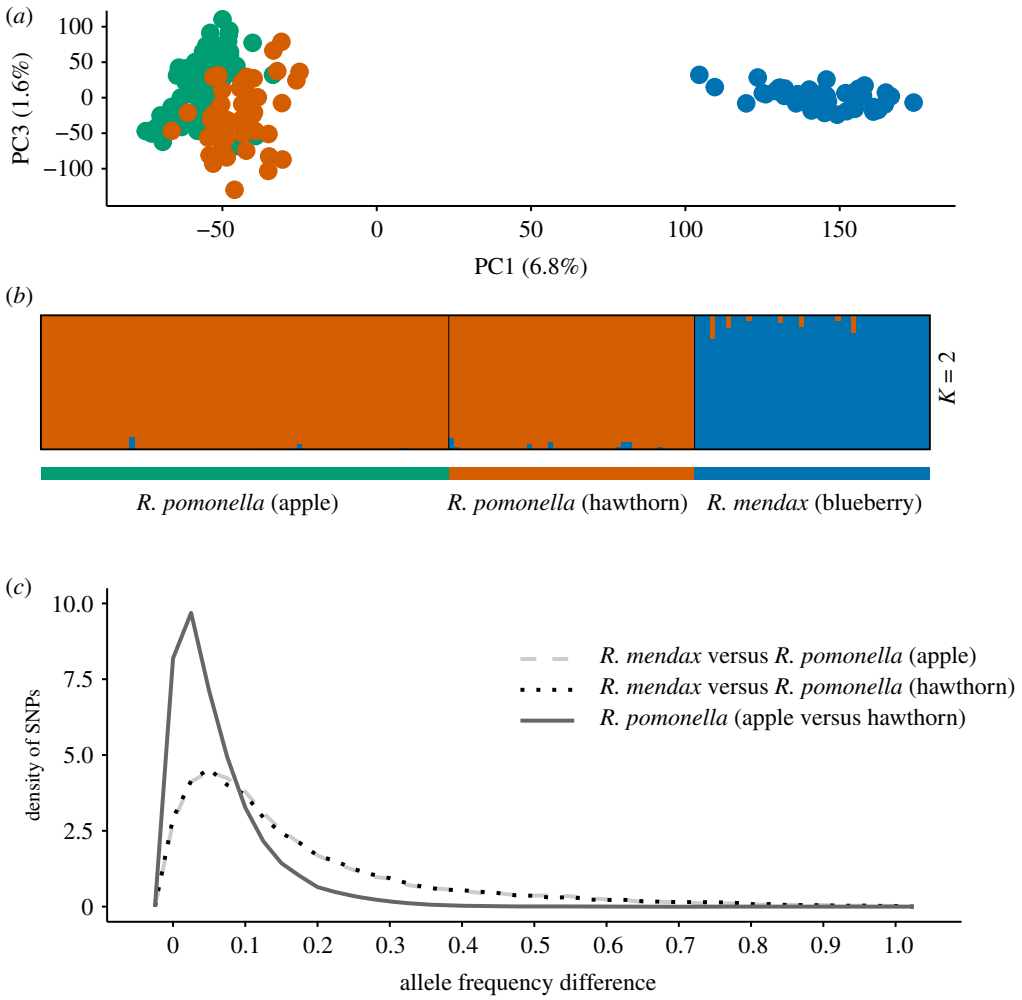

Figure 2. Population genetic structure of hawthorn- and apple-infesting host races of *R. pomonella* and blueberry-infesting sibling species *R. mendax* co-occurring at the sympatric Fennville, MI, field site. (*a*) Plot of the first two principal components for the three taxa based on PCA of all 57 857 SNPs genotyped in the study (axes are scaled relative to the proportion of variation explained); (*b*) fastSTRUCTURE plot of the three fly taxa based on all SNPs, with $K = 2$ determined to be the optimal value for the number of genetically differentiated subpopulations in the analysis; (*c*) distributions of allele frequency differences between pairs of taxa for all SNPs: apple host race and *R. mendax* (light grey dashed line), hawthorn host race and *R. mendax* (black dotted line), apple and hawthorn host races (dark grey solid line). (Online version in colour.)

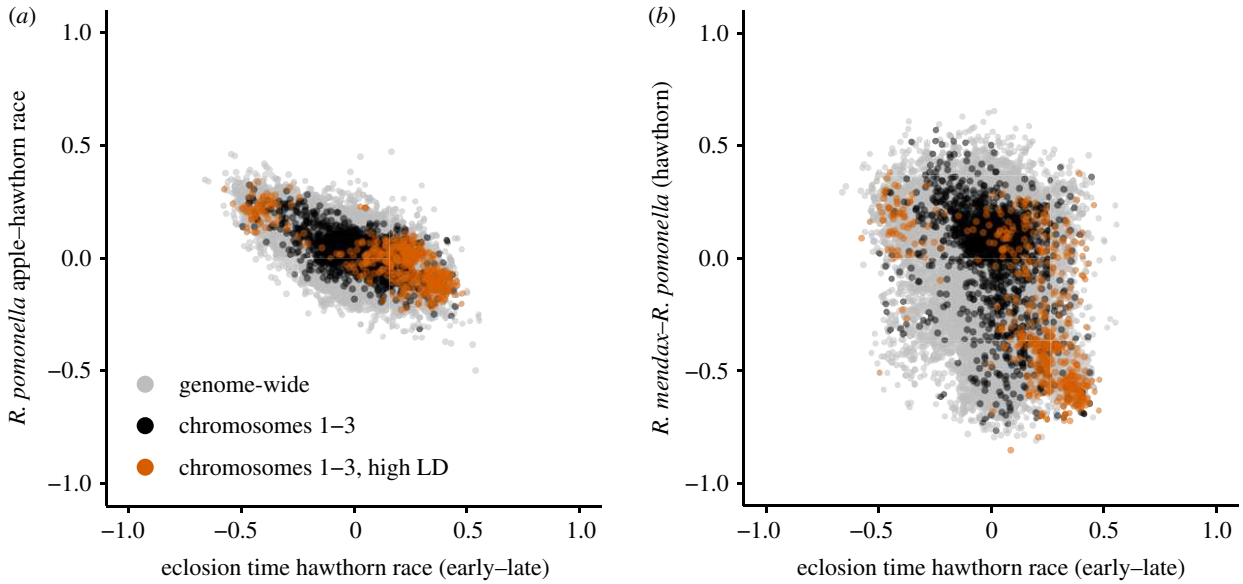

Figure 3. Relationship between the associations of SNPs in the hawthorn race eclosion time GWAS (i.e. allele frequency differences between early and late eclosing samples of flies) and their differentiation between sympatric (*a*) apple versus hawthorn host races and (*b*) *R. mendax* and the hawthorn race. Grey circles represent all SNPs genotyped in the study that were variable in the hawthorn race eclosion time GWAS (apple/hawthorn comparison: $r = -0.463$, $p < 0.0001$, *R. mendax*/hawthorn comparison: $r = -0.227$, $p < 0.0001$, 44 749 SNPs in both cases), dark circles represent loci mapping to chromosomes 1–3 (apple/hawthorn comparison: $r = -0.709$, $p < 0.0001$; *R. mendax*/hawthorn comparison: $r = -0.539$, $p < 0.0001$, 2102 SNPs in both cases) and orange circles are the subset of SNPs belonging to the high LD class of loci on chromosomes 1–3 (apple/hawthorn comparison: $r = -0.831$, $p < 0.0001$; *R. mendax*/hawthorn comparison: $r = -0.632$, $p < 0.0001$, 311 SNPs in both cases).

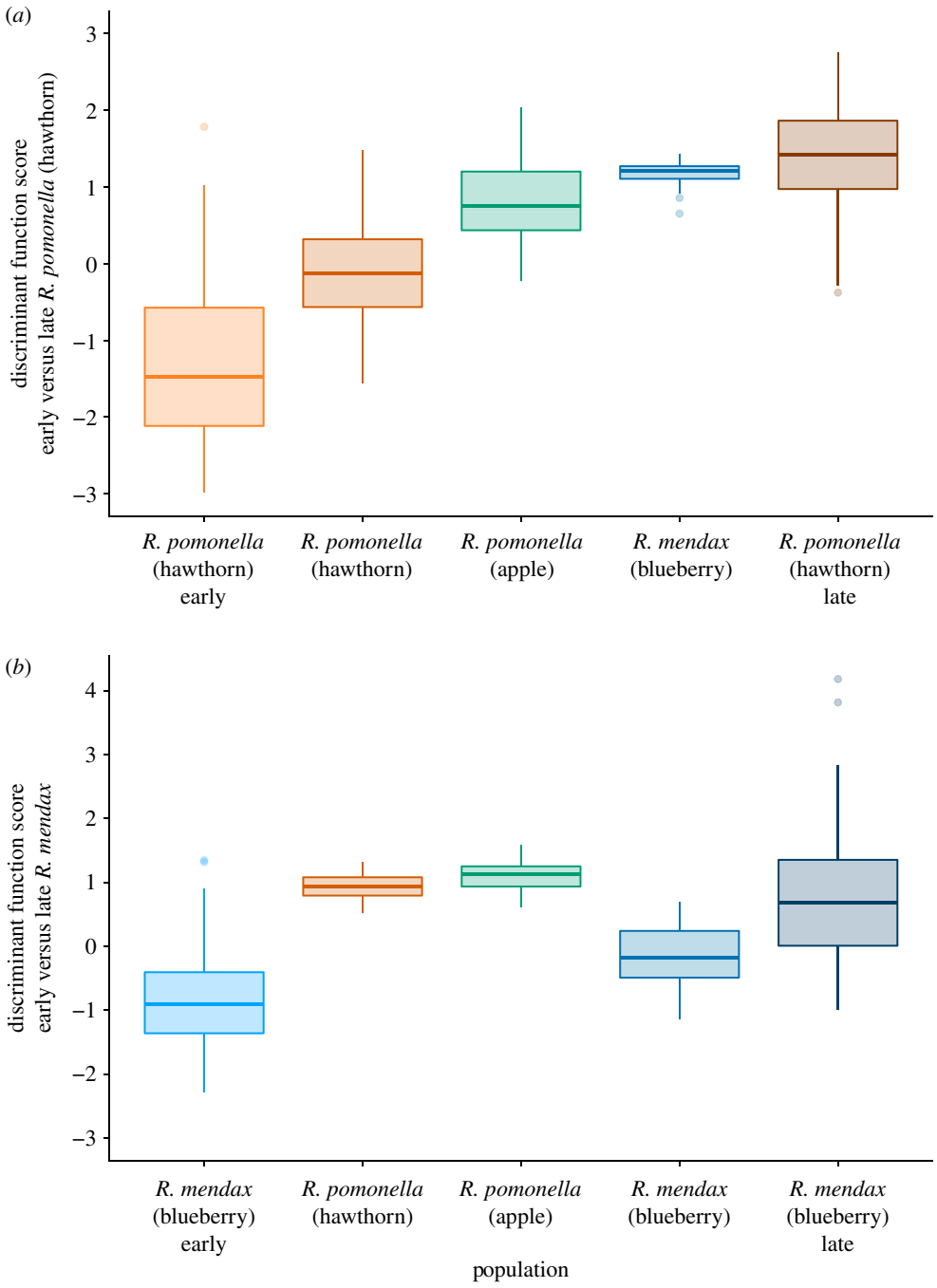

Figure 4. Projections of population data from *R. pomonella* apple and hawthorn host races and *R. mendax* onto discriminant functions of principal components (DAPC), trained to maximize genomic distance between early and late samples in each eclosion time GWAS. Projections of training data are included on the right (early sample) and left (late sample) sides of each graph for comparison. (*a*) Boxplot of scores from the discriminant function between early and late eclosing samples of hawthorn race flies. (*b*) Boxplot of scores from the discriminant function between early and late eclosing samples of *R. mendax*. (Online version in colour.)

later adult eclosion and greater initial diapause depth in the hawthorn race [27]. The correlation with the hawthorn race GWAS was weaker but also negative for divergence between *R. mendax* versus the apple race (*R. mendax*–apple; $r = -0.045$, $p < 0.0001$, $n = 44\,749$ SNPs). Consequently, as forecast, frequencies of alleles associated with later eclosion time were higher in *R. mendax* compared to the apple race. To graphically illustrate this genome-wide pattern, we performed a DAPC trained to discriminate the early and late eclosing samples of hawthorn flies in the GWAS. Population-level discriminant function scores were arrayed from most similar to early eclosing genotypes to most similar to late eclosing genotypes, in the order hawthorn host race, apple host race, *R. mendax*, with *R. mendax* being the most similar to the late eclosing sample of hawthorn flies in the GWAS (figure 4*a*).

(c) Genomic architecture of predictability from the *Rhagoletis pomonella* eclosion genome-wide-association study

The distribution of eclosion associations in the *R. pomonella* GWAS was not homogeneous across the genome. Loci displaying significant allele frequency differences between early and late eclosing hawthorn flies were concentrated on chromosomes 1–3 and, in particular, the high LD

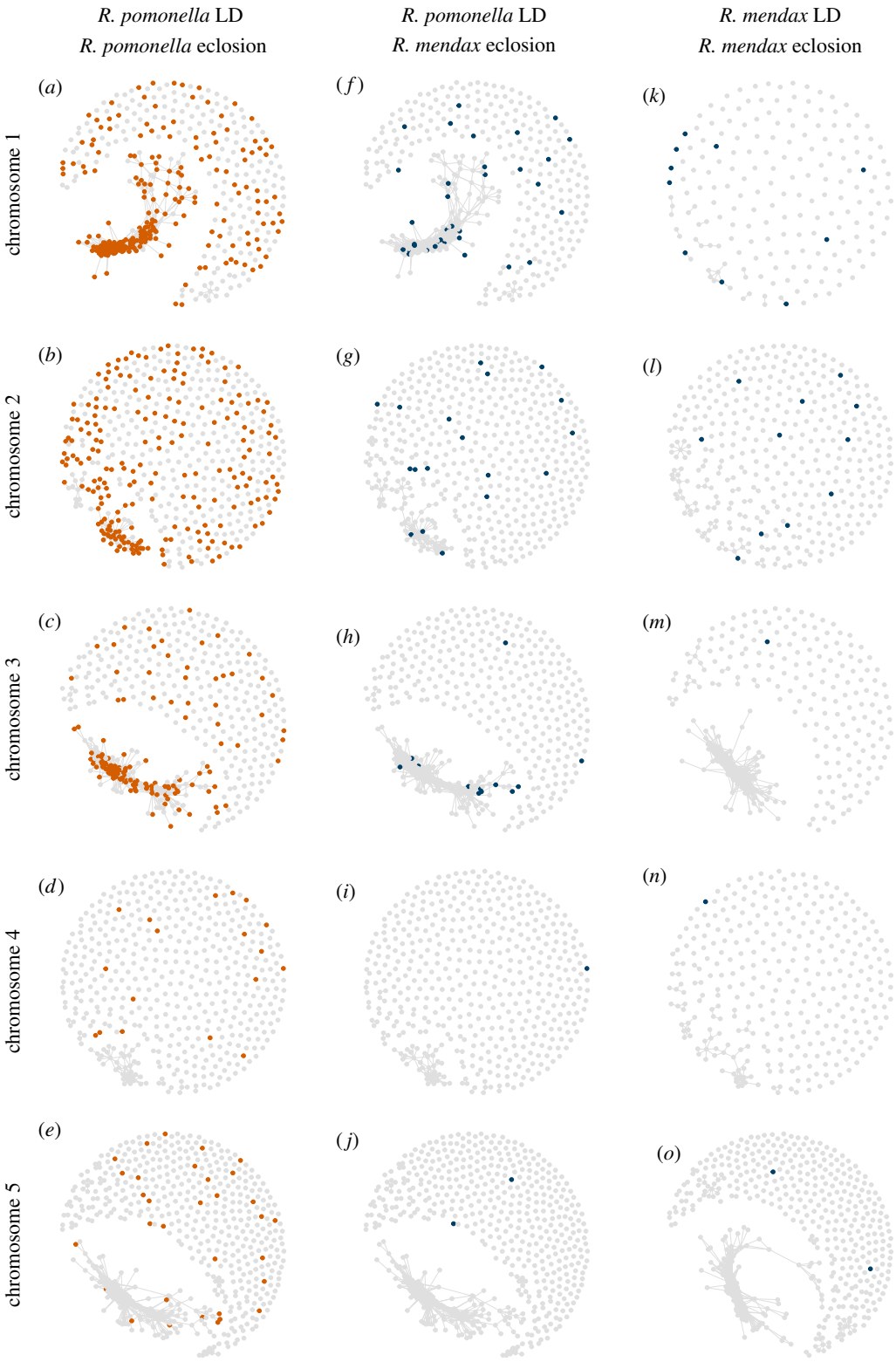

Figure 5. LD network plots for SNPs mapping to chromosomes 1–5. Loci are shown as nodes (circles) connected to one another in clusters by edges, when pairwise LD (r^2) between loci exceeds 0.8. Tight clusters indicate this high pairwise LD between many SNPs, while unconnected SNPs have a pairwise LD below 0.8 with all other SNPs on the same chromosome. Within population, LD is represented for the hawthorn host race of *R. pomonella* (*a–j*) and the blueberry-infesting population of *R. mendax* (*k–o*). Coloured circles represent SNPs showing significant allele frequency differences between the early and late eclosing samples of flies in the eclosion time GWAS performed for the hawthorn host race (orange circles in *a–e*) or for *R. mendax* (blue circles in *f–o*).

clusters putatively associated with inversions in *R. pomonella* (figure 5*a–c*) [26]. Several significantly associated SNPs also resided in lower LD regions on chromosomes 1–3, but few mapped to chromosomes 4 or 5 (figure 5*d*,*e*). Thus, as predicted, correlations between eclosion association and host race and/or species divergence were most pronounced when considering the high LD classes of loci on chromosomes 1–3 (r for apple/hawthorn difference = −0.831, $p < 0.0001$; r for *R. mendax*/hawthorn difference = −0.632, $p < 0.0001$; r for *R. mendax*/apple difference = −0.373, $p < 0.0001$, $n = 508$ SNPs in all cases; figure 3, orange circles; electronic supplementary material, table S2).

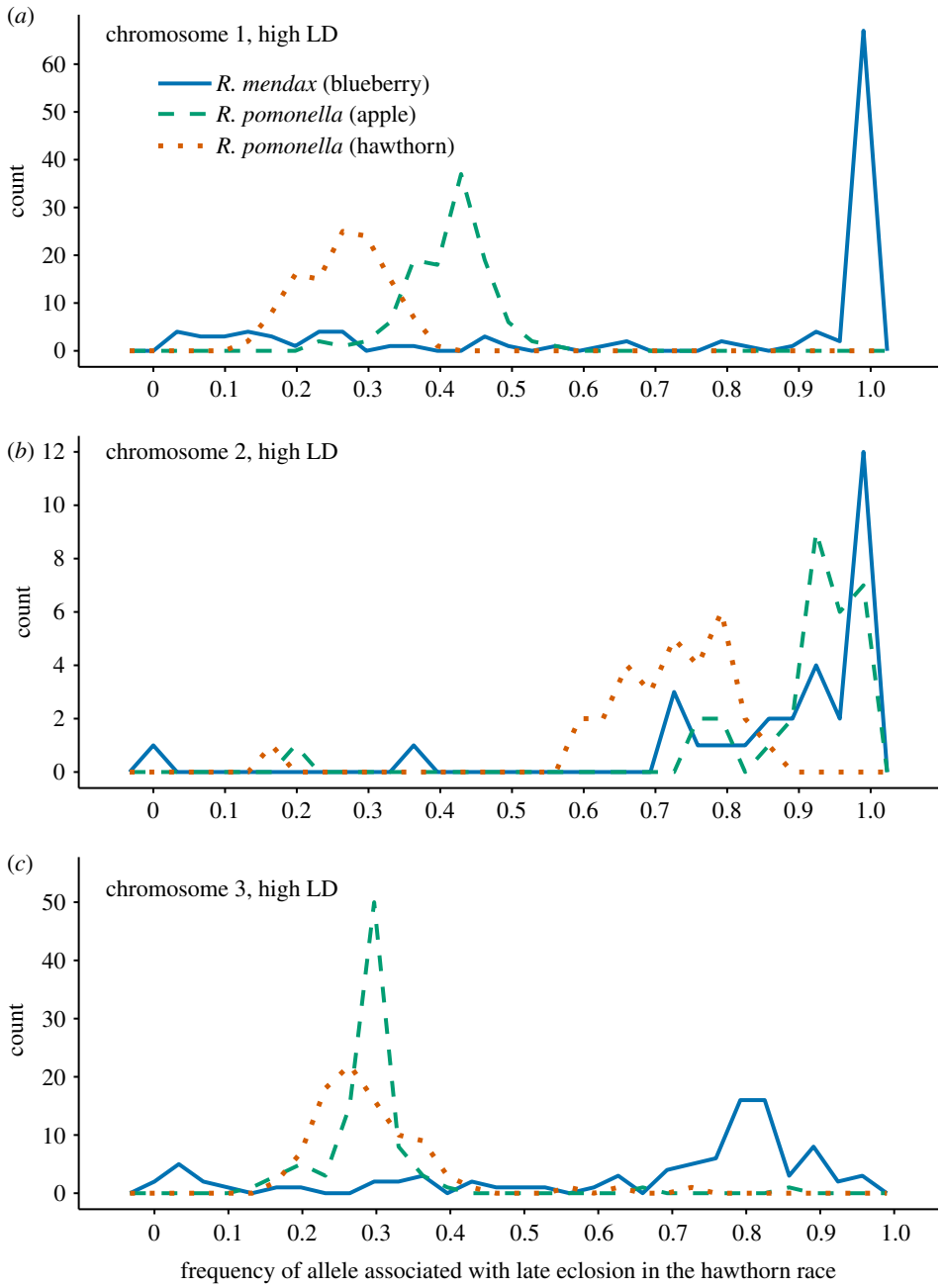

Figure 6. (*a–c*) Allele frequency distributions in *R. pomonella* apple (green, dashed lines) and hawthorn (orange, dotted lines) host races and *R. mendax* (blue, solid lines) for SNPs in the high LD classes on chromosomes 1–3 ($n = 113$, 30 and 91 SNPs, respectively), representing putative inversions. Loci were assigned to the high LD class if both designated as such by Ragland *et al*. [26,49] for *R. pomonella* from Grant, MI and in the high LD cluster for the *R. pomonella* hawthorn race at Fennville, MI (figure 5*a–c*). Frequencies were calculated for the allele associated with later adult eclosion time in the hawthorn race; distributions are shown for allele frequency bin widths of 0.033. Of note is the general increase in allele frequencies for these high LD SNPs and, by association, the putative inversions they represent, related to later adult eclosion time from hawthorn to apple to blueberry flies. Many 'late eclosion' alleles on chromosomes 1 and 2 appear fixed or near fixed in frequency at 1.0 in *R. mendax*, suggesting a loss of or low-frequency inversion polymorphism for these two chromosomes in the blueberry-infesting population. (Online version in colour.)

Together, the significant correlations between eclosion association and host race and/or species divergence (figure 3) and the concentration of eclosion variation in the high LD groups on chromosomes 1–3 (figures 3 and 5) suggest that a significant portion of the divergence among taxa may be explained by changes in the frequencies of putative inversion haplotypes on chromosomes 1–3. Frequency distributions demonstrate that late eclosion-associated alleles in the high LD classes on chromosomes 1–3 are increasingly more common from hawthorn, to apple, to blueberry fly populations (figure 6*a–c*). Thus, it can be inferred that on each of chromosomes 1–3, the inversion haplotype associated with later adult eclosion in the hawthorn race GWAS and, by pleiotropy or linkage increased initial diapause intensity [27], also increases in frequency from hawthorn race, to apple race, to *R. mendax*.

The LD networks and frequency distributions suggest that, for both chromosomes 1 and 2, only one of the inversion haplotypes segregating in *R. pomonella* is present in *R. mendax*, the other(s) being rare or absent (figures 5 and 6; electronic supplementary material, figure S3). For both chromosomes 1 and 2, the frequencies of high LD 'late eclosion' alleles in hawthorn flies were fixed at 1.0 or near so in the *R. mendax* population, but variable in the hawthorn and apple host races (figure 6*a*,*b*). This suggests the fixation of the inversion haplotype harbouring 'late eclosion' alleles

during or following the host shift from ancestral hawthorn to blueberry, allowing subsequent free recombination on chromosomes 1 and 2 in *R. mendax*. This is supported by a reduction in the numbers of SNPs in high LD clusters displaying r^2 values of greater than 0.8 with one another in *R. mendax* compared to the hawthorn race (figure 5; electronic supplementary material, figure S3). In comparison, *R. mendax* has retained some eclosion variation on chromosome 3; the high LD 'late eclosion' alleles and thus the associated inversion haplotype appear to be segregating at a higher frequency (approx. 0.80) in *R. mendax*, compared to hawthorn and apple host races (approx. 0.25 and approx. 0.30, respectively; figure 6*c*). Therefore, for chromosome 3, the LD networks for both the hawthorn race and *R. mendax* suggest the presence of an inversion polymorphism (figure 5*c*,*m*), composed of corresponding loci (electronic supplementary material, figure S3*c*,*h*).

(d) *Rhagoletis mendax* eclosion time genomics differ from those of *Rhagoletis pomonella*

The near fixation in *R. mendax* of many alleles associated with later eclosion in the ancestral *R. pomonella* suggests that the genomics of eclosion time may differ substantially between *R. pomonella* and *R. mendax*. The paradoxically earlier eclosion phenotypes of *R. mendax* also imply that this species harbours compensatory changes allowing for earlier eclosion. Therefore, an additional GWAS was conducted for *R. mendax*. In this eclosion time GWAS, the 95% CI for PVE for all SNPs was 0.710–0.999, centred on 0.964. The median PGE was estimated as 0.622, with a 95% CI from 0.068 to 0.982. The 95% CI for the number of loci of measurable effect contributing to diapause termination time ranged from 2 to 216 (median = 35). Similar to the hawthorn race, the results for *R. mendax* suggested that the eclosion phenotype is also highly polygenic in blueberry flies.

Aside from being polygenic, the genomic architecture of eclosion time largely differed between *R. mendax* and the hawthorn race, as predicted between taxa whose divergence is relatively far along the speciation continuum (table 1). For *R. mendax*, SNPs significantly associated with eclosion time were distributed relatively evenly across the genome and were not concentrated on chromosomes 1–3 (figure 5*f–j*). As a result, allele frequency differences between early and late eclosing flies in the *R. mendax* GWAS were not highly correlated with those for the hawthorn race, although the relationship was still significant ($r = 0.076$, $p < 0.0001$, $n = 24\,954$ SNPs variable in both the hawthorn race and *R. mendax* studies; electronic supplementary material, table S1). Loci mapping to chromosome 3 showed the highest correlation in eclosion time between *R. mendax* and the hawthorn race ($r = 0.272$, $p < 0.0001$, $n = 501$ SNPs; electronic supplementary material, table S1), probably because of the variation remaining in the segregating inversion polymorphism described above (figure 6*c*; electronic supplementary material, figure S3).

Unlike the results for hawthorn flies, the SNP allele frequency differences in the *R. mendax* eclosion time GWAS were not strongly predictive of genome-wide allele frequency differentiation between the host races, although the relationship was significant ($r = -0.034$, $p < 0.0001$, $n = 32\,690$ SNPs variable in *R. mendax*). Genomic correlations were also weak, but significant, between eclosion time variation in *R. mendax* and divergence between *R. mendax* and the hawthorn ($r = 0.088$, $p < 0.0001$, $n = 32\,690$ SNPs) or apple ($r = 0.095$, $p < 0.0001$, $n = 32\,690$ SNPs) host races (electronic supplementary material, table S3). In contrast with the correlations involving the hawthorn race GWAS, allele frequency differences in the *R. mendax* eclosion time GWAS were positively associated with population divergence between *R. mendax* and both host races (electronic supplementary material, table S3). Thus, *R. mendax*, the taxon with the earliest eclosion phenotype, tended to possess higher frequencies of alleles associated with earlier adult eclosion in *R. mendax* compared to both of the host races. This pattern is graphically illustrated in figure 4*b*, in which the blueberry, apple and hawthorn fly populations are projected onto a discriminant function trained on the early and late eclosing samples of *R. mendax* flies. These eclosion-associated SNPs or linked loci could potentially explain why blueberry flies eclose earlier than both apple and hawthorn flies. As predicted, they may compensate for the maladaptive consequences of selection for greater initial diapause intensity in *R. mendax*, which pulls along alleles for later adult eclosion associated with the inversions on chromosomes 1–3 [27].

4. Discussion

Here, we present evidence implying that a degree of predictability exists in the genomics of speciation (table 1). Previous studies have shown that natural selection can move populations along anticipated phenotypic and/or genetic paths, while others have attested to the repeated occurrence of convergent or parallel evolution among taxa experiencing similar selection pressures (see [3,4] for review). What distinguishes the *Rhagoletis* finding here from earlier work is that it bridges short-term deterministic tracking of the environment by populations undergoing ecological speciation with differences between more deeply diverged species. Specifically, we found that standing variation associated with diapause life-history timing in the ancestral hawthorn-infesting *R. pomonella* was predictive of genome-wide population divergence across the speciation continuum, not just for newly formed partially reproductively isolated host races, but also more distantly related sibling species (table 1). This predictability probably reflects similar selection pressures on *R. mendax* and the apple race for increased initial diapause intensity, owing to the earlier fruiting times of blueberries and apples compared to hawthorns (table 1). This was evident in the correspondence between the order of fruiting times (late to early) and frequencies of alleles associated with late eclosion and deep diapause (low to high) from hawthorn to apple to blueberry fly populations (figures 4*a* and 6). As anticipated, this predictability was driven primarily by changes in putative inversion frequencies on chromosomes 1–3, which are most strongly associated with diapause timing in *R. pomonella* [26,27].

The degree to which standing variation in the extant ancestral population was predictive of population divergence decreased with the time of separation and reduction of gene flow between taxa, as did the correlation in the genetic architecture of diapause phenotypes (table 1). Nevertheless, we identified compensatory variants in *R. mendax* that may account for the earlier adult eclosion of blueberry flies and counter the maladaptive consequences of selection for greater initial diapause intensity [27] (table 1). As hypothesized, these variants were not concentrated in the inversions on

chromosomes 1–3, for which blueberry flies harboured very high frequencies of alleles associated with later eclosion in *R. pomonella*. If such compensatory changes are generally required to counteract negative pleiotropic or epistatic effects of selection, however, then they will complicate and may limit the ability of GWAS and selection experiments on ancestral populations to predict the genomics of speciation. Indeed, adaptive evolution and speciation may follow paths of least resistance owing to pleiotropy, genetic and environmental interactions, or linkage relationships; these all may run counter to the most direct course selection could take to reach an optimal fitness peak [13]. In the case of *Rhagoletis*, we were *a priori* aware of the need for compensatory variants for earlier adult eclosion to counter the historical relationship in ancestral standing inversion polymorphism in the hawthorn race [26,27,45]. However, in the absence of such knowledge, our predictions for genomic change would have been opposite of what was observed, calling into account our forecasting ability.

Moreover, in the *R. mendax* GWAS, the alleles for earlier eclosion were not strongly associated with standing variation in the hawthorn race, except perhaps for chromosome 3, and they could not explain why apple flies eclose earlier than hawthorn flies. Therefore, these variants may represent new mutations in *R. mendax*. However, we cannot rule out that these variants are either rare or restricted to southern latitudes in *R. pomonella* and, thus, were not detected in the hawthorn race GWAS conducted at Fennville. Additionally, many SNPs in lower LD regions of the genome may not maintain linkage with causative eclosion variants in both *R. pomonella* and *R. mendax*, further reducing the chances of detecting similarities with our ddRAD-seq approach. Regardless of their origin, the genomic signature of selection for earlier eclosion in blueberry flies may not have been deduced based on information from the hawthorn race GWAS alone. Moreover, the compensatory changes in the apple race that account for its early eclosion time remain unresolved, although presently they would appear to differ from those for *R. mendax*. Thus, while we may attest to some success, not all of the adaptive changes between the host races and *R. mendax* could be predicted *a priori*.

There are a number of additional caveats concerning the current study that also require further elaboration, several of which bear generally on the topic of predicting the genomics of speciation. First, although our findings show that a predictable pattern of adaptive divergence may be detected deep into speciation, our results reflect only two data points, early and relatively late in the process. Future studies of additional taxa in the *R. pomonella* group are needed to fill in stages between the host races and *R. mendax*. Of particular interest in this regard will be the undescribed sister taxon of *R. pomonella* that attacks flowering dogwood, *C. florida*, which fruits later than hawthorn at sympatric sites in the midwestern USA [21].

Second, we focused on the genomics of a particular phenotype, diapause life-history variation, well characterized in *Rhagoletis* [26,27,40,44] that pleiotropically generates ecologically based allochronic RI (i.e. diapause is a 'magic' trait; [61]). This allowed us to employ a strategy comparing a GWAS for eclosion timing (essentially a laboratory selection experiment on diapause) with a population survey between sympatric taxa (essentially a comparative field study of different stages of the speciation process). Tests of the genomic predictability of speciation using such an approach may not be possible for all systems and phenotypes, especially in cases of complete allopatry, or when RI is not associated with divergent selection on magic traits. However, we note that such a strategy may still be useful in situations of secondary contact, where gene flow among taxa has occurred repeatedly at different locations or times. In such circumstances, predictions could be made and tested concerning the degree to which patterns of geographical or genomic clines coincide among independent instances of secondary contact (e.g. see [62]).

Third, the genomics of diapause life-history variation are conducive to testing for evolutionary predictability. Diapause timing is a highly polygenic trait in *Rhagoletis* and many attributes postulated to facilitate ecological speciation and adaptive radiation, including large stores of standing variation, hybridization and inversion polymorphism, are also associated with diapause in the presumed ancestral hawthorn population of *R. pomonella* [17,18]. Thus, in the current study, a substantial portion of the divergence of *R. mendax* from hawthorn flies appears to involve differences in the frequencies of ancestral inversion polymorphisms affecting diapause in *R. pomonella* (figure 6*a–c*). However, adaptive variation need not always be standing and associated with inversions in all systems, making it harder to predict and track evolutionary trajectories.

A final important point is that phenology presents a linear, continuous axis—so that variation in life-history timing might be maintained and diverge along that same dimension across populations or taxa. For example, in *Rhagoletis*, regardless of host, if the plant fruits earlier in the year, like blueberries and apples, then flies will experience similar selection pressures for deeper initial diapause intensity and earlier adult eclosion. However, this may not be true for other traits involved in ecological adaptation in *Rhagoletis* and for divergence in other systems, although it is possible to envision how our approach may apply more broadly to certain phenotypes (e.g. body size, colour and desiccation tolerance). *Rhagoletis* host races and species are also ecologically differentiated by variation in host preference in adults [28,63–65] and, in some cases, by variation in feeding performance in host fruit as larvae [66,67]. In contrast with fruiting phenology, blueberries and apples may differ in their chemical compositions and physical characteristics in ways that exert contrasting selection pressures on preference behaviour and feeding performance that are host specific and affected by different suites of genes in different taxa. As a result, predicting the genomics of species divergence from standing variation in ancestral populations may be more complicated for host choice and larval feeding performance than for diapause.

In conclusion, ecological speciation can connect the microevolutionary process of natural selection to macroevolutionary pattern. Here, we show that the genomes of *Rhagoletis* flies follow predicable evolutionary trajectories across the speciation continuum from host races to species. How general our findings are for other systems remains to be seen. For example, if most speciation occurs non-ecologically via a mutation order process involving intrinsic RI [14,68] and character displacement is rare, then it may be more difficult to anticipate the exact details of the genomics of population divergence. But certainly, ecological speciation is not uncommon [14]. Moreover, other actively diversifying groups share attributes akin to *Rhagoletis* flies, including large stores of standing variation, chromosomal rearrangements and a history of hybridization [17,18,69,70], making it likely that there are predictable components to the genomics of their radiations,

as well. Thus, in the grand scheme of things, life is not all due to chance.

Data accessibility. The *R. mendax* data new to this publication can be found on the Dryad Digital Repository: https://dx.doi.org/10.5061/dryad.nk98sf7pr [48], as well as the *R. pomonella* apple and hawthorn race population (https://dx.doi.org/10.5061/dryad.k42t7g2) [47] and early and late eclosing samples (https://dx.doi.org/10.5061/dryad.kn568) [49].

Authors' contributions. P.J.M. designed the study, carried out molecular laboratory work and critically revised the manuscript; M.M.D. carried out molecular laboratory work, analysed the data, helped draft the manuscript and critically revised the manuscript; G.J.R. conceived of the study, designed the study and critically revised the manuscript; G.R.H. conducted field collections, carried out the organismal laboratory work and critically revised the manuscript; S.P.E. conceived of the study and critically revised the manuscript; T.H.Q.P. conducted field collections, carried out the organismal laboratory work and critically revised the manuscript; P.N. conceived of the study and critically revised the manuscript; J.L.F. conceived of the study, designed the study, drafted the manuscript and critically revised the manuscript. All authors gave final approval for publication and agree to be held accountable for the work performed therein.

Competing interests. We declare we have no competing interests.

Funding. This work was supported by University of Notre Dame Advanced Diagnostics & Therapeutics Initiative (S.P.E., J.L.F.) and Environmental Change Initiative (G.J.R., S.P.E., J.L.F.), NSF grants to G.J.R., S.P.E. and J.L.F. and USDA grants to S.P.E. and J.L.F. P.N. was supported by a European Research Council Consolidator Grant.

Acknowledgements. The authors would like to thank James J. Smith, McCall Calvert, John Wise and the Trevor Nichols Research Station of Michigan State University for their help and assistance.

References

1. Gould SJ. 1989 *Wonderful life*. London, UK: Norton.
2. Gould SJ. 2002 *The structure of evolutionary theory*. Cambridge, MA: Belknap Press.
3. Blount ZD, Lenski RE, Losos JB. 2018 Contingency and determinism in evolution: replaying life's tape. *Science* **362**, eaam5979. (doi:10.1126/SCIENCE.AAM5979)
4. Losos JB. 2017 *Improbable destinies: fate, chance, and the future of evolution*. New York, NY: Riverhead Books.
5. Gould SJ. 1994 The evolution of life on the earth. *Sci. Am.* **271**, 84–91. (doi:10.1038/scientificamerican1094-84)
6. Barrett RDH, Schluter D. 2008 Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**, 38–44. (doi:10.1016/j.tree.2007.09.008)
7. Whitlock MC, Phillips PC, Moore FB, Tonsor SJ. 1995 Multiple fitness peaks and epistasis. *Annu. Rev. Ecol. Syst.* **26**, 601–629. (doi:10.1146/annurev.es.26.110195.003125)
8. de Visser JAGM, Krug J. 2014 Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* **15**, 480–490. (doi:10.1038/nrg3744)
9. Blount ZD, Borland CZ, Lenski RE. 2008 Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **105**, 7899–7906. (doi:10.1073/pnas.0803151105)
10. Lenormand T, Roze D, Rousset F. 2009 Stochasticity in evolution. *Trends Ecol. Evol.* **24**, 157–165. (doi:10.1016/j.tree.2008.09.014)
11. Morris SC. 2003 *Life's solution: inevitable humans in a lonely universe*. Cambridge, UK: Cambridge University Press.
12. McGhee G. 2011 *Convergent evolution: limited forms most beautiful*. Cambridge, MA: The MIT Press.
13. Schluter D. 2000 *The ecology of adaptive radiation*. Oxford, UK: Oxford University Press.
14. Schluter D. 2009 Evidence for ecological speciation and its alternative. *Science* **323**, 737–741. (doi:10.1126/science.1160006)
15. Kirkpatrick M, Barton N. 2006 Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434. (doi:10.1534/genetics.105.047985)
16. Feder JL, Nosil P, Flaxman SM. 2014 Assessing when chromosomal rearrangements affect the dynamics of speciation: implications from computer simulations. *Front. Genet.* **5**, 295. (doi:10.3389/fgene.2014.00295)
17. Berner D, Salzburger W. 2015 The genomics of organismal diversification illuminated by adaptive radiations. *Trends Genet.* **31**, 491–499. (doi:10.1016/j.tig.2015.07.002)
18. Marques DA, Meier JI, Seehausen O. 2019 A combinatorial view on speciation and adaptive radiation. *Trends Ecol. Evol.* **34**, 531–554. (doi:10.1016/j.tree.2019.02.008)
19. Feder JL, Egan SP, Nosil P. 2012 The genomics of speciation-with-gene-flow. *Trends Genet.* **28**, 342–350. (doi:10.1016/j.tig.2012.03.009)
20. Bush G. 1966 The taxonomy, cytology, and evolution of the genus *Rhagoletis* in North America (Diptera, Tephritidae). *Bull. Mus. Comp. Zool.* **134**, 431–562.
21. Berlocher SH. 2000 Radiation and divergence in the *Rhagoletis pomonella* species group: inferences from allozymes. *Evolution* **54**, 543–557. (doi:10.1111/j.0014-3820.2000.tb00057.x)
22. Xie X, Michel AP, Schwarz D, Rull J, Velez S, Forbes AA, Aluja M, Feder JL. 2008 Radiation and divergence in the *Rhagoletis pomonella* species complex: inferences from DNA sequence data. *J. Evol. Biol.* **21**, 900–913. (doi:10.1111/j.1420-9101.2008.01507.x)
23. Feder JL, Chilcote CA, Bush GL. 1988 Genetic differentiation between sympatric host races of the apple maggot fly *Rhagoletis pomonella*. *Nature* **336**, 61–64. (doi:10.1038/336061a0)
24. McPheron BA, Smith DC, Berlocher SH. 1988 Genetic differences between host races of *Rhagoletis pomonella*. *Nature* **336**, 64–66. (doi:10.1038/336064a0)
25. Feder JL, Filchak KE. 1999 It's about time: the evidence for host plant-mediated selection in the apple maggot fly, *Rhagoletis pomonella*, and its implications for fitness trade-offs in phytophagous insects. *Entomol. Exp. Appl.* **91**, 211–225. (doi:10.1046/j.1570-7458.1999.00486.x)
26. Ragland GJ, Doellman MM, Meyers PJ, Hood GR, Egan SP, Powell THQ, Hahn DA, Nosil P, Feder JL. 2017 A test of genomic modularity among life-history adaptations promoting speciation with gene flow. *Mol. Ecol.* **26**, 3926–3942. (doi:10.1111/mec.14178)
27. Calvert MB *et al*. In press. The genomics of trait combinations and their influence on adaptive divergence. *J. Evol. Biol.*
28. Feder JL, Bush GL. 1989 A field test of differential host-plant usage between two sibling species of *Rhagoletis pomonella* fruit flies (Diptera: Tephritidae) and its consequences for sympatric models of speciation. *Evolution* **43**, 1813–1819. (doi:10.1111/j.1558-5646.1989.tb02632.x)
29. Feder J, Chilcote C, Bush G. 1989 Are the apple maggot, *Rhagoletis pomonella*, and blueberry maggot, *Rhagoletis mendax*, distinct species—implications for sympatric speciation. *Entomol. Exp. Appl.* **51**, 113–123. (doi:10.1111/j.1570-7458.1989.tb01221.x)
30. Dean RW, Chapman PJ. 1973 *Bionomics of the apple maggot in eastern New York*. Ithaca, NY: Cornell University.
31. Dambroski HR, Feder JL. 2007 Host plant and latitude-related diapause variation in *Rhagoletis pomonella*: a test for multifaceted life history adaptation on different stages of diapause development. *J. Evol. Biol.* **20**, 2101–2112. (doi:10.1111/j.1420-9101.2007.01435.x)
32. Lyons-Sobaski S, Berlocher SH. 2009 Life history phenology differences between southern and northern populations of the apple maggot fly, *Rhagoletis pomonella*. *Entomol. Exp. Appl.* **130**, 149–159. (doi:10.1111/j.1570-7458.2008.00805.x)
33. Hood GR, Forbes AA, Powell THQ, Egan SP, Hamerlinck G, Smith JJ, Feder JL. 2015 Sequential divergence and the multiplicative origin of community diversity. *Proc. Natl Acad. Sci. USA* **112**, E5980–E5989. (doi:10.1073/pnas.1424717112)
34. Powell THQ, Forbes AA, Hood GR, Feder JL. 2014 Ecological adaptation and reproductive isolation in

sympatry: genetic and phenotypic evidence for native host races of *Rhagoletis pomonella*. *Mol. Ecol.* **23**, 688–704. (doi:10.1111/mec.12635)

35. Feder J, Hunt T, Bush G. 1993 The effects of climate, host-plant phenology and host fidelity on the genetics of apple and hawthorn infesting races of *Rhagoletis pomonella*. *Entomol. Exp. Appl.* **69**, 117–135. (doi:10.1111/j.1570-7458.1993.tb01735.x)
36. Taylor RS, Friesen VL. 2017 The role of allochrony in speciation. *Mol. Ecol.* **26**, 3330–3342. (doi:10.1111/mec.14126)
37. Servedio MR, Van Doorn GS, Kopp M, Frame AM, Nosil P. 2011 Magic traits in speciation: 'magic' but not rare? *Trends Ecol. Evol.* **26**, 389–397. (doi:10.1016/J.TREE.2011.04.005)
38. Powell THQ, Nguyen A, Xia Q, Feder JL, Ragland GJ, Hahn DA. 2020 A rapidly evolved shift in life history timing during ecological speciation is driven by the transition between developmental phases. *bioRxiv*, 925057. (doi:10.1101/2020.01.29.925057)
39. Feder JL, Roethele JB, Wlazlo B, Berlocher SH. 1997 Selective maintenance of allozyme differences among sympatric host races of the apple maggot fly. *Proc. Natl Acad. Sci. USA* **94**, 11 417–11 421. (doi:10.1073/pnas.94.21.11417)
40. Egan SP, Ragland GJ, Assour L, Powell THQ, Hood GR, Emrich S, Nosil P, Feder JL. 2015 Experimental evidence of genome-wide impact of ecological selection during early stages of speciation-with-gene-flow. *Ecol. Lett.* **18**, 817–825. (doi:10.1111/ele.12460)
41. Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P, Feder JL. 2010 Widespread genomic divergence during sympatric speciation. *Proc. Natl Acad. Sci. USA* **107**, 9724–9729. (doi:10.1073/pnas.1000939107)
42. Feder JL, Roethele JB, Filchak K, Niedbalski J, Romero-Severson J. 2003 Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly, *Rhagoletis pomonella*. *Genetics* **163**, 939–953.
43. Sturtevant AH. 1921 A case of rearrangement of genes in *Drosophila*. *Proc. Natl Acad. Sci. USA* **7**, 235–237. (doi:10.1073/pnas.7.8.235)
44. Doellman MM *et al*. 2018 Genomic differentiation during speciation-with-gene-flow: comparing geographic and host-related variation in divergent life history adaptation in *Rhagoletis pomonella*. *Genes* (*Basel*) **9**, 262. (doi:10.3390/GENES9050262)
45. Doellman MM *et al*. 2019 Standing geographic variation in eclosion time and the genomics of host race formation in *Rhagoletis pomonella* fruit flies. *Ecol. Evol.* **9**, 393–409. (doi:10.1002/ece3.4758)
46. Filchak KE, Roethele JB, Feder JL. 2000 Natural selection and sympatric divergence in the apple maggot *Rhagoletis pomonella*. *Nature* **407**, 739–742. (doi:10.1038/35037578)
47. Doellman MM *et al*. 2018 Data from: Standing geographic variation in eclosion time and the genomics of host race formation in *Rhagoletis pomonella* fruit flies. Dryad Digital Repository. (https://doi.org/10.5061/dryad.k42t7g2)
48. Meyers PJ, Doellman MM, Ragland GJ, Hood GR, Egan SP, Powell THQ, Nosil P, Feder JL. 2020 Data from: Can the genomics of ecological speciation be predicted across the divergence continuum from host races to species? A case study in *Rhagoletis*. Dryad Digital Repository. (https://doi.org/10.5061/dryad.nk98sf7pr)
49. Ragland GJ, Doellman MM, Meyers PJ, Hood GR, Egan SP, Powell THQ, Hahn DA, Nosil P, Feder JL. 2017 Data from: A test of genomic modularity among life-history adaptations promoting speciation with gene flow. Dryad Digital Repository. (https://doi.org/10.5061/dryad.kn568)
50. Lischer HEL, Excoffier L. 2012 PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298–299. (doi:10.1093/bioinformatics/btr642)
51. Jombart T, Ahmed I. 2011 *adegenet 1.3-1*: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071. (doi:10.1093/bioinformatics/btr521)
52. Jombart T. 2008 *adegenet*: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405. (doi:10.1093/bioinformatics/btn129)
53. Pritchard JK, Stephens M, Donnelly P. 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.
54. Raj A, Stephens M, Pritchard JK. 2014 fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589. (doi:10.1534/genetics.114.164350)
55. Francis RM. 2017 pophelper: an R package and web app to analyse and visualize population structure. *Mol. Ecol. Resour.* **17**, 27–32. (doi:10.1111/1755-0998.12509)
56. Bilton TP, McEwan JC, Clarke SM, Brauning R, van Stijn TC, Rowe SJ, Dodds KG. 2018 Linkage disequilibrium estimation in low coverage high-throughput sequencing data. *Genetics* **209**, 389–400. (doi:10.1534/genetics.118.300831)
57. Csardi G, Nepusz T. 2006 The igraph software package for complex network research. *Int. J. Complex Syst.* **1695**. See https://cran.r-project.org/web/packages/igraph/cit.
58. Kemppainen P, Knight CG, Sarma DK, Hlaing T, Prakash A, Maung Maung YN, Somboon P, Mahanta J, Walton C. 2015 Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. *Mol. Ecol. Resour.* **15**, 1031–1045. (doi:10.1111/1755-0998.12369)
59. Zhou X, Carbonetto P, Stephens M. 2013 Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264. (doi:10.1371/journal.pgen.1003264)
60. Jombart T, Devillard S, Balloux F. 2010 Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94. (doi:10.1186/1471-2156-11-94)
61. Gavrilets S. 2004 *Fitness landscapes and the origin of species*. Princeton, NJ: Princeton University Press.
62. Chaturvedi S, Lucas LK, Buerkle CA, Fordyce JA, Forister ML, Nice CC, Gompert Z. 2020 Recent hybrids recapitulate ancient hybrid outcomes. *Nat. Commun.* **11**, 2179. (doi:10.1038/s41467-020-15641-x)
63. Feder JL, Opp SB, Wlazlo B, Reynolds K, Go W, Spisak S. 1994 Host fidelity is an effective premating barrier between sympatric races of the apple maggot fly. *Proc. Natl Acad. Sci. USA* **91**, 7990–7994. (doi:10.1073/pnas.91.17.7990)
64. Linn C, Feder JL, Nojima S, Dambroski HR, Berlocher SH, Roelofs W. 2003 Fruit odor discrimination and sympatric host race formation in *Rhagoletis*. *Proc. Natl Acad. Sci. USA* **100**, 11 490–11 493. (doi:10.1073/pnas.1635049100)
65. Forbes AA, Fisher J, Feder JL. 2005 Habitat avoidance: overlooking an important aspect of host-specific mating and sympatric speciation? *Evolution* **59**, 1552–1559. (doi:10.1111/j.0014-3820.2005.tb01804.x)
66. Bierbaum TJ, Bush GL. 1990 Genetic differentiation in the viability of sibling species of *Rhagoletis* fruit flies on host plants, and the influence of reduced hybrid viability on reproductive isolation. *Entomol. Exp. Appl.* **55**, 105–118. (doi:10.1111/j.1570-7458.1990.tb01353.x)
67. Ragland GJ, Almskaar K, Vertacnik KL, Gough HM, Feder JL, Hahn DA, Schwarz D. 2015 Differences in performance and transcriptome-wide gene expression associated with *Rhagoletis* (Diptera: Tephritidae) larvae feeding in alternate host fruit environments. *Mol. Ecol.* **24**, 2759–2776. (doi:10.1111/mec.13191)
68. Nosil P. 2012 *Ecological speciation*. New York, NY: Oxford University Press.
69. Arnold ML. 1997 *Natural hybridization and evolution*. New York, NY: Oxford University Press.
70. Mallet J. 2008 Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. *Phil. Trans. R. Soc. B* **363**, 2971–2986. (doi:10.1098/rstb.2008.0081)