The genomic landscape of adaptation to a new host plant

Authors

Kalle J Nilsson* 1

Rachel A Steward* 1

Jesús Ortega Giménez* 1,2

Zachary J Nolen 1

Chao Yan 1

Yajuan Huang 1

Julio Ayala López 3

Anna Runemark 1

* Equal contributions

Affiliations:

1 Biology Department, Lund University, Lund, Sweden

2 Cavanilles Institute of Biodiversity and Evolutionary Biology, Universidad de Valencia, Paterna, Spain

3 Division of Theoretical Ecology and Evolution, Universität Bern, Bern, Switzerland

Abstract

Adaptation to novel ecological niches is known to be rapid. However, how ecological divergence translates into reproductive isolation remains a consequential question in speciation research. It is still unclear how coupling of ecological divergence loci and reproductive isolation loci occurs at the genomic level, ultimately enabling the formation of persistent species. Here, we investigate the genomic underpinning colonization of a new niche and formation of two host races in *Tephritis conura* peacock flies that persist in sympatry. To uncover the genomic differences underlying host plant adaptation, we take advantage of two independent sympatric zones, where host plant specialists using the thistle species *Cirsium heterophyllum* and *C. oleraceum* co-occur, and address what regions of the genome that diverge between the host races in a parallel fashion. Using Population Branch Statistics, dxy and BayPass we identify 2996 and 3107 outlier regions associated with host use among western and eastern Baltic populations, respectively, with 82% overlap. The majority of outliers are located within putative inversions, adding to the growing body of evidence that structural changes to the genome are important for adaptations to persist in face of gene flow. Potentially, the high contents of repetitive elements could have given rise to the inversions, but this remains to be tested. The outlier regions were enriched for genes involved in e.g. metabolism and morphogenesis, potentially due to the selection for processing different metabolites, and changes in size and ovipositor length respectively. In conclusion, this study suggests that structural changes in the genome, and divergence in independent ecological functions may facilitate the formation of persistent host races in face of gene flow.

Introduction

How biological variation and novel adaptation arise are fundamental and unresolved questions in our understanding of the origins of biodiversity. Ecological adaptation to novel environments and food sources are known to result in rapid phenotypic and genetic divergence (Bush 1969; Boag and Grant 1981; Feder et al. 1988; Filchak et al. 2000; Herrel et al. 2008; Nosil 2012), but to what extent this ecological adaptation leads to persistent speciation is contested (Schluter 2009; Anderson and Weir 2022; Anderson et al. 2023). Both the ecological context, including the consistency (Hendry 2009; Bolnick 2011) and multidimensionality of the divergent niches (Rice and Hostert 1993; Nosil and Sandoval 2008; Nosil et al. 2009; Chevin et al. 2014), and the genomic underpinnings of adaptation to those niches (Chevin et al. 2014) may influence the likelihood of divergent lineages becoming distinct species over time. Whether ecological adaptations and reproductive isolation are linked, at either the phenotypic or genomic scale, is important to the speciation process. Without reproductive isolation, ecological shifts can lead to collapses of formerly isolated population, resulting in increased gene flow and breakdown of the speciation process (Gow et al. 2006; Taylor et al. 2006; Lackey and Boughman 2017). Thus, a challenge for evolutionary biologists is to identify the genomic architecture that enables persistent species to form (Kulmuni et al. 2020).

One way to uncover the genomic architecture that enables speciation is to assess sequence differences that are robust to gene flow between populations in sympatry. Gene flow can be limited in certain regions of the genome by the presence of genes that are involved in reproductive isolation, and differences in these regions are likely contributing to the speciation process (Schluter

64 and Rieseberg 2022). However, interpreting which of the differences between existing species
65 caused speciation may be challenging, as additional genes involved in reproductive isolation may
66 have accumulated since the species diverged. Investigating the genomic differences between
67 recently diverged taxa and focusing on independent contact zones can help resolve this challenge
68 (Schluter and Rieseberg 2022). Regions that repeatedly differ among differentially-adapted taxa
69 are likely to reflect regions that have played a role in speciation (Bohutínská et al. 2021). Extensive
70 theoretical and empirical evidence has shown that genomic architecture coupling coadapted loci,
71 including inversions, is particularly important for persistence of differential adaptation in the face
72 of gene flow (Feder et al. 2003; Wellenreuther and Bernatchez 2018; Berdan et al. 2022; Schaal et
73 al. 2022). Yet, the extent to which linkage of ecological adaptation and reproductive isolation loci
74 is facilitated by pleiotropic effects of the same genes, coupling of genes through structural
75 variation, or co-inheritance of uncoupled genomic regions upheld by correlational selection is not
76 well understood.

77

78 Herbivorous insects and the plants they interact with are excellent models for studying the genomic
79 basis of speciation. These are two of the most speciose groups of eukaryotes (Mora et al. 2011;
80 Christenhusz and Byng 2016), and specialization of phytophagous insects and their hosts is
81 hypothesized to drive this divergence (Vidal and Murphy 2018). The specificity and
82 multidimensionality of the niches constituted by the host plants (Hardy and Otto 2014) and strong
83 dependency on host plants for reproduction and survival (e.g., assortative mating based on host
84 recognition; Bush 1969) may increase the chance of ecological speciation in herbivorous insects
85 (Hardy and Otto 2014). These features make phytophagous insects effective systems for
86 understanding the genomic basis of ecological adaptation. Changes in host repertoire, either

87 through host range expansion or host shifts, have previously been associated with chemosensory
88 genes underlying butterfly host choice (van Schooten et al. 2020), alter oral secretions in aphids
89 (Boulain et al. 2019; Shih et al. 2023), and code for enzymes in the digestive systems of both
90 butterflies and aphids (Nallu et al. 2018; Singh et al. 2020; Shih et al. 2023), and genes underlying
91 phenological shifts in *Rhagoletis* flies (Feder et al. 1988; Feder et al. 2003). Ecological divergence
92 resulting in reduced gene flow among sympatric or parapatric populations of herbivorous insects
93 causes in the formation of host races (Berlocher and Feder 2002). However, how the genetic
94 changes underlying ecological adaptation are coupled to or translated into reproductive isolation
95 at the genomic level, enabling co-existence and long term persistence of differentially adaptive
96 host races, remains a challenge to resolve. One possibility is that the rate at which differentially
97 adapted host races accumulate the genetic architecture enabling coupling or strong reproductive
98 isolation could determine the time required for stable co-existence. This is the case in birds, where
99 sister taxa with inversions co-exist at shorter divergence times (Hooper et al. 2019).
100
101 Here, we use the Tephritid fly *Tephritis conura* to uncover the genomic basis of ecological
102 divergence and reproductive isolation promoting stable coexistence of two differentially adapted
103 host races. These host races infest the ancestral host thistle *Cirsium heterophyllum* and the derived
104 host thistle *C. oleraceum*, respectively. They differ in phenology and host plant preference
105 (Romstock-Volkl 1997), and have host-specific larval performance and survival (Diegisser et al.
106 2008) resulting in relatively strong reproductive isolation (Seitz and Komma 1984; Romstock-
107 Volkl 1997; Diegisser et al. 2006a; Diegisser et al. 2006b). Moreover, mitochondrial networks and
108 allozyme markers suggest they are genetically differentiated (Diegisser et al. 2006a; Diegisser et
109 al. 2006b). Taken together, these findings suggest that there is strong divergent ecological selection

110 acting on these host races. This study system also exhibits independent contact zones where the
111 host races are found in sympatry both west and east of the Baltic Sea (Nilsson et al. 2022), enabling
112 us to determine the genomic regions that differ between host races in parallel, and hence are likely
113 to confer ecological adaptation. Specifically, we address which genomic regions are involved in
114 the host shift, whether structural variation is important, and to what extent these regions are
115 resistant to gene flow between the host races.
116

117 **Methods**

118 The study species *T. conura* (Loew, 1844; Diptera) is a true fruit fly in the family Tephritidae.
119 Most species in this family are specialists on one or a few plants (Aluja and Norrbom 2001) where
120 larvae feed on fruits, seeds, flowers, stems or plant roots (Christenson and Foote 1960; Headrick
121 and Goeden 1998). Within Tephritidae, several examples of host races occur in the genera *Eurosta*
122 (Craig et al. 1993), *Rhagoletis* (Bush 1969; Feder et al. 1998) and *Tephritis* (Diegisser et al. 2004;
123 Diegisser et al. 2006b). Within *T. conura*, host plant specialization has been documented in
124 populations in continental Europe, following a host shift from *C. heterophyllum* (referred to as
125 CH-flies from here on) to the derived host plant *C. oleraceum* (referred to as CO-flies) (Romstock-
126 Volkl 1997). The host races are nearly indistinguishable morphologically, with ovipositor length
127 being the only trait that enables classification in isolation (Diegisser et al. 2007; Nilsson et al.
128 2022). Sampling adult flies on their host plants or as larvae in infested buds also facilitates
129 discrimination.
130

131 In order to uncover the genomic regions involved in host plant adaptation in *T. conura*, we
132 assembled and annotated a reference genome and used whole genome resequencing of four

133 populations from each host race. For both host races, *T. conura* larvae inside infested *Cirsium* buds
134 of the respective host plants were collected during June and July 2018. We sampled infested buds
135 from eight populations, four for each host race, including sites in allopatric populations of *C.*
136 *oleraceum* in Germany and Lithuania, allopatric populations of *C. heterophyllum* in northern
137 Sweden and Finland, and sympatric populations of both host races in southern Sweden and Estonia
138 (Fig. 1; Table S1). This design enables us to identify the genomic regions that consistently differ
139 between host races, as the sympatric regions sampled on each side of the Baltic Sea constitute
140 independent contact zones. The pupae used for the reference genome was sampled from a CH-fly
141 individual from the CHST population (Fig. 1D).
142

143 Thistle buds containing *T. conura* larvae were stored individually in netted cups at 21ºC in a
144 laboratory at the Department of Biology, Lund University, Sweden. Emerging adult flies were
145 provided with honey water (1:2 ratio) *ad libitum* smeared on top of the netting. Three days after
146 the first emergence, representing a sufficient interval for all flies to eclose, cups containing adult
147 flies were relocated to a climate chamber (7ºC; 8:16 light:dark cycle). We considered all flies
148 emerging from a given bud as potential siblings and therefore sampled only a single adult male
149 from each thistle bud for genomic analysis to avoid including related individuals. Flies and the
150 pupae were euthanized through snap freezing and stored at -80ºC.
151

152 *Genome assembly and improvement*

153 DNA from a single *T. conura* pupa harvested from a *C. heterophyllum* bud using was extracted by
154 SciLifeLabs, Solna, Sweden following their in-house protocol for PacBio sequencing. Briefly, the
155 DNA was sequenced using PacBio long read technology and sequences were error-corrected using

156 SMRT Link analysis. Sequencing resulted in 4,173,684 reads with a total of 74.3Gbp and an N50
157 read length of 17.8 kb. PacBio sequences were assembled with hifiasm 0.7-dirty-r255 (Cheng et
158 al. 2021), which was run using default parameters, followed by purge_dups (v. 1.2.5; (v. 1.2.5;
159 Guan et al. 2020) to identify and remove heterozygous duplication (e.g. haplotigs) in the resulting
160 assemblies.
161
162 The reference assembly was scanned with Kraken2 to identify bacterial, viral and fungal
163 sequences, and contigs with a contamination proportion of greater than 10% were removed (150
164 contigs; 47,038,732 bp). Due to the abundance of short contigs in our assembly, we used a close
165 relative, *Rhagoletis pomonella*, to achieve a chromosome-level perspective. We aligned the *T.*
166 *conura* contigs to the *R. pomonella* assembly (GCF_013731165.1_Rhpom_1.0; accessed 22 Nov
167 2022; length = 1223 Mbp, scaffolds = 32,060, scaffold N50 = 72Mbp) using minimap2 (asm20;
168 Li et al. 2021) within the RagTag wrapper (v. 2.1.0; Alonge et al. 2022). RagTag ordered *T. conura*
169 contigs based on *R. pomonella* scaffolds. This order was used in downstream analyses to cluster
170 *T. conura* contigs within hypothetical linkage groups. Putative sex-linked contigs were identified
171 as those with an excess of hits (at least one hit per million bp, approx. top 5% of contigs) when
172 blasted (tblastn; v.2.11.0; Camacho et al. 2009) with a set of *Drosophila melanogaster* X-linked
173 proteins. This resulted in 102 putative X-enriched contigs, 68.6% of which were clustered on two
174 RagTag linkage groups: NW_023458556.1_RagTag (n = 56) and NW_023458567.1_RagTag (n
175 = 14; Fig. S1). The genome was sequenced from a pupa of unknown sex, but coverage of both
176 long and short reads mapped across these contigs suggest that the pupa was male, as *Tephritis* is
177 one of the only fly genera in which females are heterogametic (Vicoso and Bachtrog 2015).
178

179 We used RepeatModeler (v. 2.0.3; Flynn et al. 2020) and RepeatMasker (v. 4.1.2; Smit et al. 2015)
180 to identify and soft mask repetitive elements in the *T. conura* reference assembly. For
181 RepeatModeler, we specified the 'rmblast' engine and ran the LTR structural pipeline. We ran the
182 -gccalc option in RepeatMasker to account for variable GC content. To put the *T. conura* results
183 in context we accessed repeat summaries for nine Tephritinae species with published genomes
184 (Sproul et al. 2022) While Sproul et al. (2022) performed additional repeat screens, we used the
185 output from their first round of masking, in which they used RepeatModeler2.0 (search engine
186 "ncbi") and RepeatMasker4.1.0 to generate custom repeat libraries.
187
188 *Gene prediction and annotation*
189 We annotated the assembly using the MAKER pipeline (3.01.03; Cantarel et al. 2008). We
190 generated expressed sequence predictions using RNAseq reads from flash-frozen flies from
191 various life stages, sexes and populations (Table S2). Extractions were performed using Sigma
192 Aldrich's Plant RNA kit, which we have found performs best with these flies. After sample quality
193 testing, Illumina TruSeq Stranded mRNA libraries were prepared by SciLifeLab (Stockholm,
194 Sweden, https://ngisweden.scilifelab.se) and sequenced using NovaSeq6000. We used two tools
195 to generate EST predictions from this RNAseq evidence. First, we mapped the reads to the soft-
196 masked reference genome using the splice aware mapper HISAT2 (v. 2.2.1; Kim et al. 2019).
197 Mapped reads were sorted and indexed with SAMtools (v. 1.14; Danecek et al. 2021). Transcripts
198 were predicted and merged into a single annotation with StringTie (v. 2.1.4; Kovaka et al. 2019).
199 We also generated a *de novo* transcriptome prediction with Trinity (v. 2.11.0; Grabherr et al. 2011).
200 Protein evidence was downloaded from Uniprot database (UniProt Consortium 2021; accessed
201 June 2021) under the taxonomy"Acalyptratae 99 [43741]" and only reviewed homologs were used.
202 Protein and EST predictions were synthesized into a final annotation with 27588 transcripts. The

203 final annotation was assessed using BUSCO and the diptera_odb10 database, which found 84.6%
204 completeness (82.9% single copy, 1.7% duplicated), 3.7% fragmented and 11.7% missing. We
205 functionally annotated the predicted genes with eggNOG emmaper (v. 2.1.5; Huerta-Cepas et al.
206 2018) using default settings. Functional annotations were returned for 14812 transcripts, including
207 gene ontology (GO) terms for 9894 of these.
208
209 *DNA extraction and sequencing*
210 We extracted DNA for whole genome sequencing from entire flies using Qiagen DNeasy Blood
211 and Tissue Kit (Qiagen Corp., Valencia, CA), following the standard protocol with some small
212 modifications (Supplementary Methods 1). Frozen flies were briefly thawed and ground with
213 plastic pestles in the tissue lysis buffer. Samples were incubated with proteinase K (3.5h, 56°C),
214 and then RNAse was added and the pellet eluted using a 56µl of EB buffer. Library preparation
215 and whole genome resequencing were performed by SciLifeLab (NGI-Sweden, Solna,Sweden)
216 using Illumina TruSeq PCR-free prep kit with an insert size of 350bp. Sequencing of 2x150 bp
217 paired-end reads was carried out on 8 Illumina HiSeqX lanes.
218
219 *Mapping, variant calling and allele frequencies*
220 We aligned the whole genome sequencing data of all individuals to the reference genome using
221 bwa mem (v.7.17-r1188; Li and Durbin 2009), then merged, sorted and indexed using samtools
222 (v. 1.14; Li et al. 2009). PCR duplicates were identified and removed using MarkDuplicates in
223 Picard tools (v2.10.3; Broad Institute). Genome-wide mean coverage was calculated with samtools
224 (v 1.9; Fig. S2).
225

226 To account for low coverage, we utilized genotype likelihood based methods for analyses when
227 possible. To accommodate both genotype likelihood and genotype call based analyses, we
228 produced both (1) a Beagle file containing genotype likelihoods for the 96 *T. conura* individuals
229 and (2) a VCF containing genotype calls for the 96 individuals as well as three outgroup individuals
230 using ANGSD (v. 0.940; Korneliussen et al. 2014). The filters and steps to produce these two files
231 are described in Supplementary Methods 2. The final Beagle file and VCF contained 18,722,827
232 and 53,722,613 SNPs, respectively.
233
234 Per population, we generated site allele frequency indices (SAF) and inferred minor allele
235 frequencies (MAFs) using ANGSD, relying on the same filters as described for the Beagle file,
236 excluding those used for SNP calling (Supplementary Methods 2). We produced a consensus
237 FASTA from the mapped reads of one of the outgroup individuals (P18705_115, *Tephritis*
238 *kogardtauica*) using ANGSD and polarized the SAF files to this sequence (-anc outgroup.fa). As
239 this FASTA is produced from mapped reads of an outgroup individual, many positions did not
240 have sufficient coverage to be called in the FASTA. As such, the SAFs only estimated frequencies
241 at positions that both passed filters set during SAF production and had a non-ambiguous base call
242 in the outgroup FASTA file. We produced unfolded 1D and 2D site frequency spectra (SFS) for
243 each population and population pair using the SAF files in realSFS (Nielsen et al. 2012).
244
245 *Population clustering*
246 We investigated the genetic clustering of the individuals using principal component analysis
247 performed in PCAngsd (v. 0.982; Skotte et al. 2013) for exploratory and illustration purposes, and
248 a formal clustering analysis for 1-10 clusters (K) as implemented in NGSAdmix (v. 33; Skotte et

249 al. 2013). To prepare the input files for these tools, we estimated pairwise linkage disequilibrium
250 between all SNPs within 100 kb of each other using ngsLD (Fox et al. 2019). We then used the
251 included prune_ngsLD.py script to produce a list of linkage pruned SNPs, excluding edges from
252 the initial graph between sites with a distance greater than 50 kb and with an r^2 of less than 0.1 (--
253 max_dist 50000 --min_weight 0.1). We then subset our Beagle file to these linkage pruned sites
254 and used it as input for both PCAngsd and NGSAdmix. For the clustering analysis, we performed
255 between 10-100 replicate optimizations for each K with a maximum of 4000 iterations per
256 replicate. Replicates were halted when the replicates converged or 100 replicates had completed,
257 considering replicates as converging when the three highest likelihood replicates came within 2
258 log-likelihood units of each other, as in Pečnerová et al. (2021). We selected the best K as the
259 highest level of K achieving convergence.
260
261 *Demographic history*
262 Demographic history was estimated for each population using MSMC2 (v. 2.1.2; Schiffels and
263 Durbin 2014). We generated the individual VCF and mask bed file by using bcftools (v. 1.14; Li
264 2011) and *bamCaller.py* script from msmc-tools (https://github.com/stschiff/msmc-tools;
265 Accessed on 2023-02-09). We also generated a reference genome mask file to eliminate sites with
266 extremely low or high coverage, defined as 0.5x and 2x mean coverage respectively. We used
267 Samtools (v1.16) to compute the average sample coverage and generate a BED file per-sample.
268 We merged the VCF and mask files (msmc-tools, *generate_multihetsep.py* script) and used the
269 resulting file as input for MSMC2. To scale the result, we used a mutation rate of 3.46×10^{-9}
270 mutations per site per generation, which is derived from estimates in *Drosophila melanogaster*
271 (Keightley et al. 2009) and a mean generational interval of 1 year (Romstock-Volkl 1997). Because

272 this software is sensitive to read coverage, neutrality, and repetitive content, we compared the
273 MSMC2 output for three sets of samples (the full set of samples, the individuals with highest
274 coverage in each of the eight populations, and the individuals with coverage closest to the global
275 mean in each of the eight populations) at three levels of VCF filtering (the VCF output from
276 bcftools + *bamCaller.py* script described above; the VCF excluding all repetitive content, and the
277 VCF excluding all outlier windows identified by population branch structure analysis described
278 below).
279
280 *Population differentiation and divergence*
281 To measure population differentiation, *Fst* was calculated for each site, first by generating the site
282 frequency spectrum (2D-SFS) for each population pair, based on which we calculated per-site *Fst*
283 using a Bhatia estimator (Bhatia et al. 2013) as implemented in ANGSD (realSFS *Fst*). Average
284 pairwise *Fst* was estimated genome-wide and for nonoverlapping 10kb windows (realSFS stats2).
285 We used population branch statistics (PBS) to assess differentiation between sympatric
286 populations of the two host races when corrected for divergence between populations within host
287 races. This accounts for heterogeneity in background divergence between populations. We
288 calculated PBS for each site (realSFS stats2) for four population triads, where each sympatric
289 population was the target in one comparison. Triads focused on populations where host plants are
290 sympatric, for which we expect selection favoring host specificity to be strongest, and were
291 composed of a sympatric population specializing on host plant A compared against the sympatric
292 and allopatric populations specializing on host B in either the east or west transects. Thus, PBS_{CHES}
293 quantifies the differentiation between sympatric CH-flies (CHES) and sympatric CO-flies
294 (COES), corrected for divergence from allopatric CO-flies (COLI) populations (Fig. S3

295 Supplementary Methods 3). PBS was estimated for nonoverlapping 10kb windows. Outlier
296 windows in each of these triads were identified as those with PBS values greater than four standard
297 deviations away from the genome-wide mean (Salmón et al. 2021; Montejo-Kovacevich et al.
298 2022).
299
300 *Genome-wide association analysis with BayPass*
301 Genome-wide scans for association with ecotype were performed with BayPass (v. 2.3; Gautier
302 2015). We converted the filtered VCF to allele frequencies for each population with bcftools. The
303 dataset was further filtered to account for linkage by subsampling every 20th variant across each
304 contig into 20 unique sets of SNPs. The data were analyzed first under the BayPass core model
305 using default option, then under the auxiliary model using a Markov Chain Monte Carlo (MCMC)
306 algorithm and ecotype as a covariate. Three independent runs (using the option -seed) were
307 performed for each of the 20 SNP sets. Support for association with ecotype was evaluated using
308 the mean Bayes Factor of these three runs. Results from the 20 independent runs were combined
309 into a single dataset for all sites, and reported in deciban units, with 10db corresponding to 10:1
310 odds, 20db to 100:1 odds, etc.).
311
312 *Outlier analysis and gene ontology term enrichment*
313 To identify genomic regions associated with host use we identified outliers for three different
314 metrics, PBS and dxy that were assessed independently for the western and eastern populations
315 respectively, and BayPass that was estimated based on all eight populations. Windows with PBS
316 and dxy that exceeded the genome-wide mean by four standard deviations or had a Bayes Factor
317 (in decibans) of 20 or higher were classified as outliers. As sex chromosomes and autosomes differ

318 in effective population size and sex chromosomes evolve faster, we performed all outlier analyses
319 on three different data sets, one data set including the entire genome, one including only putatively
320 X-enriched contigs, and one including only autosomal contigs. This approach allowed us to
321 identify enriched functions that were associated with either X-enriched or autosomal contigs.
322
323 For each of these outlier sets, GO-enrichment analyses were performed using TopGO (Alexa and
324 Rahnenfuhrer 2022), with a minimum node size of 5. Significant enrichment was tested with one-
325 sided Fisher's exact tests corrected with the parent-child algorithm (Grossmann et al. 2007). GO
326 terms were considered significantly enriched with a p-value < 0.01. We also performed GO-
327 analyses on the outliers that occurred in two of the three outlier sets in the west or in the east to
328 capture the outliers that are most consistently diverged.
329
330 *Nucleotide diversity and divergence from neutrality*
331 We evaluated nucleotide diversity (π) and divergence from neutrality, as characterized by Tajima's
332 D, to understand genome-wide and local differences among populations. ANGSD was used to
333 calculate pairwise nucleotide diversity for each site (thetaD, realSFS saf2theta). These were used
334 to estimate Tajima's D and pairwise differences across the entire genome and over 10Kb
335 nonoverlapping windows (thetaStat do_stat). To calculate π, pairwise differences were divided by
336 the total number of sites used in each window. For both metrics, we set a minimum coverage
337 fraction of 0.4 for our window-based analysis, meaning we required genotypes for at least 4kb of
338 the 10kb in each window. Windows below this threshold were excluded from downstream
339 analyses.
340

341 *Selection statistics*

342 To identify SNPs associated with host races and different biogeographic scenarios we used Selscan
343 (v. 1.3; Szpiech and Hernandez 2014) to calculate the cross-population extended haplotype
344 homozygosity (XP-EHH). XP-EHH is a cross-population statistic that test for differential local
345 adaptation and calculates the extended haplotype homozygosity (EHH) for two populations
346 separately, integrated over recombination distance and compared in order to identify loci under
347 selection in one of the two populations. To prepare the data for XP-EHH analysis, SNPs unpruned
348 for linkage disequilibrium were phased with SHAPEIT2 v 2.r837 (Delaneau et al. 2013) using
349 default MCMC parameters (7 burn-in MCMC iterations, 8 pruning iterations, and 20 main
350 iterations), conditioning states for haplotype estimation (K = 100), and window size set at 0.5Mb
351 as recommended for whole genome sequence data. We also used an effective population size (Ne)
352 of 494, based on estimates calculated with SNeP (v. 1.1; Barbato et al. 2015). Here, we used the
353 mean across populations, as all estimates were very similar. XP-EHH was calculated in windows
354 of size 100 kb in each direction from core SNPs, allowing decay curves to extend up to 1 Mb from
355 the core, and SNPs with MAF < 0.05 were excluded from consideration as a core SNP. As we
356 lacked a fine-scale genetic map for *T. conura*, we assumed a constant recombination rate of 1.73
357 cM/Mb as found for the related species *Bactrocera cucurbitae*. Scores were then normalized
358 within contigs with the norm version of selscan version 1.3.0 program.

359

360 The XP-EHH method identifies genomic regions that underwent a selective sweep in one
361 population but remained variable in the second population. We calculated XP-EHH for six
362 population pairs to test for differences in selection between host races in sympatry and allopatry,
363 and to test for differences between western and eastern transects. For each pair, we identified

364 putative locality-specific sweeps by selecting those SNPs above the 99th percentile of the genome-
365 wide distribution. Genes under selection were detected by comparing the SNPs detected by the
366 XP-EHH test against a gene annotation for *T.conura* with bedtools v 2.29.2 window 2 kb upstream
367 and downstream these genes (Quinlan and Hall 2010).
368
369 *Introgression statistics*
370 We investigated if there is evidence of introgression between populations by estimating genome-
371 wide Patterson's D and f_4-ratios (Patterson et al. 2012) for all possible population trios, as
372 implemented in Dsuite's DTrios function (Malinsky et al. 2021). To assess variation in
373 introgression across the genome in sympatric populations, we investigated f_{dM} in windows across
374 the genome for eight focal trios, as implemented in the script ABBABABAwindows.py
375 (https://github.com/simonhmartin/genomics_general). We selected this statistic as it is designed to
376 detect introgression in localized regions of the genome and is suitable for detecting introgression
377 between both P1 & P3, as well as P2 & P3. Briefly, for each of the four sympatric populations as
378 P2, we estimated f_{dM} in non-overlapping 10 kb windows, assigning the closest allopatric population
379 of the same host plant to P1, with P3 being either the closest allopatric or sympatric population of
380 the alternate host plant. For all analyses we utilized a *Tephritis hyoscyami* individual
381 (P18705_127) as the outgroup.
382
383 **Results**
384 Our final assembly of the *T. conura* genome was 1.99 Gbp in 2713 contigs (N50 = 1.75 Mbp).
385 Benchmarking with Dipteran single copy orthologs found 97.0% completeness with low
386 duplication (2.4%), making it not only the longest but among the most complete Tephritidae

387 genomes available to date (Fig. S4). It was also the most repeat rich genome by more than 20%,
388 driven by an expansion of long terminal repeats that make up over 22% of the genomic content
389 (Fig. S5; Table S5). This history of genomic expansion and high repetitive content, which increase
390 the opportunity for both structural change and chromatin state differences, suggests that *T. conura*
391 has the potential for extensive and rapid genomic divergence driven by structural variants.
392
393 *Genomically divergent host races*
394 We generated whole genome sequence data from eight continental *T. conura* populations
395 specializing on either *C. oleraceum* and *C. heterophyllum*. Populations were located along two
396 parallel transects west and east of the Baltic Sea, covering regions where host plants can be found
397 in either allopatry or sympatry (Fig. 1D). We found considerable genomic divergence, with *C.*
398 *heterophyllum* (CH) and *C. oleraceum* (CO) host races representing distinct lineages. The two host
399 races separated in a formal clustering analysis implemented in Admixture (Fig. 1E) and
400 consistently, in a genome-wide exploratory analysis of variants, the first PC axis, explaining 4.7%
401 of the variance, separated individuals by host race (Fig. 1F). Within the derived CO host race,
402 individuals cluster by geography, with western and eastern populations separating along the second
403 PC axis and in the clustering analysis (Fig. 1F). The western sympatric CH population (CHSK)
404 appears to be the most distinct lineage, separating from both CO individuals and the remaining CH
405 individuals along PC1, forming a separate cluster under the best supported model in the clustering
406 analysis (K=3; Table S4). Individuals from the remaining populations were grouped largely by
407 ecotype, with some evidence of shared ancestry (Fig. S6). These results were supporeted by
408 pairwise genome-wide *Fst* and dxy comparisons, which showed the largest differentiation and
409 divergence occurred between populations of different host races (Fig. S7).

410

411 In contrast to earlier suggestions of the derived host race arising before the onset of the last glacial
412 maximum c. 18000 years ago (Diegisser et al. 2006b) , an MSMC2 analysis suggests host races
413 diverged over one million years ago (Fig. 1G). These results broadly confirm findings from
414 clustering analyses, including western and eastern clustering of the CO host race. In particular, the
415 highly divergent CHSK population again appears as a distinct lineage, perhaps having diverged
416 from the other populations much earlier. However, similar patterns can arise when populations
417 have been affected by recent bottlenecks, and we have been unable to disentangle these alternative
418 explanations at this time. Finally, we found that our MSMC2 analyses were highly sensitive to
419 both genomic content (e.g., nonneutral sites, repetitive content) and coverage (Fig. S8)

420

421 *Genomic regions underlying host plant specialization*

422 The genomic regions that separated the host races were highly consistent between the western and
423 eastern transect for each of the tests we performed (PBS, dxy and BayPass Fig. 2A-B, C-D and E
424 respectively; Fig. S9-S12). We found consistently higher differentiation and divergence for
425 nonoverlapping 10kb windows located on contigs enriched for X-linked *D. melanogaster*
426 orthologs (X-enriched, putatively sex-linked) than the remainder of the genome. Both west and
427 east of the Baltic, population branch statistics (PBS) for sympatric populations recover many
428 outliers exceeding the mean by more than 4SD, both on putatively sex-linked and autosomal
429 contigs and in eastern and western host race comparisons (Fig. 2A-B; Fig. S10). Many of these
430 outliers clustered into distinct peaks, often affecting a majority of the 10kb windows within a given
431 contig. There was a remarkably high degree of overlap between western and eastern PBS outliers.
432 Fewer and less clustered outlier regions were identified using dxy, but dxy outliers were also highly

433 consistent across the two host race comparisons (Fig. 2C-D; Fig. S11-S12). BayPass, which
434 identified sites associated with host use among all eight populations, recapitulated many of the
435 divergent peaks found using the PBS analyses (Fig. 2E). However, using the conservative
436 recommended cutoff of BF > 20, only 60% of the BayPass outlier windows overlapped PBS
437 outliers in the west and east transects, and no BayPass outliers overlapped dxy outliers. We found
438 surprisingly low overlap among PBS, dxy and BayPass outliers within the eastern and western
439 host race comparison. As a result, we identified 'host race outlier' windows as those with support
440 from at least one statistic within each transect.
441
442 Host race outlier windows were more abundant on X-enriched contigs than those falling on
443 putative autosomes, and a greater proportion overlapped annotated genes (Fig. S13).
444 Gene ontology (GO) analyses recovered enrichment of early morphological development (e.g.,
445 GO:0048598 embryonic morphogenesis; GO:0001704 formation of primary germ layer) and
446 growth and mitosis (e.g., GO:1901722 regulation of cell proliferation, GO:0051782 negative
447 regulation of cell division; Fig. 3). RNA surveillance and molecular transport were also enriched
448 in both the west and the east, and these were similarly supported by enriched cellular components
449 (GO:0005921-gap junction; GO:0000701-condensed nuclear chromosome) and molecular
450 functions (GO:0070034 telomerase RNA binding; GO:0015078 proton transmembrane transporter
451 activity; GO:0005342-organic acid transmembrane transporter activity). A minority of outliers
452 were located on putative autosomes, and as a result the genes overlapping these windows in both
453 the west and east were enriched for few GO terms. However, appendage morphogenesis
454 (GO:0035107) was enriched among both X-enriched and autosomal outliers (Fig. 3), which may
455 support our hypothesis that nonphysical linkage underlies some host-specific adaptations in *T.*

456 *conura*. When we limited outlier windows to those with support from two or more statistics (i.e.,
457 PBS and Dxy or PBS and BayPass), we found that very few of the same biological process terms
458 were enriched compared to outliers supported by one or more statistic (Fig. 2F; Fig. S14). Rather,
459 a large number of terms involved in neurological processes were enriched, suggesting highly
460 differentiated and divergent regions may be involved in cue processing, related to interaction with
461 host plants, or communication, related to interaction with conspecifics (Fig. S14).
462
463 *Selection for host race divergence*
464 To understand the selective pressures acting on the regions of the genome that differentiate the
465 host races, we tested for evidence for selective sweeps, and whether these are stronger in the
466 derived CO host race. To do this, we assessed Tajima's D, nucleotide diversity (π) and extended
467 haplotype homozygosity. When considering all windows across the genome, Tajima's D was
468 similar among populations (Fig. S15-16). The exception was CHSK, which had a much higher
469 Tajima's D, supporting the conclusion from Admixture and MSCM analyses that this population
470 has experienced a recent bottleneck. In contrast to results for all windows, Tajima's D in outlier
471 windows is lower in CO populations than in CH populations, which is consistent with these regions
472 having experienced strong positive or purifying selection. Similarly, π within outlier windows is
473 also lower for CO populations (Fig. S17-19), suggesting selection has disproportionately decreased
474 nucleotide diversity in these regions in the CO population.
475
476 Across the genome, we found low Dxy, π, Tajima's D associated with repeat-dense regions of the
477 genome, even after imposing a coverage fraction filter that excluded windows missing genotypes
478 for 60% or more of the bases. However, we were able to confirm that this association is unlikely

479 to be driving the evidence for selective sweeps described above. We found that there was no
480 relationship between the proportion of repetitive content in outlier windows and estimates of Dxy
481 (Fig. S20) , π (Fig. S21), or Tajima's D (Fig. S22), suggesting these results are robust to the effects
482 of repetitive content.
483

484 Extended haplotype statistics reveal considerably stronger selection among the top selected SNPs
485 in CO populations when contrasted with CH populations. With a cutoff of the top 1% of SNPs, we
486 we recover a similar number of SNPs under selection in all the populations (Range: 253525 –
487 279839, Table S5). In all comparisons between CH and CO population pairs, the mean normalized
488 XP-EHH statistic of selected SNPs was higher in the CO host race. Meanwhile, in comparisons
489 within host race, there was evidence of stronger selection in sympatric populations in the western
490 transect (CHSK and COSK). We identified selected genes as those overlapping selected SNPs. We
491 found more genes overlapped selected SNPs in the CH host race (Table S5). This contrasting
492 pattern between SNPs and genes was due to the fact that the higher number of SNPs under selection
493 in CO populations affected a reduced number of annotated genes compared to CH populations
494 (Table S5). Furthermore, these genes were more likely to overlap with genes containing at least
495 one highly differentiated outlier window (as in Fig 3F; 4) in CO populations than in CH
496 populations (Fig. S23), again supporting that selected SNPs fall in a more concentrated, more
497 differentiated set of genes. This pattern is strongest in the COSK population, where in the number
498 of SNPs in annotated genes is an order of magnitude higher than in the other populations (Table
499 S5) and almost doubles the number of SNPs in annotated genes in CHSK.
500

501 *Regions permeable to introgression*

502 We find evidence consistent with introgression between the host races in both the eastern and
503 western sympatric regions. Using D-statistics, we found introgression between host races was
504 higher in the eastern than the western transect, consistent with this being an older contact zone for
505 these host races (Fig. S24, Table S6) We also found greater evidence of introgression into
506 sympatric populations of the alternate host race than into more distant allopatric populations.
507
508 Introgression was not evenly distributed across the genome, with outlier windows showing distinct
509 patterns of introgression compared to nonoutliers (Fig. 5). We used f_{dM} statistics to compare
510 introgression in 10kb windows across the genome. This statistic is expected to be positive when
511 introgression occurred between P3 and P2 and negative values if it occurred between P3 and P1.
512 After filtering, non-outlier windows had a much wider spread of f_{dM} values than outlier windows.
513 Similar to genome-wide results, f_{dM} in nonoutlier regions were consistently higher in trees testing
514 introgression into sympatric populations of the other host race than in trees testing introgression
515 into allopatric populations. In contrast, outlier regions appeared more resistant to introgression.
516 Across most triads, f_{dM} in outlier regions was lower, and in western transects significantly more
517 negative, than in nonoutliers. Negative values are consistent with greater introgression between
518 the more geographically distant populations, which may reflect historical gene flow. These
519 patterns suggest outlier regions have increased resistance to introgression between sympatric
520 populations. The exception to this result occurred in triads comparing introgression from CH
521 populations (either CHFI or CHES) into the two CO populations in the east. In both cases, outliers
522 had higher f_{dM} than nonoutliers.
523
524 **Discussion**

525 We find strong and consistent genomic divergence between the two *T. conura* host races in both
526 allopatry and sympatry. These results were highly parallel between contact zones both west and
527 east of the Baltic Sea. Putatively sex-linked contigs consistently show high divergence between
528 the host races. Furthermore, these contigs form discrete clusters when aligned to contiguous
529 regions in the genome of the closely related species *R. pomonella*. The recent and extensive
530 expansions of long terminal repeats, and discrete nature of the divergent genomic areas could
531 suggest a role for repeat-facilitated sex-linked inversions in separating the two host races of *T.*
532 *conura*. Potentially, these regions could couple the multifarious ecological traits that are associated
533 with host plant, including phenology (Romstock-Volkl 1997), host preference (Diegisser et al.
534 2008), morphology including ovipositor length (Diegisser et al. 2007; Nilsson et al. 2022), and
535 physiology including digestion of plant defense chemicals (Diegisser et al. 2008) and thus facilitate
536 coexistence between the host races. Recent research has shown that inversions are important for
537 maintaining both discrete morphs within species (Jones et al. 2012; Küpper et al. 2016; Tuttle et
538 al. 2016; Faria et al. 2019a), facilitate speciation (Wellenreuther and Bernatchez 2018; Faria et al.
539 2019b) and enable coexistence with sister taxa after a shorter period of time (Hooper and Price
540 2017). Our findings suggest that inversions may be important for coupling (Butlin and Smadja
541 2018) of ecologically co-adapted traits also in *T. conura*. However, we also found that discrete
542 autosomal regions show elevated PBS, *Fst* and Dxy, and were associated with host race divergence
543 in BayPass analyses. This supports that, in addition to physical coupling facilitated by inversions,
544 nonphysical coupling of host race-associated loci may be contributing to ecological divergence
545 and reproductive isolation in peacock flies.
546

547 Consistent with expectations under strong divergent ecological selection, the divergent areas are
548 highly consistent across the eastern and western sympatric zones, both for PBS and dxy.
549 Introgression statistics suggest low levels of gene flow across many, but not all, regions of the
550 genome in the sympatric areas. Regions of the genome not associated with host use should
551 exchange freely, and the presence of discrete genomic regions that remain divergent between host
552 races in sympatry suggests they contain divergent adaptations in the two host races. Moreover,
553 genomic regions important for host adaptation are expected to diverge in parallel in the sympatric
554 zones on both sides of the Baltic, as parallelism in response to a similar environment is a hallmark
555 of selection (Johannesson 2001; Stuart et al. 2017). Hence, the genomic regions identified in our
556 analyses are highly likely to be important for host race specific ecological adaptation.
557
558 Interestingly, there is also an overlap in functions of the genes enriched in outlier sets based on
559 PBS, dxy, BayPass or the intersect. Functions that frequently were enriched include metabolic
560 biological processes, neural signaling, and morphogenesis. The divergence in genes associated
561 with metabolic processes is expected as the host races suffer strong extrinsic inviability as larvae
562 when feeding on the wrong host plants (Diegisser et al. 2008). While the exact function of the
563 genes contributing to the enrichment of morphogenesis among diverged genes are not known, it
564 would be interesting to examine the expression of these genes during ovipositor formation, as the
565 relative length of the ovipositor is the most defining morphological difference among the host
566 races (Diegisser et al. 2007; Nilsson et al. 2022). There are also size differences between the host
567 races that potentially could be affected by these genes. Interpretation of neural signaling remains
568 speculative, but examining the expression of these genes in the brain when exposed to different

host plants would uncover if the divergence in neural signaling could be of importance in host plant preference.

Tephritis conura has parallels with the co-existence of different host races of *Rhagoletis* infesting hawthorne and apple, respectively (Feder et al. 1988; Filchak et al. 2000; Feder et al. 2003), the parapatric ecomorphs of *Littorina* snails adapted to either wave-exposure or crab predation (Johannesson and Johannesson 1996; Hollander et al. 2005; Johannesson et al. 2010; Faria et al. 2019a; Perini et al. 2020), and *Timema* walking sticks adapted to be cryptic on different host plants (Nosil et al. 2002; Nosil 2007; Nosil and Sandoval 2008; Villoutreix et al. 2020). Each involve strong selection pressures that maintain ecologically divergent population in sympatry despite incomplete reproductive isolation. Rhagoletis flies are isolated by phenological difference and host preference (Feder et al. 2003), but the genomic differences that underlie the differently adapted ecomorphs is polygenic and can rapidly be recovered through selection on standing genetic variation within the ancestral hawthorn host race (Egan et al. 2015). The genomic basis conferring mimicry on different host plants in Timema walking consists of few loci of large effects, and the background genomic divergence is lower (Villoutreix et al. 2020). The genomic basis of the Littorina ecotypes is similar to that in *T. conura* in that variants within inversions contribute to approximately half of the ecotype divergence (Koch et al. 2022). Compared to those systems, the study races in *T. conura* have relatively high levels of genetic divergence and strong reproductive isolation, arising from the combination of preference, phenology and performance in the alternative niche (Diegisser et al. 2006a; Diegisser et al. 2006b, 2007; Diegisser et al. 2008; Nilsson et al. 2022). Potentially, the adaptation is more multifarious, or the adaptation may have taken place during a longer period of time. In contrast to previous divergence time estimates based

592 on mitochondria, MSMC analyses suggest that host race divergence may be old. Potentially, the
593 host races could be older than 500 kya as there is no convergence between the host races in the
594 MSMC plots even at this time scale. Further exploration of the dependency of the MSMC inference
595 of which fractions of the genome are included would be necessary to draw any firm conclusions
596 though. Given the morphological similarity of the species (Nilsson et al. 2022), it may seem
597 surprising that divergence times exceed previous estimates of just before the last glacial maximum.
598 However, our new estimates are consistent with the high level of genomic divergence and limited
599 introgression on the putatively sex-linked contigs observed in this study. Furthermore, we believe
600 this timeframe is also in accordance with strong reproductive isolation in this system, evidenced
601 by phenological differences (Romstock-Volkl 1997), oviposition preference and larval mortality
602 on the new host plant (Diegisser et al. 2008). Hence, *T. conura* provides a window into a later
603 stage of ecological speciation continuum (Stankowski and Ravinet 2021) with the potential to offer
604 interesting insights into the development of reproductive barriers following ecological
605 specialization.
606
607 Interestingly, the scaffolds that were enriched for *Drosophila melanogaster* X-linked proteins
608 showed on average higher divergence. Several different factors, including the lower effective
609 population sizes of sex chromosomes (Charlesworth et al. 2018), their potential lower rates of
610 recombination (Charlesworth 2017) and the faster Z (Mank et al. 2009) are expected to contribute
611 to this pattern, but differ depending on the evolutionary history of the sex chromosomes. These
612 processes could complicate the interpretation of genomic regions or genes important for ecological
613 adaptation on these contigs. In addition, while *Tephritis* flies are known to have ZW sex
614 determination (shown for *T. californica*; Vicoso and Bachtrog 2015), the origins of this switch to

615 female heterogamety are not known. Whether the Z-chromosome was converted directly from the
616 ancestral X-chromosome, or originated from an autosome while the ancestral X-chromosome lost
617 its sex determination role remains to be discovered, limiting our understanding of the evolutionary
618 dynamics of X-enriched contigs in *T. conura.* An additional intriguing possibility is that lower
619 levels of gene flow could result in the typically more divergent sex chromosomes with a large role
620 in reproductive isolation representing the ancestral species tree in face of introgression (Fontaine
621 et al. 2015). Future comparisons of coverage across the genomes of males and females and gene
622 expression data from reproductive tissues can shed further light on the sex chromosome in *T.*
623 *conura* and facilitate interpreting its importance for host race formation. The GO-terms for
624 putatively sex-linked scaffolds rarely overlapped with those enriched for outliers on the autosomes.
625
626 Long-read assemblies of the derived CO host race would be an interesting future venue to gain
627 further insights into the structural variation underlying host race formation and persistence in *T.*
628 *conura*. Moreover, the potential for transposable elements to contribute to reproductive isolation
629 and/or serve as a source for novel variation through altering regulation of gene expression and
630 remodeling chromosomes, has recently been recognized (Serrato-Capuchina and Matute 2018).
631 To which extent recent TE releases has contributed to the potential to adapt to novel niches, and
632 to reproductively isolate the host races of *T. conura* is an interesting venue for future research that
633 could help shed light on the role of TE releases in ecological speciation.
634
635 In a rapidly changing climate, understanding the ability to adapt to a novel niche is more important
636 than ever, as it has critical implications for conservation and our power to predict potential host
637 range expansions and new pest species following northward shifts of inset distributions. *Tephritis*

638 *conura* is an exemplar ongoing ecological speciation of two host races and future research on this
639 study system will offer unique insights into the changes in the genome necessary for rapid
640 colonization of novel niches.
641
642 **Data availability**
643 All code used will be deposited on github upon acceptance. The reference genome will be available
644 from NCBI and all resequencing data will be available as bam-files at NCBI/ENA upon manuscript
645 acceptance.
646
647 **Author contributions**
648 J. O. performed field work, extracted DNA and RNA, calculated recent effective population sizes
649 with SNeP and performed extended haplotype statistics analyses. K. J. N. performed field work,
650 performed bioinformatic analyses including PCAngsd, NGSadmix, *Fst*, dxy, pi and Tajima's D
651 with support from Z N and RAS, plotted figures and prepared the supplement. R. S. improved the
652 reference genome with analyses of microbial content and alignment to the *Rhagoletis pomonella*
653 genome, performed PBS analyses, BayPass analyses, GO-analyses and helped with other analyses
654 and plotting and wrote a first draft of the methods. Z. J. N. paved the pipeline used to generate *Fst*,
655 dxy, pi and Tajima's D as well as PCAngsd and NGSadmix. C. Y. generated alignments and the
656 maker annotation. Y. H. performed MSMC analyses. J. A. L. helped analyze *Fst*, dxy, pi and
657 Tajima's D. H. V. extracted DNA for the first long-read assembly. A. R. Conceived of, designed
658 and funded the study, performed field work, outlined the bioinformatics analyzes to be performed
659 and supported the students performing these, interpreted data, took a lead on the writing and wrote
660 the first draft of the introduction, results and discussion.

Acknowledgements

This work was funded by a fellowship from the Wenner-Gren foundations, a Swedish Research Council starting grant, and additional grants from the Crafoord foundation, Erik Philip Sörensens Stiftelse and Carl Tryggers Stiftelse covering sequencing to A.R. The authors acknowledge support from the National Genomics Infrastructure in Stockholm funded by Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council, and SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. We thank Jes Johanneson for support when starting the work on this study system, and Emma Kärrnäs and Mathilde Schnuriger for assistance during field work.

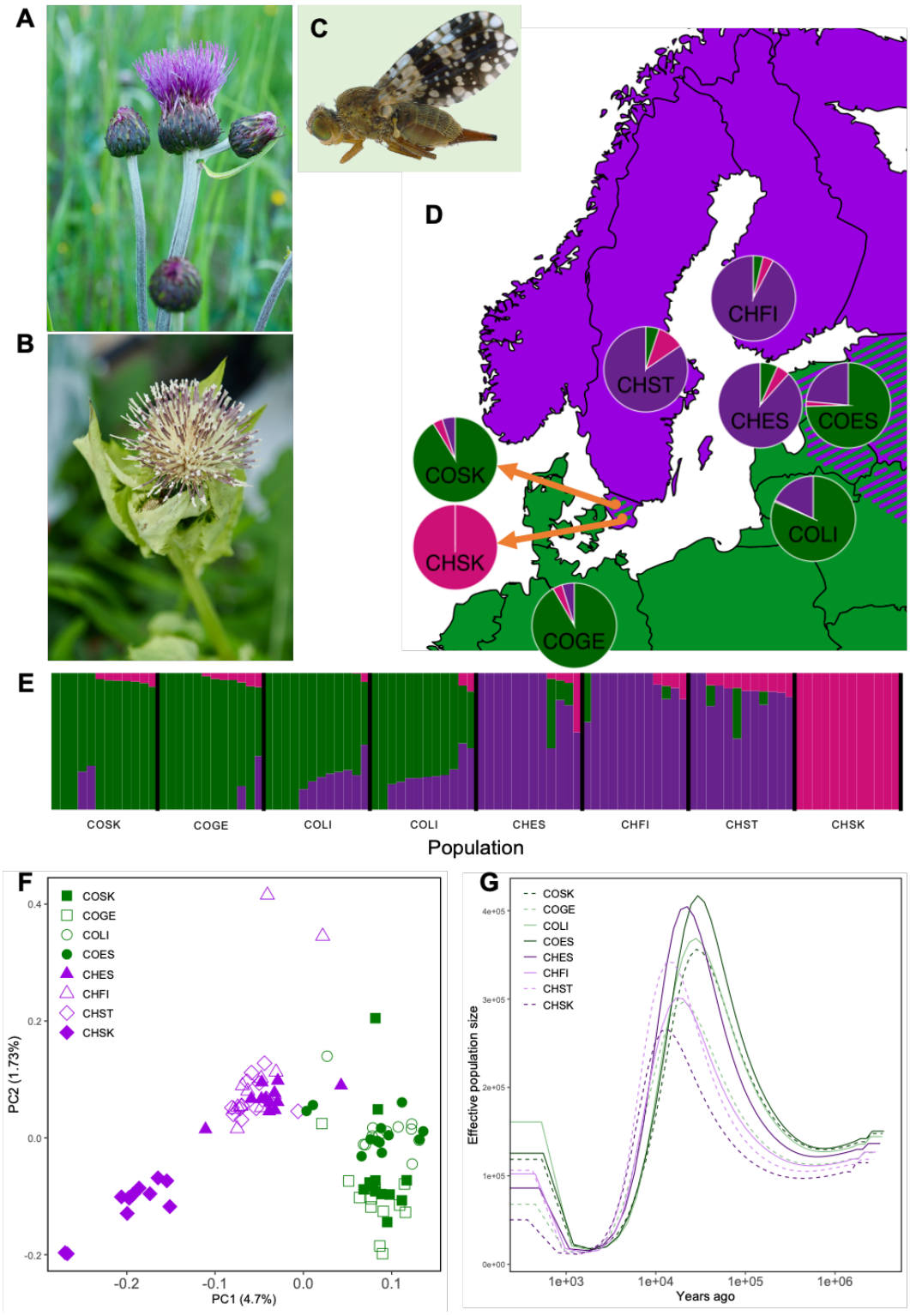

Fig. 1. Genomic divergence of *Tephritis conura* host races. Host races of *T. conura* specialize on *Cirsium heterophyllum* (CH; **A**) or *C. oleraceum* (CO; **B**) thistle buds. **(C)** A female *T. conura.* **(D)** Distribution and admixture proportions of populations sampled for this study. Purple represents regions of northern Europe where *C. heterophyllum* exists in allopatry, green where *C.*

677 *oleraceum* exists in allopatry, and striped where both host plants are found in sympatry. **(E)**
678 Admixture proportions estimated for K=3 (see Fig. S6A for K2-9). **(F)** Principal component (PC)
679 analysis of genetic differences among populations separates CH (purple) and CO (green)
680 individuals first PC axis (see Fig. S6B for a comparison of the second and third PC axes). **(G)**
681 MSMC analysis of demographic history.
682

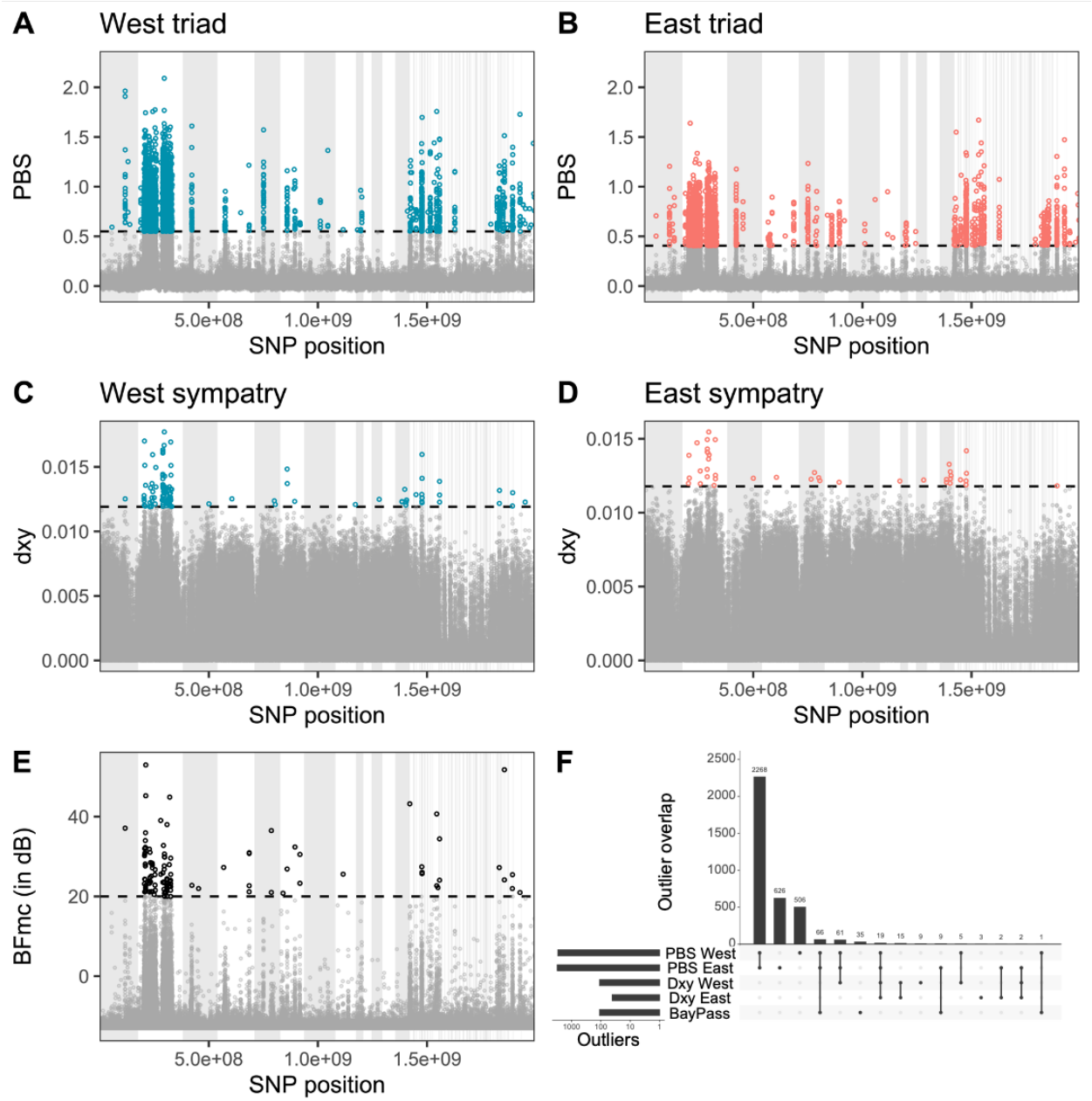

Fig. 2. Highly divergent genomic windows between host races based on **(A,B)** population branch statistics (PBS), **(C,D)** dxy, and **(E)** BayPass analyses for association with host race. Statistics were evaluated over 10kb non-overlapping windows. PBS outliers were greater than four standard deviations above the mean. Dxy comparisons were between CHSK and COSK in the west, CHES and COES in the easta and outliers were also greater than four standard deviations above the mean. BayPass outliers had an average BayesFactor over 20 (in decibans). **(F)** Overlap between outlier sets for eastern and western transects.

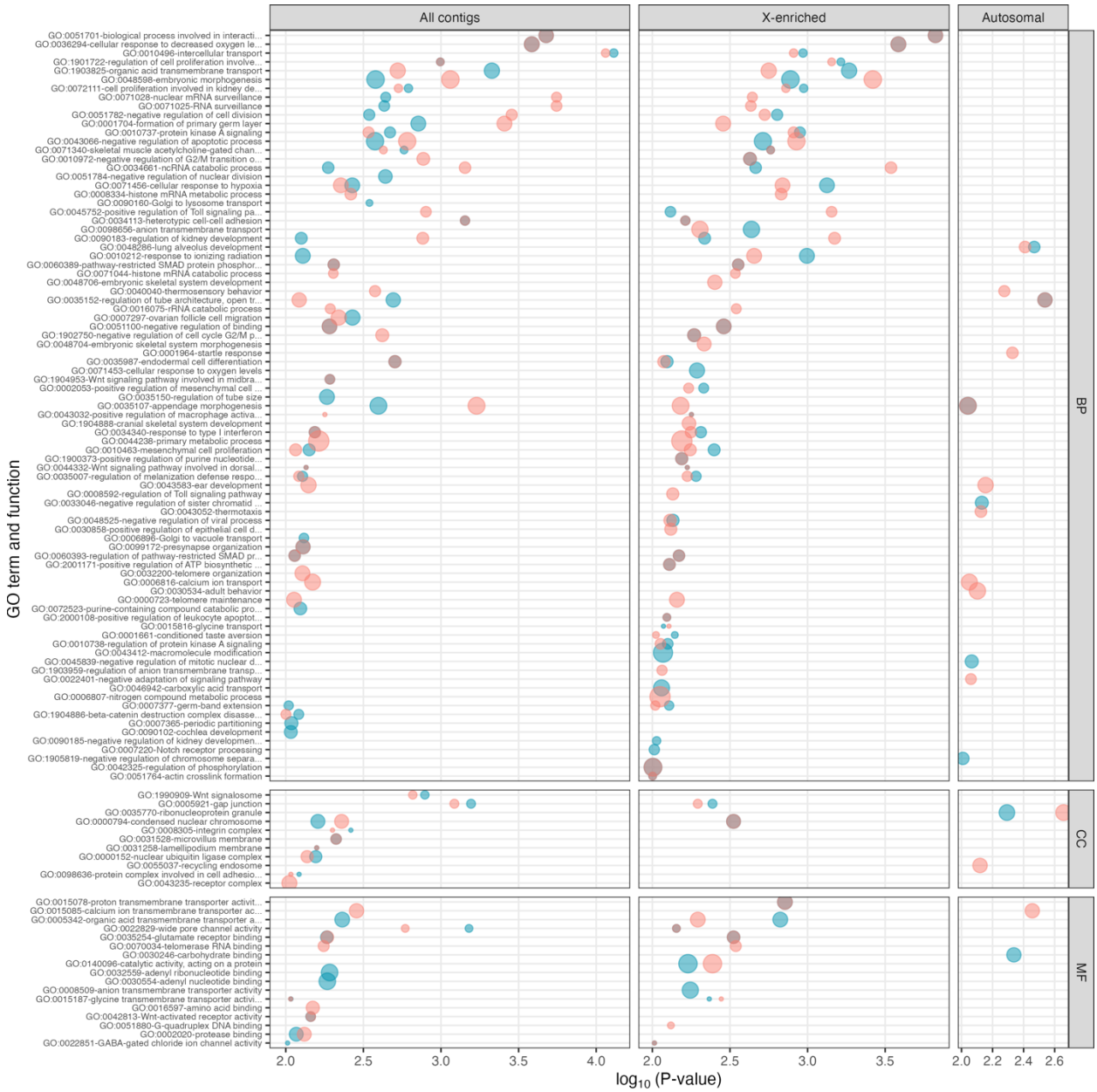

Fig. 3. Functional enrichment of outlier genes. Gene set enrichment was performed using genes (+/- 2000bp) overlapping 10kb outlier regions identified using PBS, Dxy or BayPass in the western (blue) and eastern (coral) comparisons. BayPass outliers were included in both east and west, as this analysis was performed on all populations. Point size is scaled to reflect the number of genes annotated with that GO term (log10-transformed).

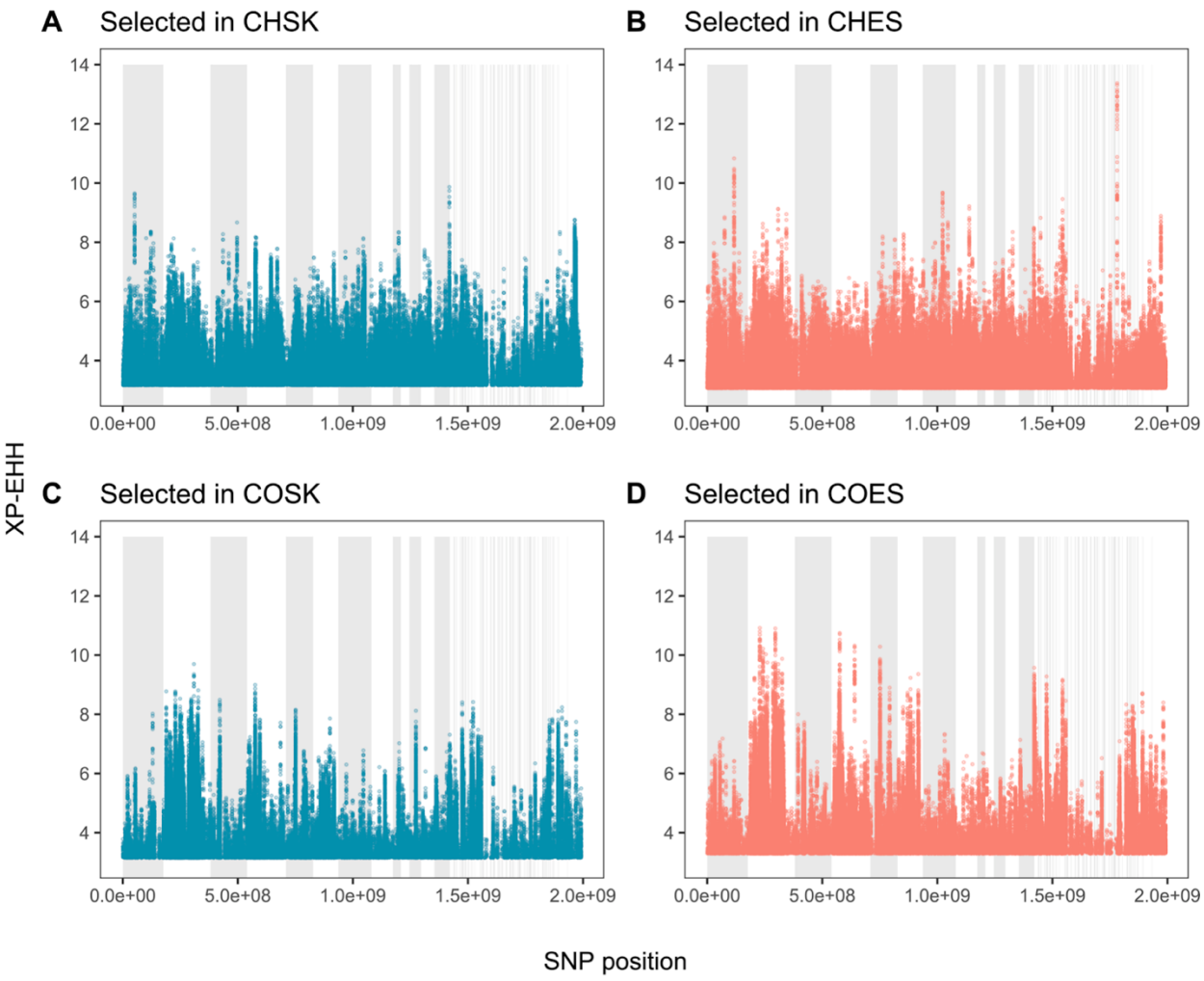

Fig. 4. Distribution of selected SNPs across the genome. Selection was estimated using extended haplotype statistics. Panels show top 1% of SNPs identified as under selection in each sympatric population when compared against the sympatric population of alternative host race (CHSK vs. COSK; CHES vs. COES). For all comparisons see Table S5.

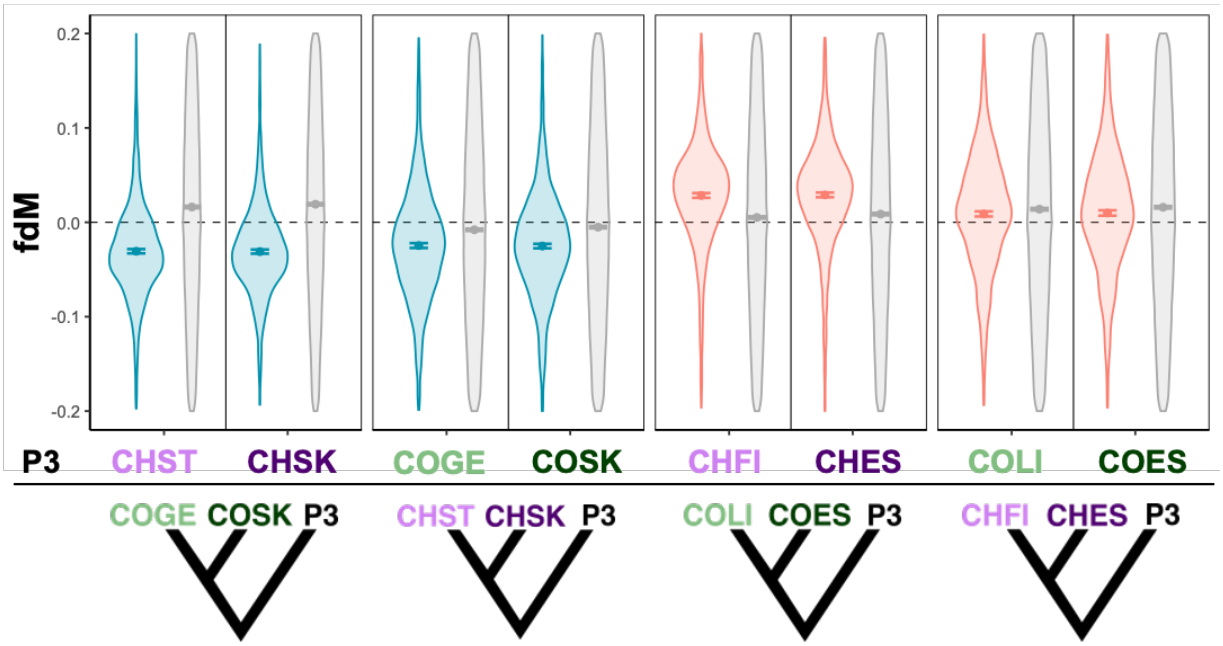

Fig. 5. Introgression in outlier windows compared to genome-wide. Introgression was compared between triads of *T. conura* populations (trees below x-axis) using f_{dM}. Values close to zero indicate little to no introgression, negative values indicate greater relative introgression between the left-most population and P3, while positive values indicate greater relative introgression between the middle population and P3, where P3 is one of the two populations above each tree.

712 **References**

713 Alexa, A. and J. Rahnenfuhrer. 2022. topGO: enrichment analysis for gene ontology. R package
714 version 2.48.0.

715 Alonge, M., L. Lebeigle, M. Kirsche, K. Jenike, S. Ou, S. Aganezov, X. Wang, Z. B. Lippman,
716 M. C. Schatz, and S. Soyk. 2022. Automated assembly scaffolding using RagTag elevates
717 a new tomato system for high-throughput genome editing. Genome Biology 23:258.

718 Aluja, M. and A. L. Norrbom. 2001. Fruit flies (*Tephritidae*): phylogeny and evolution of
719 behaviour. CRC Press, Boca Raton.

720 Anderson, S. A. S., H. López-Fernández, and J. T. Weir. 2023. Ecology and the origin of
721 nonephemeral species. Am. Nat. 0:000-000.

722 Anderson, S. A. S. and J. T. Weir. 2022. The role of divergent ecological adaptation during
723 allopatric speciation in vertebrates. Science 378:1214-1218.

724 Barbato, M., P. Orozco-terWengel, M. Tapio, and M. W. Bruford. 2015. SNeP: a tool to estimate
725 trends in recent effective population size trajectories using genome-wide SNP data. Front
726 Genet 6:109.

727 Berdan, E. L., T. Flatt, G. M. Kozak, K. E. Lotterhos, and B. Wielstra, eds. 2022. Genomic
728 architecture of supergenes: causes and evolutionary consequences [Theme Issue],
729 Philosophical Transactions of the Royal Society B-Biological Sciences, 377:1856.

730 Berlocher, S. H. and J. L. Feder. 2002. Sympatric speciation in phytophagous insects: moving
731 beyond controversy? Annu. Rev. Entomol. 47:773-815.

732 Bhatia, G., N. Patterson, S. Sankararaman, and A. L. Price. 2013. Estimating and interpreting FST:
733 the impact of rare variants. Genome Res 23:1514-1521.

734 Boag, P. T. and P. R. Grant. 1981. Intense natural selection in a population of Darwin's finches
735 (*Geospizinae*) in the Galápagos. Science 214:82-85.
736 Bohutínská, M., J. Vlček, S. Yair, B. Laenen, V. Konečná, M. Fracassetti, T. Slotte, and F. Kolář.
737 2021. Genomic basis of parallel adaptation varies with divergence in *Arabidopsis* and its
738 relatives. Proc. Natl. Acad. Sci. 118:e2022713118.
739 Bolnick, D. I. 2011. Sympatric speciation in threespine stickleback: why not? International Journal
740 of Ecology 2011:e942847.
741 Boulain, H., F. Legeai, J. Jaquiéry, E. Guy, S. Morlière, J.-C. Simon, and A. Sugio. 2019.
742 Differential expression of candidate salivary effector genes in pea aphid biotypes with
743 distinct host plant specificity. Frontiers in Plant Science 10.
744 Broad Institute. Picard Tools, http://broadinstitute.github.io/picard/.
745 Bush, G. L. 1969. Sympatric host race formation and speciation in frugivorious flies of genus
746 *Rhagoletis* (*Diptera, Tephritidae*). Evolution 23:237-251.
747 Butlin, R. K. and C. M. Smadja. 2018. Coupling, reinforcement, and speciation. Am. Nat. 191:155-
748 172.
749 Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden.
750 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.
751 Cantarel, B. L., I. Korf, S. M. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. Sánchez Alvarado,
752 and M. Yandell. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging
753 model organism genomes. Genome Res 18:188-196.
754 Charlesworth, B., J. L. Campos, and B. C. Jackson. 2018. Faster-X evolution: theory and evidence
755 from *Drosophila*. Mol. Ecol. 27:3753-3771.

756 Charlesworth, D. 2017. Evolution of recombination rates between sex chromosomes.
757 Philosophical Transactions of the Royal Society B: Biological Sciences 372:20160456.
758 Cheng, H., G. T. Concepcion, X. Feng, H. Zhang, and H. Li. 2021. Haplotype-resolved de novo
759 assembly using phased assembly graphs with hifiasm. Nature Methods 18:170-175.
760 Chevin, L.-M., G. Decorzent, and T. Lenormand. 2014. Niche dimensionality and the genetics of
761 ecological speciation. Evolution 68:1244-1256.
762 Christenhusz, M. J. M. and J. W. Byng. 2016. The number of known plants species in the world
763 and its annual increase. Phytotaxa 261:201-217.
764 Christenson, L. D. and R. H. Foote. 1960. Biology of fruit flies. Annu. Rev. Entomol. 5:171-192.
765 Craig, T. P., J. K. Itami, W. G. Abrahamson, and J. D. Horner. 1993. Behavioral evidence for host-
766 race formation in *Eurosta solidaginis*. Evolution 47:1696-1710.
767 Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T.
768 Keane, S. A. McCarthy, R. M. Davies, and H. Li. 2021. Twelve years of SAMtools and
769 BCFtools. Gigascience 10.
770 Delaneau, O., B. Howie, Anthony J. Cox, J.-F. Zagury, and J. Marchini. 2013. Haplotype
771 estimation using sequencing reads. The American Journal of Human Genetics 93:687-696.
772 Diegisser, T., J. Johannesen, C. Lehr, and A. Seitz. 2004. Genetic and morphological
773 differentiation in *Tephritis bardanae (Diptera: Tephritidae)*: evidence for host-race
774 formation. J. Evol. Biol. 17:83-93.
775 Diegisser, T., J. Johannesen, and A. Seitz. 2006a. The role of geographic setting on the
776 diversification process among *Tephritis conura (Tephritidae)* host races. Heredity 96:410-
777 418.

778 Diegisser, T., J. Johannesen, and A. Seitz. 2008. Performance of host-races of the fruit fly,
779 *Tephritis conura* on a derived host plant, the cabbage thistle *Cirsium oleraceum*:
780 Implications for the original host shift. J. Insect Sci. 8:1-6.
781 Diegisser, T., A. Seitz, and J. Johannesen. 2006b. Phylogeographic patterns of host-race evolution
782 in *Tephritis conura (Diptera: Tephritidae)*. Mol. Ecol. 15:681-694.
783 Diegisser, T., A. Seitz, and J. Johannesen. 2007. Morphological adaptation in host races of
784 *Tephritis conura*. Entomol. Exp. Appl. 122:155-164.
785 Egan, S. P., G. J. Ragland, L. Assour, T. H. Q. Powell, G. R. Hood, S. Emrich, P. Nosil, and J. L.
786 Feder. 2015. Experimental evidence of genome-wide impact of ecological selection during
787 early stages of speciation-with-gene-flow. Ecol. Lett. 18:817-825.
788 Faria, R., P. Chaube, H. E. Morales, T. Larsson, A. R. Lemmon, E. M. Lemmon, M. Rafajlović,
789 M. Panova, M. Ravinet, K. Johannesson, A. M. Westram, and R. K. Butlin. 2019a. Multiple
790 chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. Mol.
791 Ecol. 28:1375-1393.
792 Faria, R., K. Johannesson, R. K. Butlin, and A. M. Westram. 2019b. Evolving inversions. Trends
793 Ecol. Evol. 34:239-248.
794 Feder, J. L., S. H. Berlocher, and S. B. Opp. 1998. Sympatric host-race formation and speciation
795 in *Rhagoletis* (*Diptera: Tephritidae*): a tale of two species for Charles D.
796 Feder, J. L., C. A. Chilcote, and G. L. Bush. 1988. Genetic differentiation between sympatric host
797 races of the apple maggot fly *Rhagoletis pomonella*. Nature 336:61-64.
798 Feder, J. L., J. B. Roethele, K. Filchak, J. Niedbalski, and J. Romero-Severson. 2003. Evidence
799 for inversion polymorphism related to sympatric host race formation in the apple maggot
800 fly, *Rhagoletis pomonella*. Genetics 163:939-953.

801 Filchak, K. E., J. B. Roethele, and J. L. Feder. 2000. Natural selection and sympatric divergence
802 in the apple maggot *Rhagoletis pomonella*. Nature 407:739-742.
803 Flynn, J. M., R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, and A. F. Smit. 2020.
804 RepeatModeler2 for automated genomic discovery of transposable element families. Proc
805 Natl Acad Sci U S A 117:9451-9457.
806 Fontaine, M. C., J. B. Pease, A. Steele, R. M. Waterhouse, D. E. Neafsey, I. V. Sharakhov, X.
807 Jiang, A. B. Hall, F. Catteruccia, E. Kakani, S. N. Mitchell, Y.-C. Wu, H. A. Smith, R. R.
808 Love, M. K. Lawniczak, M. A. Slotman, S. J. Emrich, M. W. Hahn, and N. J. Besansky.
809 2015. Extensive introgression in a malaria vector species complex revealed by
810 phylogenomics. Science 347:1258524.
811 Fox, E. A., A. E. Wright, M. Fumagalli, and F. G. Vieira. 2019. ngsLD: evaluating linkage
812 disequilibrium using genotype likelihoods. Bioinformatics 35:3855-3856.
813 Gautier, M. 2015. Genome-wide scan for adaptive divergence and association with population-
814 specific covariates. Genetics 201:1555-1579.
815 Gow, J. L., C. L. Peichel, and E. B. Taylor. 2006. Contrasting hybridization rates between
816 sympatric three-spined sticklebacks highlight the fragility of reproductive barriers between
817 evolutionarily young species. Mol Ecol 15:739-752.
818 Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L.
819 Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind,
820 F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev.
821 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome.
822 Nature Biotechnology 29:644-652.

823 Grossmann, S., S. Bauer, P. N. Robinson, and M. Vingron. 2007. Improved detection of
824 overrepresentation of gene-ontology annotations with parent child analysis. Bioinformatics
825 23:3024-3031.
826 Guan, D., S. A. McCarthy, J. Wood, K. Howe, Y. Wang, and R. Durbin. 2020. Identifying and
827 removing haplotypic duplication in primary genome assemblies. Bioinformatics 36:2896-
828 2898.
829 Hardy, N. B. and S. P. Otto. 2014. Specialization and generalization in the diversification of
830 phytophagous insects: tests of the musical chairs and oscillation hypotheses. Proceedings
831 of the Royal Society B: Biological Sciences 281:20132960.
832 Headrick, D. H. and R. D. Goeden. 1998. The biology of nonfrugivorous tephritid fruit flies. Annu
833 Rev Entomol 43:217-241.
834 Hendry, A. P. 2009. Ecological speciation! Or the lack thereof? Can. J. Fish. Aquat. Sci. 66:1383-
835 1398.
836 Herrel, A., K. Huyghe, B. Vanhooydonck, T. Backeljau, K. Breugelmans, I. Grbac, R. Van
837 Damme, and D. J. Irschick. 2008. Rapid large-scale evolutionary divergence in
838 morphology and performance associated with exploitation of a different dietary resource.
839 Proc. Natl. Acad. Sci. 105:4792-4795.
840 Hollander, J., M. Lindegarth, and K. Johannesson. 2005. Local adaptation but not geographical
841 separation promotes assortative mating in a snail. Animal Behaviour 70:1209-1219.
842 Hooper, D. M., S. C. Griffith, and T. D. Price. 2019. Sex chromosome inversions enforce
843 reproductive isolation across an avian hybrid zone. Mol. Ecol. 28:1246-1262.
844 Hooper, D. M. and T. D. Price. 2017. Chromosomal inversion differences correlate with range
845 overlap in passerine birds. Nat. Ecol. Evol. 1:1526-1534.

846 Huerta-Cepas, J., D. Szklarczyk, D. Heller, A. Hernández-Plaza, S. K. Forslund, H. Cook, D. R.
847 Mende, I. Letunic, T. Rattei, Lars J. Jensen, C. von Mering, and P. Bork. 2018. eggNOG
848 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based
849 on 5090 organisms and 2502 viruses. Nucleic Acids Research 47:D309-D314.
850 Johannesson, B. and K. Johannesson. 1996. Population differences in behaviour and morphology
851 in the snail *Littorina saxatilis*: Phenotypic plasticity or genetic differentiation? Journal of
852 Zoology 240:475-493.
853 Johannesson, K. 2001. Parallel speciation: a key to sympatric divergence. Trends Ecol. Evol.
854 16:148-153.
855 Johannesson, K., M. Panova, P. Kemppainen, C. André, E. Rolán-Alvarez, and R. K. Butlin. 2010.
856 Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of
857 speciation. Philos Trans R Soc Lond B Biol Sci 365:1735-1747.
858 Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli, J. Johnson, R. Swofford, M.
859 Pirun, M. C. Zody, S. White, E. Birney, S. Searle, J. Schmutz, J. Grimwood, M. C. Dickson,
860 R. M. Myers, C. T. Miller, B. R. Summers, A. K. Knecht, S. D. Brady, H. Zhang, A. A.
861 Pollen, T. Howes, C. Amemiya, J. Baldwin, T. Bloom, D. B. Jaffe, R. Nicol, J. Wilkinson,
862 E. S. Lander, F. Di Palma, K. Lindblad-Toh, and D. M. Kingsley. 2012. The genomic basis
863 of adaptive evolution in threespine sticklebacks. Nature 484:55-61.
864 Keightley, P. D., U. Trivedi, M. Thomson, F. Oliver, S. Kumar, and M. L. Blaxter. 2009. Analysis
865 of the genome sequences of three *Drosophila melanogaster* spontaneous mutation
866 accumulation lines. Genome Res 19:1195-1201.

867 Kim, D., J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg. 2019. Graph-based genome
868 alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907-
869 915.
870 Koch, E. L., M. Ravinet, A. M. Westram, K. Johannesson, and R. K. Butlin. 2022. Genetic
871 architecture of repeated phenotypic divergence in *Littorina saxatilis* ecotype evolution.
872 Evolution 76:2332-2346.
873 Korneliussen, T. S., A. Albrechtsen, and R. Nielsen. 2014. ANGSD: analysis of next generation
874 sequencing data. BMC Bioinformatics 15:356.
875 Kovaka, S., A. V. Zimin, G. M. Pertea, R. Razaghi, S. L. Salzberg, and M. Pertea. 2019.
876 Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome
877 Biology 20:278.
878 Kulmuni, J., R. K. Butlin, K. Lucek, V. Savolainen, and A. M. Westram, eds. 2020. Towards the
879 completion of speciation: the evolution of reproductive isolation beyond the first barriers
880 [Theme issue], Philosophical Transactions of the Royal Society B-Biological Sciences,
881 375:1806.
882 Küpper, C., M. Stocks, J. E. Risse, N. dos Remedios, L. L. Farrell, S. B. McRae, T. C. Morgan, N.
883 Karlionova, P. Pinchuk, Y. I. Verkuil, A. S. Kitaysky, J. C. Wingfield, T. Piersma, K. Zeng,
884 J. Slate, M. Blaxter, D. B. Lank, and T. Burke. 2016. A supergene determines highly
885 divergent male reproductive morphs in the ruff. Nature Genetics 48:79-83.
886 Lackey, A. C. R. and J. W. Boughman. 2017. Evolution of reproductive isolation in stickleback
887 fish. Evolution 71:357-372.

888 Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and
889 population genetical parameter estimation from sequencing data. Bioinformatics 27:2987-
890 2993.
891 Li, H. and R. Durbin. 2009. Fast and accurate short read alignment with Burrows–Wheeler
892 transform. Bioinformatics 25:1754-1760.
893 Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R.
894 Durbin, and G. P. D. P. Subgroup. 2009. The sequence alignment/map format and
895 SAMtools. Bioinformatics 25:2078-2079.
896 Malinsky, M., M. Matschiner, and H. Svardal. 2021. Dsuite - fast D-statistics and related
897 admixture evidence from VCF files. Molecular Ecology Resources 21:584-595.
898 Mank, J. E., K. Nam, and H. Ellegren. 2009. Faster-Z evolution is predominantly due to genetic
899 drift. Mol. Biol. Evol. 27:661-670.
900 Montejo-Kovacevich, G., J. I. Meier, C. N. Bacquet, I. A. Warren, Y. F. Chan, M. Kucka, C.
901 Salazar, N. Rueda-M, S. H. Montgomery, W. O. McMillan, K. M. Kozak, N. J. Nadeau, S.
902 H. Martin, and C. D. Jiggins. 2022. Repeated genetic adaptation to altitude in two tropical
903 butterflies. Nat. Commun. 13:4676.
904 Mora, C., D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm. 2011. How many species are
905 there on earth and in the ocean? PLoS. Biol. 9:8.
906 Nallu, S., J. A. Hill, K. Don, C. Sahagun, W. Zhang, C. Meslin, E. Snell-Rood, N. L. Clark, N. I.
907 Morehouse, J. Bergelson, C. W. Wheat, and M. R. Kronforst. 2018. The molecular genetic
908 basis of herbivory between butterflies and their host plants. Nat. Ecol. Evol. 2:1418-1427.

909 Nielsen, R., T. Korneliussen, A. Albrechtsen, Y. Li, and J. Wang. 2012. SNP calling, genotype
910 calling, and sample allele frequency estimation from new-generation sequencing data.
911 PLoS One 7:e37558.
912 Nilsson, K. J., J. Ortega, M. Friberg, and A. Runemark. 2022. Non-parallel morphological
913 divergence following colonization of a new host plant. Evol. Ecol. 36:859-877.
914 Nosil, P. 2007. Divergent host plant adaptation and reproductive isolation between ecotypes of
915 *Timema cristinae* walking sticks. Am. Nat. 169:151-162.
916 Nosil, P. 2012. Ecological speciation. Oxford University Press, Oxford.
917 Nosil, P., B. J. Crespi, and C. P. Sandoval. 2002. Host-plant adaptation drives the parallel evolution
918 of reproductive isolation. Nature 417:440-443.
919 Nosil, P., L. J. Harmon, and O. Seehausen. 2009. Ecological explanations for (incomplete)
920 speciation. Trends Ecol Evol 24:145-156.
921 Nosil, P. and C. P. Sandoval. 2008. Ecological niche dimensionality and the evolutionary
922 diversification of stick insects. PLoS One 3:e1907.
923 Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster,
924 and D. Reich. 2012. Ancient admixture in human history. Genetics 192:1065-1093.
925 Pečnerová, P., G. Garcia-Erill, X. Liu, C. Nursyifa, R. K. Waples, C. G. Santander, L. Quinn, P.
926 Frandsen, J. Meisner, F. F. Stæger, M. S. Rasmussen, A. Brüniche-Olsen, C. Hviid Friis
927 Jørgensen, R. R. da Fonseca, H. R. Siegismund, A. Albrechtsen, R. Heller, I. Moltke, and
928 K. Hanghøj. 2021. High genetic diversity and low differentiation reflect the ecological
929 versatility of the African leopard. Curr Biol 31:1862-1871.e1865.

930 Perini, S., M. Rafajlovic, A. M. Westram, K. Johannesson, and R. K. Butlin. 2020. Assortative
931 mating, sexual selection, and their consequences for gene flow in *Littorina*. Evolution
932 74:1482-1497.
933 Quinlan, A. R. and I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic
934 features. Bioinformatics 26:841-842.
935 Rice, W. R. and E. E. Hostert. 1993. Laboratory experiments on speciation: what have we learned
936 in 40 years? Evolution 47:1637-1653.
937 Romstock-Volkl, M. 1997. Host race formation in *Tephritis conura*: determinants from three
938 trophic levels. Ecol. Stud. 130:21-38.
939 Salmón, P., A. Jacobs, D. Ahrén, C. Biard, N. J. Dingemanse, D. M. Dominoni, B. Helm, M.
940 Lundberg, J. C. Senar, P. Sprau, M. E. Visser, and C. Isaksson. 2021. Continent-wide
941 genomic signatures of adaptation to urbanisation in a songbird across Europe. Nat.
942 Commun. 12:2983.
943 Schaal, S. M., B. C. Haller, and K. E. Lotterhos. 2022. Inversion invasions: when the genetic basis
944 of local adaptation is concentrated within inversions in the face of gene flow. Philos Trans
945 R Soc Lond B Biol Sci 377:20210200.
946 Schiffels, S. and R. Durbin. 2014. Inferring human population size and separation history from
947 multiple genome sequences. Nature Genetics 46:919-925.
948 Schluter, D. 2009. Evidence for ecological speciation and its alternative. Science 323:737-741.
949 Schluter, D. and L. H. Rieseberg. 2022. Three problems in the genetics of speciation by selection.
950 Proc. Natl. Acad. Sci. 119:e2122153119.

951 Seitz, A. and M. Komma. 1984. Genetic polymorphism and its ecological background in tephritid
952 populations (*Diptera: Tephritidae*). Pp. 143-158 *in* K. Wöhrmann, and V. Loeschcke, eds.
953 Population Biology and Evolution. Springer Berlin Heidelberg, Berlin, Heidelberg.
954 Serrato-Capuchina, A. and D. R. Matute. 2018. The role of transposable elements in speciation.
955 Genes 9.
956 Shih, P.-Y., A. Sugio, and J.-C. Simon. 2023. Molecular mechanisms underlying host plant
957 specificity in aphids. Annu. Rev. Entomol. 68:431-450.
958 Singh, K. S., B. J. Troczka, A. Duarte, V. Balabanidou, N. Trissi, L. Z. Carabajal Paladino, P.
959 Nguyen, C. T. Zimmer, K. M. Papapostolou, E. Randall, B. Lueke, F. Marec, E. Mazzoni,
960 M. S. Williamson, A. Hayward, R. Nauen, J. Vontas, and C. Bass. 2020. The genetic
961 architecture of a host shift: an adaptive walk protected an aphid and its endosymbiont from
962 plant chemical defenses. Science Advances 6:eaba1070.
963 Skotte, L., T. S. Korneliussen, and A. Albrechtsen. 2013. Estimating individual admixture
964 proportions from next generation sequencing data. Genetics 195:693-702.
965 Smit, A. F., R. Hubley, and P. Green. 2015. RepeatMaster Open-4.0,
966 http://www.repeatmasker.org.
967 Sproul, J. S., S. Hotaling, J. Heckenhauer, A. Powell, A. M. Larracuente, J. L. Kelley, S. U. Pauls,
968 and P. B. Frandsen. 2022. Repetitive elements in the era of biodiversity genomics: insights
969 from 600+ insect genomes. bioRxiv:2022.2006.2002.494618.
970 Stankowski, S. and M. Ravinet. 2021. Defining the speciation continuum. Evolution 75:1256-
971 1273.
972 Stuart, Y. E., T. Veen, J. N. Weber, D. Hanson, M. Ravinet, B. K. Lohman, C. J. Thompson, T.
973 Tasneem, A. Doggett, R. Izen, N. Ahmed, R. D. H. Barrett, A. P. Hendry, C. L. Peichel,

974 and D. I. Bolnick. 2017. Contrasting effects of environment and genetics generate a
975 continuum of parallel evolution. Nat. Ecol. Evol. 1:0158.
976 Szpiech, Z. A. and R. D. Hernandez. 2014. selscan: an efficient multithreaded program to perform
977 EHH-based scans for positive selection. Mol Biol Evol 31:2824-2827.
978 Taylor, E. B., J. W. Boughman, M. Groenenboom, M. Sniatynski, D. Schluter, and J. L. Gow.
979 2006. Speciation in reverse: morphological and genetic evidence of the collapse of a three-
980 spined stickleback *(Gasterosteus aculeatus)* species pair. Mol. Ecol. 15:343-355.
981 Tuttle, Elaina M., Alan O. Bergland, Marisa L. Korody, Michael S. Brewer, Daniel J. Newhouse,
982 P. Minx, M. Stager, A. Betuel, Zachary A. Cheviron, Wesley C. Warren, Rusty A. Gonser,
983 and Christopher N. Balakrishnan. 2016. Divergence and functional degradation of a sex
984 chromosome-like supergene. Curr. Biol. 26:344-350.
985 UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids
986 Res 49:D480-d489.
987 van Schooten, B., J. Meléndez-Rosa, S. M. Van Belleghem, C. D. Jiggins, J. D. Tan, W. O.
988 McMillan, and R. Papa. 2020. Divergence of chemosensing during the early stages of
989 speciation. Proc. Natl. Acad. Sci. 117:16438-16447.
990 Vicoso, B. and D. Bachtrog. 2015. Numerous transitions of sex chromosomes in Diptera. PLoS.
991 Biol. 13:e1002078.
992 Vidal, M. C. and S. M. Murphy. 2018. Bottom-up vs. top-down effects on terrestrial insect
993 herbivores: a meta-analysis. Ecol. Lett. 21:138-150.
994 Villoutreix, R., C. F. de Carvalho, V. Soria-Carrasco, D. Lindtke, M. De-la-Mora, M. Muschick,
995 J. L. Feder, T. L. Parchman, Z. Gompert, and P. Nosil. 2020. Large-scale mutation in the
996 evolution of a gene complex for cryptic coloration. Science 369:460-466.

997 Wellenreuther, M. and L. Bernatchez. 2018. Eco-evolutionary genomics of chromosomal
998 inversions. Trends Ecol. Evol. 33:427-440.

999