Genomic islands of divergence between *Drosophila yakuba* subspecies predominantly overlap with chromosomal inversions

Erina A. Ferreira[1,2], Cathy C. Moore[3], David Ogereau[1], Arnaud Suwalski[2], Rebekah L. Rogers[3], Amir Yassin[1,2]

[1]Laboratoire Évolution, Génomes, Comportement et Écologie, CNRS, IRD, Université Paris-Saclay – Institut Diversité, Écologie et Évolution du Vivant (IDEEV), 91190 Gif-sur-Yvette, France.

[2]Institut Systématique, Evolution, Biodiversité (ISyEB), Muséum National d'Histoire Naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, CP50, 75005 Paris, France.

[3]Department of Bioinformatics and Genomics, University of North Carolina, Charlotte NC, USA.

* Corresponding author: Amir Yassin (amir.yassin@universite-paris-saclay.fr)

Laboratoire Évolution, Génomes, Comportement et Écologie, CNRS, IRD, Université Paris-Saclay – Institut Diversité, Ecologie et Evolution du Vivant (IDEEV), 12 route 128, 91190 Gif-sur-Yvette, France.

Abstract

During the early stages of speciation, genetic differences tend to accumulate at certain regions of the genome leading to the formation of genomic islands of divergence (GIDs). This pattern may be due to selection and/or difference in the rate of recombination. Here, we investigate the possible causes of GIDs in *Drosophila yakuba mayottensis*, a subspecies of *Drosophila yakuba* specializing on toxic noni (*Morinda citrifolia*) fruits on the island of Mayotte. Field collection reconfirmed the association of this subspecies with noni fruits. Population genomics did not find evidence for genetic structure on the island, but fine-scale analysis comparing these sequences to two mainland *D. y. yakuba* populations identified multiple structured genomic regions. We assembled a new genome for *D. y. mayottensis* using a combination of long and short reads. We identified six chromosomal inversions in high recombining regions, of which two were unique to *D. y. mayottensis*. Genome-wide differentiation analyses revealed that 92% of outlier genes overlapped with the inversions. We also compared population genomics sequences of *Drosophila sechellia*, the only other noni-specialist, to its generalist homosequential relatives of the *simulans* clade. Although we identified several common outlier genes between *D. sechellia* and *D. y. mayottensis*, we did not detect traces of GIDs in this clade. Thus, different genomic architectures can underlie complex convergence in response to common selection in independent lineages, leading to distinct genomic divergence patterns.

Keywords: ecological speciation; evolutionary predictability; population genomics; genetic parallelism; insect-plant interactions.

Introduction

During the early stages of speciation, genetic differences tend to accumulate more rapidly in certain genomic regions which can lead to the formation of genomic islands of divergence (GIDs) (Feder and Nosil 2009; Wolf and Ellegren 2017). Those islands were originally hailed as result of divergent selection acting on genes that may underlie adaptation to distinct environment and/or reproductive isolation (Turner et al. 2005). However, a plethora of mechanisms may also lead to heterogeneous patterns of genetic differentiation across the genome even in the absence of divergent selection. Most importantly, variation in recombination rate, due to heterochromatin content or structural variation, strongly affects the rate of the accumulation of genetic differences in natural populations (Feder and Nosil 2009; Cruickshank and Hahn 2014; Quilodrán et al. 2019; Huang et al. 2020). Chromosomal inversions are among the most common forms of structural variation, and a number of studies have unraveled an overlap between inversions and GIDs (Feder and Nosil 2009; Michel et al. 2010; Ellegren et al. 2012; Sodeland et al. 2016; Huang et al. 2020) . Inversions often suppress recombination hence reducing the rate by which slightly-deleterious mutations could be purged out from the population (Berdan et al. 2019; Faria et al. 2019; Jay et al. 2021). Conversely, the reduction of recombination could conserve one or more favorable allelic combinations for survival in contrasting environments, hence becoming targets of diverging selection. Untangling the complex interactions of selection and inversions in driving GIDs and ultimately speciation therefore requires proper understanding of the genomic architecture of the study model and its natural variation.

In herbivorous insects, which constitute a large portion of the biodiversity of the planet, speciation is usually associated with host shifts or changes in the width of the ecological niche, e.g., specialization (Hébert et al. 2016; Forbes et al. 2017). The complexity of this phenomenon, which should involve the simultaneous evolution of genes underlying multiple behavioral and physiological phenotypes with an ultimate impact on reproductive isolation, has long been a major area of research (Forbes et al. 2017). One of the best well-characterized systems at the genomic level is the specialization of the fly *Drosophila sechellia* on toxic fruits of noni (*Morinda*

66 *citrifolia*) on the Seychelles islands in the Indian Ocean. *Drosophila sechellia* is a member of the
67 *simulans* species complex, which comprises two additional species, namely the cosmopolitan
68 *Drosophila simulans* and *Drosophila mauritiana*, which is endemic to Mauritius. Unlike *D. sechellia*,
69 both species are generalists and are intolerant to noni main toxins, octanoic acid causing death
70 and hexanoic acid causing reversible coma (Legal et al. 1994). The three species represent a sister
71 clade to *Drosophila melanogaster*. This facilitated the extension of the unique genetic and genomic
72 tools from this paragmatic species to the study of *D. sechellia* specialization. Besides, the three
73 species show partial reproductive isolation and are karyologically homosequential, making them
74 ideal for quantitative trait loci (QTLs) mapping through interspecific crosses (Auer et al. 2021). In
75 spite of these numerous advantages, the divergence of *D. sechellia* from its generalist sister species
76 *D. simulans* is old enough, estimated at around 89.000 years ago (Schrider et al. 2018); its entire
77 genome has accumulated so many differences that traces of possible GIDs during the early stages
78 of divergence may be difficult to observe. Besides, a particularly low level of genetic diversity has
79 been detected in this species (Legrand et al. 2009; Legrand et al. 2011; Schrider et al. 2018),
80 further complicating population genomics analyses.

81 In January 2013, we discovered a second case of a *Drosophila* species associated with toxic
82 noni in Mayotte, another island in the Indian Ocean. This was a population of *Drosophila yakuba*,
83 which had higher preference for and survival on noni unlike its continental conspecifics (Yassin
84 et al. 2016). The noni-specialist population has likely diverged from generalist continental
85 populations nearly 28,000 years ago and specialization was accompanied by precopulatory
86 isolation between Mayotte and continental flies, earning them subspecific status as
87 *D. y. mayottensis* and *D. y. yakuba*, respectively (Yassin et al. 2016; Yassin 2017). Genomic
88 comparisons between this single population of *D. y. mayottensis* and two *D. y. yakuba* continental
89 populations identified multiple GIDs, widespread across the genome, consistent with theoretical
90 expectations, with many GIDs overlapping with large noni survival QTLs in *D. sechellia* (Yassin,
91 Debat, et al. 2016). *Drosophila yakuba* is a member of the *melanogaster* species subgroup, and it is
92 thought to have diverged from *D. sechellia* nearly 7 million years ago (Suvorov et al. 2022). The

93 two species therefore represent an ideal case of two time points along an 'ecological speciation
94 continuum', wherein two independent lineages specialize on the same host. In spite of those
95 similarities, *D. yakuba* is a species where, unlike the *simulans* species complex, multiple
96 polymorphic chromosomal inversions have been identified in continental populations
97 (Lemeunier and Ashburner 1976). Several GIDs in *D. y. mayottensis* may therefore be result from
98 such inversions.

99 In this paper, we report the population genomics analysis of three additional populations
100 of *D. y. mayottensis* all associated with noni on Mayotte. We identified 18 GIDs distinguishing the
101 three populations from continental *D. y. yakuba*. We assembled a genome for *D. y. mayottensis* and
102 identified multiple inversions that predominantly overlap with the GIDs. We also performed
103 population genomics analyses in the three species of the *simulans* complex. Despite minor overlap
104 between outlier genes in both noni-specialist taxa that may indicate genetic parallelism, our
105 results show that their potential genic convergence was attained through distinct genomic
106 architectures.

107

108 **Results**

109 Drosophila yakuba mayottensis *is strictly associated with noni on Mayotte*

110 In April 2017, we returned to Mayotte to better investigate the association between
111 *D. y. mayottensis* and noni. Out of 14 collection sites, we identified noni trees in 9 locations, mostly
112 near the coastline. However, the trees were not always at the same fruiting stage, with fallen ripe
113 fruits found at only three locations, namely Soulou (the same locality for the 2013 sample,
114 hereafter SOU), M'tsangamouji (MTS) in the humid northwest and Bambo Est (BE) in the arid
115 southeast (Figure 1A). *D. y. mayottensis* flies were collected through net sweeping or direct
116 aspiration over fallen, rotting noni fruits only at those three locations. No single *D. y. mayottensis*
117 fly was collected using fermenting banana traps at those sites or at any other site on the island,
118 although the noni trees at Bambo Est were present amid a banana plantation. On continental
119 Africa, such traps would have collected hundreds of *D. y. yakuba*. The total number of collected

120 *D. y. mayottensis* was low, ~200 flies throughout the 10-days duration of the expedition, compared
121 to ~600 flies that were collected during the last 2 days of the January 2013 at a single location
122 (David et al. 2014). This indicates a strong seasonality in the abundance of the species, consistent
123 with other observations on *D. y. yakuba* on the continent (Prigent et al. 2013), and highlights the
124 necessity of year-round investigations. We also observed that flies were more abundant on noni
125 fruits that have been nibbled and thrown to the ground by the brown lemur (*Eulemur fulvus*) than
126 on intact fallen fruits (Figure 1B).
127
128 *Pooled and isofemale line sequencing of the three Mayotte populations reveal low level of*
129 *differentiation on the island*
130 We established 26, 22 and 29 isofemale lines from wild-caught flies from the three
131 collection sites, *i.e.,* BE, SOU and MTS, respectively. As with the 2016 study (Yassin et al. 2016), we
132 sequenced a single F_1 fly per line in three pools according to the populations. The level of genetic
133 differentiation in SOU between the 2013 and 2017 samples was low (F_{ST} = 0,0272), as well as the
134 level of differentiation between the three 2017 geographical samples (F_{ST} = 0.0290 ± 0.0057). On
135 continental Africa, we estimated the level of genetic differentiation between two *D. y. yakuba*
136 populations from Kenya (NY) and Cameroon (CY) that were sequenced by Rogers et al. (2014) to
137 be F_{ST} = 0.0609, a value that is slightly lower from that distinguishing both populations from the
138 three Mayotte samples (F_{ST} = 0.0650 ± 0.015). Because F_{ST} estimates from Pool-Seq data could be
139 slightly biased (Hivert et al. 2018), we also sequenced 7, 6 and 6 individual isofemale lines from
140 BE, SOU and MTS, respectively. We found a strong correspondence between F_{ST} estimates from
141 both approach (F_{ST} = 0.0343 ± 0.0062 between the Mayotte populations and 0.0698 ± 0.0162
142 between Mayotte and the continent).
143
144 *Fine-scale genetic structure analysis identifies 12 structured tracts in* D. yakuba
145 We used the individual lines sequences to investigate genome-wide levels of population
146 genetic structure within and between *D. yakuba* subspecies. Following Reilly's (2020) study on

147 the two continental populations, we run the MavericK ver. 1.0 software package (Verity and
148 Nichols 2016) to infer for each 100-kb window of the v1.05 reference genome
149 (http://ftp.flybase.net/genomes/Drosophila_yakuba/) the likeliest number of demes (*k*) among
150 three possibilities 1, 2 and 3. 854 (72%), 272 (23%) and 59 (5%) of the windows had 1, 2 and 3
151 demes, respectively. Visual depiction of the distribution of the demes among the lines across the
152 genome identified structured genomic regions (
153 Figure 1C). Based on our previous determination of low-recombining genomic regions
154 from the number of segregating sites in CY, *i.e.* the most polymorphic *D. y. yakuba* population
155 inhabiting the ancestral range (CY), we identified 12 structured regions falling in normally high
156 recombining regions (Table S1). Remarkably, the regions greatly differed in length, with 6 regions
157 only encompassing from 1-2 windows (*i.e.* 100-200 kb) and six regions going from 0.7 to 11.6 Mb
158 long. The longest structured regions were on chromosome 2. Of which two regions (S3 and S4)
159 fell within the cytogenetic limits of the *2Lh* chromosomal inversion according to Lemeunier and
160 Ashburner (1976; considered new inversion by Reilly (2020)). The longest region, S7, almost
161 covered the entirety of the high recombining region of chromosomal arm 2R. It was the only
162 region with windows with 3 demes and it encompassed the cytogenetic limits of three large
163 inversions, *2Rj*, *2Rk* and *2Rn*, that were bioinformatically delimited by Reilly (2020). We noted,
164 however, that for those 2L and 2R regions, no structure among *D. y. mayottensis* lines was
165 observed, indicating that this subspecies is likely monomorphic for two inversions that are
166 polymorphic in *D. y. yakuba*. By contrast, all the remaining structured regions mostly
167 distinguished *D. y. mayottensis* from *D. y. yakuba* populations. We also noted other genomic
168 regions distinguishing the two species falling in the low recombining long telomeric region of
169 chromosome X and on the boundaries of the centromeric region of chromosome 3.
170
171 *Novel genome assembly identifies subspecies-specific inversions in* D. y. mayottensis
172 To investigate whether the remaining genomic regions distinguishing Mayotte from
173 continental flies fall within specific inversions, we assembled a new genome from a laboratory

174 mass culture of *D. y. mayottensis* that was formed from the pooling of isofemale lines collected in
175 Soulou in 2013. The assembly was made following a hybrid strategy that combined Oxford
176 Nanopore long read and Illumina HiSeq short read sequences under the MaSurCa software
177 package v3.4.0 (Zimin et al. 2013). The Nanopore MinION run has generated 143 587 reads with
178 a N50 of 33,25 kb and with the final assembly we had obtained 563 fragments with an N50 of
179 6,9Mb, and a BUSCOv5 score of 99,5% (Manni et al. 2021). We aligned the *D. y. mayottensis*
180 assembly on a novel assembly of the West-African reference *D. y. yakuba* strain that was
181 assembled using the same hybrid Nanopore-Illumina approach (Miller et al. 2018; David et al.
182 2022) (Figure 2A). This new assembly corrects the order of genes falling between 8,586,415 and
183 11,101,068 on chromosomal arm 2L (Figure S1). We also aligned the *D. y. mayottensis* assembly
184 on a PiggyBac long-read assembly for the Kenyan NY73PB strain (NCBI Assembly
185 GCA_016746335.2), which differed from the reference genome in being homozygous for different
186 karyotypes of the 2L and 2R inversions, namely *2Lh* and *2Rj* (Reilly 2020) (Figure 2B). From these
187 dual comparisons we were able to identify 11 inversions in low and high recombining regions.
188 Five of them were specific to *D. y. mayottensis*.
189 Nine *D. y. mayottensis* scaffolds aligned to the X chromosome. The longest scaffold
190 (scf7180000002491, 8090556 bp) covered the low recombining subtelomeric region. No signals
191 of inversions or breakpoints were detected on this scaffold. On the contrary, evidence of inversion
192 was detected in the middle of the scaffold scf7180000003411, which spans from X:9696909 to
193 the end of the scaffold at X:10135060 bp on the reference v1.05 genome, and which overlaps with
194 the short region strongly differentiating the two subspecies (
195 Figure 1C). A single breakpoint with inverted gene order was detected at this scaffold. We
196 were however unable to identify the second breakpoint of this inversion, which we call *Xtb*, on the
197 other scaffolds. The only other aberration that we detected was a connection between regions
198 X:12576920-13477771 and X:17070823-20685121 on the scf7180000002569 scaffold. Genes on
199 the two connected regions followed the same order as on the reference genome, which could
200 result from double inversions. However, we do not find population genetic support for such

201 hypothetical double inversions. It is therefore possible that the second breakpoint of the inversion
202 we detected at X:8671033 fell at intermediate regions between the scaffolds that we were not able
203 to assemble or align. No previous inversions on chromosome X were ever detected in *D. y. yakuba*,
204 and there is no evidence for genetic structure on this chromosome in the continental populations
205 (Figure 1C).

206 For chromosome 2, which is the most polymorphic in *D. y. yakuba*, alignment on both the
207 reference and NY73 genomes confirmed the population genomic analysis suggestion that
208 *D. y. mayottensis* was fixed for the *2Rj* inversion. Indeed, the breakpoints of the three common
209 inversions on 2R all fell within a single scaffold (scf7180000002462), which showed consistent
210 patterns of alignment with the reference genome, homozygous for *2Rn*, and the homokaryotpic
211 NY73, homozygous for *2Rj*. On chromosomal arm 2L, *D. y. mayottensis* was homokaryotypic with
212 the reference genome for inversion *2Lh*, and distinct from NY73 which is homozygous for
213 inversion *2Ll*. Nonetheless, we found an aberrant connection on the longest scaffold
214 (scf7180000002456) of this arm connecting regions 2L:14779-8423810 and 2L:10283568-
215 20352247 of the reference genome with opposite gene order at the later region. We, however,
216 note that the region at the left breakpoint is situated on a misassembled region on the 1.05
217 reference genome (see above; Figure S1). Indeed, the breakpoint at 2L:20352247 overlapped with
218 a genomic region differentiating *D. y. mayottensis* from the continent (
219 Figure 1C). We did not detect a second breakpoint on chromosomal arm 2L. However, a scaffold
220 near the centromeric region of chromosome 2R also showed evidence for a breakpoint. Because
221 the two breakpoints are separated by the highly repetitive centromeric scaffolds, with many of
222 such scaffolds also harboring evidence for inversion breakpoints, we cannot define at this stage
223 with certainty the boundaries of the unique 2L inversion in *D. y. mayottensis*, which we call *2Lo*.

224 Chromosome 3 is the least polymorphic autosome in *D. y. yakuba*. Lemeunier and
225 Ashburner (1976) found two inversions, *3Lj* and *3Rh*, at low frequencies mostly in West-African
226 populations but none was found between CY and NY populations. *D. y. mayottensis* had the
227 standard configuration found in the reference and NY73b genomes for those inversions. In

228 *D. y. mayottensis*, we detected two novel subcentromeric inversion breakpoints, on scaffold
229 scf7180000003504 on chromosomal arm 3L and on scaffold scf7180000003512 on chromosomal
230 arm 3R. Fine-scale genetic structure analyses did not suggest an association between the two
231 breakpoints and signal of subspecific differentiation was detected only on the 3R breakpoint,
232 suggesting the two breakpoints to likely be independent. As with chromosome 2, several
233 breakpoints in single scaffolds of the centromeric region of chromosome 3 were also detected.
234
235 *Genomic islands of divergence between* D. yakuba *subspecies overlap with inversions*
236 We conducted Population Branch Excess (*PBE*) analysis (Yassin et al. 2016) to identify
237 genomic islands of divergence specifically distinguishing *D. y. mayottensis* from the two
238 continental populations CY and NY. Because of the lack of genetic structure among the three
239 populations of *D. y. mayottensis* (
240 Figure 1C), we pooled either the individual lines sequences or the Pool-Seq sequences of
241 the three populations in a single population, referring to them as Pool of Individuals (PoI) and
242 Pool of Pools (PoP), respectively. Results did not substantially differ between the two methods
243 and we will show here PBE analysis on the PoP data. For sake of comparison with the three species
244 of the *simulans* complex (see below), we conducted the PBE analyses at each gene ± 2 kb. *PBE* is a
245 standardized version of the Population Branch Statistic (*PBS*), which was proposed to identify
246 population-specific allelic changes from F_{ST} pairwise comparisons between three populations (Yi
247 et al. 2010). *PBE* measures how *PBS* at a particular gene deviate from the chromosome-wide
248 median *PBS* divided by the median F_{ST} estimate from the two non-focal populations. Therefore,
249 *PBE* is less sensitive than *PBS* to genome-wide variation in polymorphism levels.
250 The *PBE* analyses showed 18 genomic islands of divergence (GIDs) with genes falling
251 above the 97.5% or below the 2.5% quantiles in *D. y. mayottensis* (Figure 3A; Table S1). Of these,
252 7 and 8 GIDs had predominantly positive or negative *PBE* values, respectively. Two GIDs had both
253 positive and negative *PBE* values. 12 out of the 18 GIDs fell within or were adjacent to
254 chromosomal inversions breakpoints, amounting to ~92% of the sum of the lengths of all GIDs.

255 Among these, the only two positive GIDs (GID1 and 8) fell within the *Xtb* and *2Lo* inversions
256 specific to *D. y. mayottensis*, respectively. Five negative GIDs fell at the breakpoints of the *2Lh* and
257 *2Ll* (GID3-6), and the breakpoints between *2Rn* and *2Rj/k* (GID10 and 13). Their negative *PBE*
258 values agree with the fact that the most widespread of those inversions were fixed in
259 *D. y. mayottensis*, while being polymorphic between the two *D. y. yakuba* populations. Two GIDs,
260 both falling within the polymorphic chromosomal arm 2R had positive *PBE* values at one side and
261 negative on the other. The positive *PBE* sides define the breakpoints between the *2Rj* inversion
262 (fixed in Mayotte) and the *2Rk* inversion (predominant in NY), since the reference genome is
263 homozygous for the *2Rn* inversion. Five out of the six remaining GIDs, which did not co-localized
264 with inversions, had positive *PBE* values (GID2 on 2L, GID14 on 2R and GID16-18 on 3R). Because
265 those GIDs do not fall within low recombining telomeric and centromeric region or in
266 chromosomal inversions, their high positive *PBE* values are likely due to selection specific to
267 *D. y. mayottensis*.
268
269 *Population genomics analyses in* D. sechellia *unravel distinct divergence patterns with*
270 D. y. mayottensis
271 For sake of comparison, we aligned on the reference chromosome v.2.02 of *D. simulans*
272 (Flybase ; Thurmond et al. 2019) recently published population genomics data for the three species
273 of the *simulans* complex, namely *D. mauritiana* (Nolte et al. 2012), *D. simulans* (Rogers et al. 2014)
274 and *D. sechellia* (Schrider et al. 2018). We conducted the *PBE* analysis on the topology
275 (*mauritiana*,(*simulans*,*sechellia*)) suggested by Cruickshank and Hahn (2014). In contrast to
276 *D. y. mayottensis*, no distinct genomic island of divergence was observed in *D. sechellia*; variation
277 in *PBE* values were almost equally distributed across the genome (Figure 3B). Applying the same
278 stringent threshold, we used for *D. y. mayottensis*, *i.e.* quantiles ≥97.5 and ≤2.5, outlier genes were
279 found scattered across the genomes. Of these, 28 genes were outliers in both *D. sechellia* and
280 *D. y. mayottensis*. This proportion is not significantly different from what would be expected if the
281 5% most outlier genes were randomly drawn from the two species, most likely because a great

282 majority of *D. y. mayottensis* outliers were strongly linked and consequently non-independent.
283 Nonetheless, those overlapping genes may still be candidate for parallel genetic changes
284 underlying the convergent specialization on noni fruits (see Discussion below).

285 Leveraging the knowledge from functional genetic studies, we investigated the *PBE* values
286 for genes for which evidence for contribution in *D. sechellia* divergence was demonstrated. These
287 include the odorant receptor *Or22a*, which is responsible for olfactory attraction to noni (Auer et
288 al. 2020), the ionotropic receptor *Ir75b*, which underlies short-range attraction to noni (Prieto-
289 Godino et al. 2017), the odorant binding proteins *Obp57d* and *Obp57e*, which underlie preference
290 for noni taste (Matsuo et al. 2007) and *Catsup*, which contributes to the higher oviposition rate in
291 *D. sechellia* when reared on noni (Lavista-Llanos et al. 2014). *PBE* values were -0.198, -0.016, -
292 0.049, -0.058 and -0.036, respectively. *Obp57d/e* which falls amid the major *2R* chromosomal
293 inversions, whereas none of the remaining genes overlap with any of the *D. y. mayottensis* GIDs.
294 Nonetheless, the very low *PBE* value of *Or22a* was astonishing. This gene had one of the most
295 extreme absolute *PBE* value in *D. sechellia*. The negative sign is in complete agreement with the
296 scenario deduced from functional studies wherein mutations in the remaining four genes had
297 likely originated in the *D. sechellia* lineage whereas mutations in *Or22a* conferring noni attractions
298 in *D. sechellia* were likely ancestral (Auer et al. 2020).

299 In summary, whereas no major overlap could be detected between the outliers of the two
300 species, population genomics analyses in the *simulans* clade was still capable to identify some
301 early targets of selection driving *D. sechellia* specialization on noni. Those analyses do not support
302 a role of GIDs at the early stages of speciation in this homosequential clade.

303

304 **Discussion**

305 *Distinct genomic architectures between* D. sechellia *and* D. y. mayottensis

306 How much could we learn about the early stages of *Drosophila sechellia* specialization on
307 noni, a primary genetics model for insect-plant interactions, from population genomics analyses
308 of its distant relative, yet another noni-specialist, *D. yakuba mayottensis*? In theory, relative

309 divergence at parallel genetic changes between species should correlate in population pairs
310 inhabiting similar contrasting environments (e.g., Bohutínská et al. (2021); Montejo-Kovacevich
311 et al. (2021)). If specialization on noni has involved common genetic changes, one should expect
312 an enrichment of shared outlier genes between the two taxa, although those genes would be
313 members of large "genomic islands of divergence" in the recently diverging taxon,
314 *i.e. D. y. mayottensis*, and of less distinct "genomic atolls of divergence" in the older taxon,
315 *i.e. D. sechellia*. The rational is that at early stages of selection, variation in large genomic regions
316 tends to hitchhike with the selected gene(s). At later stages the level of average genome-wide
317 divergence accumulate to almost eroding through recombination the signal surrounding the
318 selected gene(s). Our results demonstrate that this straightforward expectation is not to be
319 encountered between *D. sechellia* and *D. y. mayottensis* mainly due to their distinct genomic
320 architectures.

321 Recombination landscape strongly differs between the two species. In *D. yakuba*, nearly
322 35.3% of the genome has low number of segregating sites, extending from the telomeres and
323 centromeres to region that are usually highly recombining in *D. melanogaster*. In the three species
324 of the *simulans* clade, only 10.5% of the genome has low number of segregating sites, a proportion
325 that is even lower than that encountered in *D. melanogaster* (Pool et al. 2012). *Drosophila yakuba*
326 is also highly polymorphic for large chromosomal inversions that are known to usually suppress
327 recombination. Of the 13 rearrangements identified by Lemeunier and Ashburner (1976) for
328 autosomal chromosomes, we encountered here five that were polymorphic in the two *D. y. yakuba*
329 populations. Besides we identified two inversions that were specific, and nearly fixed in
330 *D. y. mayottensis*. The three species of the *simulans* clade are, on the other hand, homosequential
331 and chromosomal inversions were very rarely reported in *D. simulans* (Chakraborty et al. 2021).
332 Such difference in recombination landscape complicates cross-species comparisons because
333 outlier genes will tend to predominantly accumulate in regions with low recombination in one
334 species, *i.e., D. yakuba*, and disperse across the genome in the other species, *i.e., D. sechellia*. In
335 some instances, such as in the intensively investigated "supergenes", inversions often involve

336 multiple genes that simultaneously control the diverging phenotype (Thompson and Jiggins 2014;
337 Durmaz et al. 2018; Villoutreix et al. 2021). In those instances, an enrichment of outlier genes may
338 be expected regardless to the recombination landscape difference between taxa. This however
339 does not seem to be the case between *D. sechellia* and *D. y. mayottensis*. There is no enrichment of
340 shared outliers in the two inversions on the X and 2L chromosomes that are specific to
341 *D. y. mayottensis*. Moreover, *D. y. mayottensis* has the standard karyotype of the highly
342 polymorphic chromosomal arm *2R*, which is thought to even predate the speciation between
343 *D. yakuba* and its closest relative *D. santomea* (Reilly 2020).
344
345 *Potential parallelism at the gene-level between the two noni-specialist species*
346 Despite the genomic distinctness between the two noni-specialists, some traces of
347 potential genetic parallelism are still detectable. For example, common outlier genes at GID1,
348 which is adjacent to inversion *Xtb*, directly flank *Flo2*, a long gene that responds to insecticide
349 exposure (Zhang et al. 2021). *Flo2* has a long intron, which contains multiple paralogs of the
350 glucose-methanol-choline oxidoreductases family, some of which have been associated with
351 insect-plant interactions (Iida et al. 2007; Rahfeld et al. 2014) or ethanol tolerance in
352 *D. melanogaster* (Glaser-Schmitt and Parsch 2018). Two shared outliers in the *2Lo* inversions are
353 genes fallen within the long intron of the *pdm3* gene. This long intron is known to harbor elements
354 that regulate *pdm3* expression (Yassin, Delaney, et al. 2016). This gene codes a transcription factor
355 that can determine the number of sensilla expressing the ionotropic receptor *Ir75b*, a gene that is
356 responsible for *D. sechellia* preference for noni's hexanoic acid odor (Prieto-Godino et al. 2017;
357 Arguello et al. 2021). The highest *PBE* value among the different outlier genes from the
358 polymorphic chromosomal arm *2R* concerns *GstE9*, a gene that is down-regulated in *D. sechellia*
359 antenna upon exposure to noni (Dworkin and Jones 2009; Shiao et al. 2015).
360 Under lesser stringent thresholds to detect potential parallelism, GIDs that were not
361 associated with inversions also shared some interesting outliers with *D. sechellia*. Most notably,
362 syntenic members of the *Osiris* gene family constitute GID13 in *D. y. mayottensis*. The role of those

363 genes in conferring resistance against noni main toxin, octanoic acid, has been demonstrated
364 through genetic mapping (Hungate et al. 2013) and transcriptomic analyses in *D. sechellia* (Lanno
365 et al. 2017), as well as functional gene silencing in *D. melanogaster* (Andrade López, J. M. et al.
366 2017). Taken together, our comparative population genomics highlighted a number of candidate
367 genes potentially involved in parallel genetic changes for future functional analyses.
368
369 *Population genomics signals of early targets of selection in* D. sechellia and *repeatability of genomic*
370 *divergence during convergent speciation*
371 The role of the genomic islands of divergence in driving speciation is tightly linked with
372 an oligogenic mode of speciation wherein strong selection on a few loci initiates the earliest stages
373 of speciation (Via 2009). As in hard selective sweep models, the amplitude and the width of the
374 islands will depend on the allelic difference between the diverging populations and the
375 recombination rate at their genomic positions. The alternative mode of speciation genomics is a
376 polygenic view wherein multifarious selection simultaneously acts on multiple loci. This mode can
377 also generate large islands under certain circumstances, but genomic divergence is mostly
378 expected to proceed through the formation of smaller islands that become rapidly submerged by
379 genome-wide average divergence (Michel et al. 2010; Nosil et al. 2021). Detection of such small
380 islands during the early stages is difficult, especially if selection proceeded on standing variation
381 with moderate frequencies and/or high number of loci, *i.e.,* a soft selective sweep model (Pritchard
382 et al. 2010). Paradoxically, if the selection pressure persists, freezing the ancestral variation in one
383 population while allowing the locus to cumulate variations in other populations, targets of this
384 selection could readily be detected at later stages of speciation. This is likely the case of the *Or22a*
385 olfactory receptor gene, which controls *D. sechellia* attraction to noni odor. Our population
386 genomics analysis confirmed earlier functional genetics hypothesis emitted by Auer et al. (2020)
387 that *D. sechellia* has retained an ancestral functionally relevant variation at this gene. In a
388 population genomics scan for signals of soft selective sweep specifically targeting chemosensory
389 genes in the 2013 population of *D. y. mayottensis*, *Or22a* was the most diverging olfactory receptor

390 (Ferreira et al. 2021). Hence, genic parallelism outside the genomic islands of divergence remains
391 potentially possible.

392 In summary, the genomic landscape of divergence in *D. yakuba* is predominated by large
393 islands overlapping inversion breakpoints. We cannot rule out that such islands may still harbor
394 noni-specialization genes, but they are so large that we cannot use population genomics data alone
395 to identify them. By contrast, a few mostly narrow islands that do not overlap with inversion
396 breakpoints harbor strong candidates for genetic parallelism between *D. y. mayottensis* and *D.*
397 *sechellia*, e.g., the *Osiris* gene family. For *D. sechellia*, which lacks large inversions, it is most likely
398 that a few genomic islands of divergence did not characterize the early stages of its speciation.
399 Indeed, QTL analyses of noni-use traits in *D. sechellia* suggested specialization to be highly
400 polymorphic; at least 15 loci on the five chromosomal arms underlie preference and performance
401 phenotypes (Jones 2005; Auer et al. 2021). QTL analyses will be complicated in *D. yakuba* by the
402 prevalence of chromosomal inversions. Future functional genetics analyses, which should be
403 simultaneously carried in both species, would certainly clarify potential genic parallelism. At
404 present, our results clearly show that distinct genomic architectures do not impede complex
405 convergent response to common selection between independent lineages.

406

407 **Materials and methods**

408 *Fly collection and establishment of isofemale lines from Mayotte island*

409 Flies were prospected in April 2017 from 14 localities on the Grande Terre and Petite
410 Terre islands of Mayotte. We searched fruiting noni trees. Whereas noni trees in multiple sites,
411 trees in only three localities had fruits at an advanced stage of ripening falling aground. *Drosophila*
412 *yakuba* flies were collected by net sweeping and direct aspiration over fallen fruits or bred from
413 ripe fruits brought back to the laboratory. Standard Drosophila baits with fermenting banana did
414 not yield any *D. yakuba*. Isofemale lines were established by placing one female and one male in a
415 small vial with instant Carolina medium and the F_1 progeny was transferred to an axenic
416 laboratory medium in the laboratory.

417

418 *Pooled and individual lines genome sequencing*

419 *Drosophila yakuba mayottensis* were sequenced both as pooled (Pool-Seq) as well as
420 individual lines (Ind-Seq). For Pool-Seq analyses, we pooled single F_1 female per line in three pools
421 according to the populations, namely BE, MTS and SOU. Pools were sent in ethanol to End2end
422 Genomics LLC, Davis, California where DNA was extracted, PE library prepared using NEBNext
423 Protocol and 75x covergage double-paired Illumina NovoSeq sequencing was conducted.

424 For Ind-Seq analyses, DNA was extracted from single flies flash frozen in liquid nitrogen
425 following QIAamp Mini Kit (Qiagen) protocol without using RNase A. The resulting DNA samples
426 were quantified (Qubit dsDNA HS assay kit, ThermoFisher Scientific), assessed for quality
427 (Nanodrop ND-2000; A260/A280$>$1.8), and stored at -20°C. Illumina TruSeq Nano DNA
428 libraries were prepared manually following the manufacturer's protocol (TruSeq Nano DNA,
429 RevD; Illumina). Briefly, samples were normalized to 100ng DNA and sheared by sonication with
430 Covaris M220 (microTUBE 50; Sage Science). The samples were end repaired, purified with
431 Ampure XP beads (Agencourt; Beckman Coulter), adaptors adenylated, and Unique Dual Indices
432 (Table) ligated. Adaptor enrichment was performed using eight cycles of PCR. Following Ampure
433 XP bead cleanup, fragment sizes for all libraries were measured using Agilent Bioanalyzer 2100
434 (HS DNA Assay; Applied Biosystems). The libraries were diluted 1:10 000 and 1:20 000 and
435 quantified in triplicate using the KAPA Library Quantification Kit (Kapa Biosystems). Equimolar
436 samples were pooled, and the libraries were size selected targeting 400-700bp range to remove
437 adaptor monomers and dimers using Pippen Prep DNA Size Selection system (1.5% Agarose Gel
438 Cassette #CDF1510; Sage Sciences). Library pools (24 samples per lane) were run on an llumina
439 HiSeq 4000 platform using the 150bp paired end (PE) Cluster Kit.

440

441

442

443 *Population genomics analyses and delimitation of the genomic islands of divergence (GIDs) in*
444 Drosophila yakuba

445 As in Yassin et al. (2016) and Ferreira et al. (2021), we used previously published genomic
446 sequences from two continental *D. y. yakuba* populations, namely Cameroon and Kenya (Rogers
447 et al. 2014), for comparisons with *D. y. mayottensis*. We also used previous Pool-Seq reads from a
448 single *D. y. mayottensis* population collected from Soulou in 2013 for temporal comparisons
449 within the subspecies (Yassin et al. 2016). For all comparisons, reads were mapped to the *D.*
450 *yakuba* v.1.05 and the *D. simulans* v.2.02 reference genome obtained from Flybase
451 (https://flybase.org/, Thurmond et al., 2019) using Minimap2 software package (Li 2018).
452 Minimap2-generated SAM files were converted to BAM format using samtools 1.9 software (Li et
453 al. 2009). The BAM files were then cleaned and sorted using Picard v.2.0.1
454 (http://broadinstitute.github.io/picard/). Popoolation 2 software package (v.1.20162, Kofler,
455 Pandey, and Schlötterer 2011) was used to generate a synchronized mpileup file, from which
456 intra-population polymorphism (π) and between-populations genetic differentiation (F_{ST}) were
457 estimated using a customized perl script given in (Ferreira et al. 2021).

458 We used the MavericK ver. 1.0 software package (Verity and Nichols 2016) to infer for
459 each 100-kb window of the v1.05 reference genome the likeliest number of demes (k) among
460 three possibilities 1, 2 and 3. For each window, we extracted bi-allelic sites for the 20 *D. y. yakuba*
461 and 19 *D. y. mayottensis* lines and ran MavericK using the admixture model k_{min} = 1, k_{max} = 3,500
462 burn-in iterations, and 250 sampled iterations for the main MCMC run, 250 burn-in iterations and
463 250 sampled iterations for the thermodynamic integrations MCMC, with 21 evenly spaced values
464 of β for thermodynamic integration (TI). These were the same parameters that were used by
465 Reilly (2020) to delimit chromosomal inversion polymorphisms within *D. y. yakuba*. Admixture
466 proportions for the optimal k for each line, chosen by maximum TI likelihood, were collated using
467 custom Perl script and plotted using R (Schie et al. 2014)(XXX). Structured regions were delimited
468 as consecutive windows with k > 1 fallen less then 500-kb apart in regions with high

469 recombination that were previously defined by Yassin et al. (2016)from polymorphism patterns
470 in the CY population.
471
472 *De novo assembly of* D. y. mayottensis *genome and identification of chromosomal rearrengements*
473 A mass culture of *D. y. mayottensis* was used to assemble a reference genome for this
474 subspecies. Genomic DNA was extracted using the Nucleobond AXG20 kit and buffer set IV from
475 Macherey-Nagel (ref 740544 and 740604, https://www.mn-net.com, Düren, Germany), and size
476 selection was conducted using the SRE XS kit from Circulomics (https://www.circulomics.com/,
477 Baltimore, Maryland, USA). Fragments less than 10 Kb were eliminated. The SQK-LSK109 kit from
478 Oxford nanopore technology (Lu, Giordano, and Ning 2016, https://nanoporetech.com/) was then
479 used to prepare the samples for nanopore sequencingfollowing manufacturer's protocol. The
480 library waq then loaded on a R9.4 flow cell for sequencing. On MinION. Raw datas were basecalled
481 with Guppy v4.0.11.
482 Illumina short-read sequences have also been produced to reconstruct our genome, using
483 a hybrid assembly strategy, combining those short accurate Illumina reads to our long Nanopore
484 reads. Illumina sequencing was realized by Novogene Company Limited
485 (https://en.novogene.com, Cambridge, UK) on our previously extracted DNA.
486 DNA sequences, from Nanopore and Illumina technologies, were then assembled using
487 MaSuRCA genome assembler software v3.4.0 (Zimin et al. 2013). We aligned our newly
488 constructed *D. y. mayottensis* genome using minimap2 software package to a newly assembled
489 genome of the reference *D. yakuba* strain (David et al. 2022) in order to correct for early
490 missassemblies in the v.1.05 reference genome (Miller et al. 2019). *D. y. mayottensis* scaffolds were
491 then ordered and oriented according to the reference genome, and chromosomal rearrangements
492 were visually detected from a dotplot constructed by R. The newly constructed *D. y. mayottensis*
493 genome was also aligned, following the same methodology, with a genome of *D. y. yakuba* (strain
494 NY73PB, NCBI Assembly GCA_016746335.2) from Kenya, which differs from the reference

genome in two rearrangements, namely *2Ll* and *2Rj*, to highlight inversions specific to the population of the island of Mayotte from those that could be polymorphic in *D. y. yakuba*.

Delimitation of the genomic islands of divergence (GIDs) in D. y. mayottensis

Pairwise F_{ST} measurements were used to estimate *Population Branch Statisitics (PBS)* and *Population Branch Excess (PBE)* using the formulas presented in Yi et al. (2010) and Yassin et al. (2016), respectively. Due to the lack of genetic structure within *D. y. mayottensis*, Pool-Seq reads of the three populations were pooled into a single Pool of Pools (PoP). *PBS* and *PBE* were made assuming an (CY,(NY,PoP) tree. Genes with *PBE* values falling ≥ the upper 97.5 or ≤ 2.5% quantiles were considered outliers, and the width of a genomic island of divergence was considered if two outlier genes did not fall more than 500-kb apart.

Population genomics analyses in the simulans clade

To test for possible similarities at the gene-level in divergence pattern between *D. y. mayottensis* and *D. sechellia*, we obtained short-read sequences of populations from the three species of the *simulans* clade. For *D. simulans*, we obtained data from Rogers et al. (2014) concerning 10 inbred lines from two geographical populations, namely Kenya (NS) and Madagascar (MS). For *D. mauritiana*, we obtained Pool-Seq data from 107 isofemale lines published by Nolte et al. (2013). For *D. sechellia*, we pooled individual fly genomes published by Schrider et al. (2018) from three islands, namely Mahé, La digue and Marianne. Schrider et al. (2018)found no evidence for genetic structure between the three islands, so sequences from those islands were pooled together for subsequent analyses. By contrast, these authors found genetic structure when sequences from a fourth island, Praslin, were analyzed. Therefore, we did not include any sequence from this island.

For the three species, short reads were mapped to the *Drosophila simulans* v.2.02 reference genome (Flybase) using minimap2, and subsequent population genomic analyses were conducted to estimate π and F_{ST} as detailed above for *D. yakuba*. For sake of comparison, a list of

522 orthologous genes was obtained from Flybase between *D. melanogaster*, *D. simulans* and *D. yakuba*
523 and a further run of BLAST (Altschul et al. 1990) was conducted to confirm orthology. F_{ST} was
524 calculated for each gene from pairwise species comparisons. A (*mauritiana*,(*simulans*,*sechellia*))
525 tree, which presumably reflects the phylogeny of the three species (Cruickshank and Hahn 2014),
526 was used to estimate *PBS* and *PBE* values. Similar quantile boundaries as in *D. yakuba* were used
527 to delimit outlier genes in the *simulans* clade.
528

529 **Acknowledgments**

530 The authors are grateful for Claire Golléty for help identifying noni plantation places in
531 Mayotte. Our late colleague, Jean R. David (1931-2021), who initially discovered *D. y.*
532 *mayottensis* in 2013, participated to the 2017 expedition. This work was funded by the
533 Agence Nationale de la Recherche grant no. ANR-18-CE2-0008, a Richard Lounsbery
534 Foundation grant and an ATM bursary from the Muséum National d'Histoire Naturelle to
535 AY, and a National Institutes of Health grant NIH NIGMS R35 GM133376 to RLR.
536

537 **Conflict of interest**

538 The authors declare no conflict of interest.

539 **References**

540 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J.*
541 *Mol. Biol.* 215:403–410.

542 Andrade López, J. M., Lanno, S. M., Auerbach, J. M., Moskowitz, E. C., Sligar, L. A., Wittkopp, P. J.,
543 Coolon JD. 2017. Genetic basis of octanoic acid resistance in Drosophila sechellia: functional
544 analysis of a fine-mapped region. *Mol. Ecol.* 26:1148–1160.

545 Arguello JR, Abuin L, Armida J, Mika K, Chai PC, Benton R, Ramaswami M. 2021. Targeted
546 molecular profiling of rare olfactory sensory neurons identifies fate, wiring, and functional
547 determinants. *Elife* 10:1–33.

548 Auer TO, Khallaf MA, Silbering AF, Zappia G, Ellis K, Álvarez-Ocaña R, Arguello JR, Hansson BS,
549 Jefferis GSXE, Caron SJC, et al. 2020. Olfactory receptor and circuit evolution promote host
550 specialization. *Nature* [Internet] 579:402–408. Available from:
551 http://dx.doi.org/10.1038/s41586-020-2073-7

552 Auer TO, Shahandeh MP, Benton R. 2021. Drosophila sechellia : A Genetic Model for Behavioral
553 Evolution and Neuroecology. *Annu. Rev. Genet.* [Internet] 55:527–554. Available from:
554 https://www.annualreviews.org/doi/10.1146/annurev-genet-071719-020719

555 Bamba M, Kawaguchi YW, Tsuchimatsu T. 2018. Plant adaptation and speciation studied by
556 population genomic approaches. *Dev. Growth Differ.*:12–24.

557 Berdan E, Enge S, Nyluns GM, Wellenreuther M, Martens GA, Pavia H. 2019. Genetic divergence
558 and phenotypic plasticity contribute to variation in cuticular hydrocarbons in the eaweed
559 fly Coelopa frigida. *Ecol. Evol.* 9:12156–12170.

560 Bohutínská M, Vlček J, Yair S, Laenen B, Konečná V, Fracassetti M, Slotte T, Kolár F. 2021.
561 Genomic basis of parallel adaptation varies with divergence in Arabidopsis and its relatives.
562 *Proc. Natl. Acad. Sci. U. S. A.* 118.

563 Chakraborty M, Chang C-H, Khost DE, Vedanayagam J, Adrion JR, Liao Y, Montooth KL,
564 Meiklejogn CD, Larracuente AM, Emerson JJ. 2021. Evolution of genome structure in the
565 Drosophila simulans species complex. *Genome Res.* 31:380–396.

566 Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due
567 to reduced diversity, not reduced gene flow. *Mol. Ecol.* 23:3133–3157.
568 David JR, Ferreira EA, Jabaud L, Ogereau D, Bastide H, Yassin A. 2022. Evolution of assortive
569 mating following selective introgression of pigmentation genes between two Drosophila
570 species. *Ecol. Evol.* 12:1–13.
571 David JR, Yassin A, Gidaszewski N, Debat V. 2014. Drosophilids (Diptera) from Mayotte island:
572 An annotated list of species collected in 2013 and comments on the colonisation of Indian
573 Ocean Islands. *Ann. la Soc. Entomol. Fr.* [Internet] 50:336–342. Available from:
574 http://dx.doi.org/10.1080/00379271.2014.938548
575 Durmaz E, Benson C, Kapun M, Schmidt P, Flatt T. 2018. An inversion supergene in Drosophila
576 underpins latitudinal clines in survival traits. *J. Evol. Biol.* [Internet] 31:1354–1364.
577 Available from: https://pubmed.ncbi.nlm.nih.gov/29904977/
578 Dworkin I, Jones CD. 2009. Genetic changes accompanying the evolution of host specialization in
579 Drosophila sechellia. *Genetics* 181:721–736.
580 Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, Künstner A, Mäkinen H,
581 Nadachowska-Brzyska K, Qvarnström A, et al. 2012. The genomic landscape of species
582 divergence in Ficedula flycatchers. *Nature* 491:756–760.
583 Faria R, Johannesson K, Butlin RK, Westram AM. 2019. Evolving Inversions. *Ecol. Evol.* 34:239–
584 248.
585 Feder JL, Nosil P. 2009. Chromosomal inversions and species differences: When are genes
586 affecting adaptive divergence and reproductive isolation expected to reside within
587 inversions? *Evolution (N. Y).* 63:3061–3075.
588 Ferreira EA, Lambert S, Verrier T, Marion-Poll F, Yassin A. 2021. Soft selective sweep on
589 chemosensory genes correlates with ancestral preference for toxic noni in a specialist
590 drosophila population. *Genes (Basel).* 12:1–13.
591 Forbes AA, Devine SN, Hippee AC, Tvedte ES, Ward AKG, Widmayer HA, Wilson CJ. 2017.
592 Revisiting the particular role of host shifts in initiating insect speciation. *Evolution (N. Y).*

593 71:1126–1137.
594 Glaser-Schmitt A, Parsch J. 2018. Functional characterization of adaptative variation within a cis-
595 regulatory element influencing Drosophila melanogaster growth. *PLoS Biol.* 16:1–28.
596 Hébert JB, Scheffer SJ, Hawthorne DJ. 2016. Evidence for ecological speciation via a host shift in
597 the holly leaf miner, Phytomyza glabricola (Diptera: Agromyzidae). *Ecol. Evol.* 6:6565–
598 6577.
599 Hivert V, Leblois R, Petit EJ, Gautier M, Vitalis R. 2018. Measuring Genetic Differentiation from
600 Pool-seq Data. *Genetics* 210:315–330.
601 Huang K, Andrew RL, Owens GL, Ostevik KL, Rieseberg LH. 2020. Multiple chromosomal
602 inversions contribute to adaptive divergence of a dune sunflower ecotype. *Mol. Ecol.*
603 29:2535–2549.
604 Hungate EA, Earley EJ, Boussy IA, Turissini DA, Ting CT, Moran JR, Wu ML, Wu CI, Jones CD.
605 2013. A locus in Drosophila sechellia affecting tolerance of a host plant toxin. *Genetics*
606 195:1063–1075.
607 Iida K, Cox-Foster DL, Yang X, Ko WY, Cavener DR. 2007. Expansion and evolution of insect GMC
608 oxidoreductases. *BMC Evol. Biol.* 7:1–12.
609 Jay P, Chouteau M, Whibley A, Bastide H, Parrinello H, Llaurens V, Joron M. 2021. Mutation load
610 at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. *Nat.*
611 *Genet.* 53:288–293.
612 Jones CD. 2005. The genetics of adaptation in Drosophila sechellia. *Genetica* 123:137–145.
613 Kofler R, Pandey RV, Schlötterer C. 2011. PoPoolation2: Identifying differentiation between
614 populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* 27:3435–
615 3436.
616 Lachaise D, Harry M, Solignac M, Lemeunier F, Benassi V, Cariou ML. 2000. Evolutionary
617 novelties in islands: Drosophila santomea, a new melanogaster sister species from Sao
618 Tome. *Proc. R. Soc. B Biol. Sci.* 267:1487–1495.
619 Lanno SM, Gregory SM, Shimshak SJ, Alverson MK, Chiu K, Feil AL, Findley MG, Forman TE,

620 Gordon JT, Ho J, et al. 2017. Transcriptomic analysis of octanoic acid response in Drosophila
621 sechellia using RNA-sequencing. *G3 Genes, Genomes, Genet.* 7:3867–3873.
622 Lavista-Llanos S, Svatoš A, Kai M, Riemensperger T, Birman S, Stensmyr MC, Hansson BS. 2014.
623 Dopamine drives Drosophila sechellia adaptation to its toxic host. *Elife* 3:3785–3802.
624 Legal L, Chappe B, Jallon JM. 1994. Molecular basis of Morinda citrifolia (L.): Toxicity on
625 drosophila. *J. Chem. Ecol.* 20:1931–1943.
626 Legrand D, Tenaillon MI, Matyot P, Gerlach J, Lachaise D, Cariou M. 2009. Species-Wide Genetic
627 Variation and Demographic History of Drosophila sechellia , a Species Lacking Population
628 Structure. *Genetics* 182:1197–1206.
629 Legrand D, Vautrin D, Lachaise D, Cariou ML. 2011. Microsatellite variation suggests a recent
630 fine-scale population structure of Drosophila sechellia, a species endemic of the Seychelles
631 archipelago. *Genetica* 139:909–919.
632 Lemeunier F, Ashburner M. 1976. Relationships within the melanogaster species subgroup of the
633 genus Drosophila (Sophophora). II. Phylogenetic relationships between six species based
634 upon polytene chromosome banding sequences. *Proc. R. Soc. London - Biol. Sci.* [Internet]
635 193:275–294. Available from:
636 https://royalsocietypublishing.org/doi/abs/10.1098/rspb.1976.0046
637 Li H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–
638 3100.
639 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009.
640 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.
641 Lu H, Giordano F, Ning Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly.
642 *Genomics, Proteomics Bioinforma.* 14:265–279.
643 Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel and
644 Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring
645 of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* 38:4647–4654.
646 Matsuo T, Sugaya S, Yasukawa J, Aigaki T, Fuyama Y. 2007. Odorant-binding proteins OBP57d

647 and OBP57e affect taste perception and host-plant preference in Drosophila sechellia. *PLoS*
648 *Biol.* 5:0985–0996.
649 Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P, Feder JL. 2010. Widespread genomic
650 divergence during sympatric speciation. *Proc. Natl. Acad. Sci. U. S. A.* 107:9724–9729.
651 Miller DE, Staber C, Zeitlinger J, Hawley RS. 2018. Highly Contigous Genome Assemblies of 15
652 Drosophila Species Generated Using Nanopore Sequencing. *G3 Genes, Genomes, Genet.*
653 8:3131–3141.
654 Montejo-Kovacevich G, Meier JI, Bacquet CN, Warren IA, Chan YF, Kucka M, Salazar C, Rueda N,
655 Montgomery SH, McMillan WO, et al. 2021. Repeated genetic adaptation to high altitude in
656 two tropical butterflies. *bioRxiv* [Internet]:2021.11.30.470630. Available from:
657 https://www.biorxiv.org/content/10.1101/2021.11.30.470630v1%0Ahttps://www.biorxi
658 v.org/content/10.1101/2021.11.30.470630v1.abstract
659 Nolte V, Vinay Pandey R, Kofler R, Christian S. 2013. Genome-wide patterns of natural variation
660 reveal strong selective sweeps and ongoing genomic conflict in Drosophila mauritiana.
661 *Genome Res.* 23:99–110.
662 Nosil P, Feder JL, Gompert Z. 2021. How many genetic changes create new species? *Science (80-.*
663 *).* 371:777–779.
664 Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchen P, Emerson
665 JJ, Saelao P, Begun DJ, et al. 2012. Population Genomics of Sub-Saharan Drosophila
666 melanogaster: African Diversity and Non-African Admixture. *PLoS Genet.* 8.
667 Prieto-Godino LL, Rytz R, Cruchet S, Bargeton B, Abuin L, Silbering AF, Ruta V, Dal Peraro M,
668 Benton R. 2017. Evolution of Acid-Sensing Olfactory Circuits in Drosophilids. *Neuron*
669 93:661-676.e6.
670 Prigent SR, Le Gall P, Mbunda SW, Veuille M. 2013. Seasonal and altitudinal structure of
671 drosophilid communities on Mt Oku (Cameroon volcanic line). *Comptes Rendus - Geosci.*
672 345:316–326.
673 Pritchard JK, Pickrell JK, Coop G. 2010. The Genetics of Human Adaptation: Hard Sweeps, Soft

674 Sweeps and Polygenic Adaptation. *Curr. Biol.* 20:208–215.
675 Quilodrán CS, Ruegg K, Sendell-price AT, Anderson E, Coulson T, Clegg S. 2019. The multiple
676 population genetic and demographic routes to islands of genomic divergence.
677 R'Kha S, Capy P, David JR. 1990. Host-plant specialization in the Drosophila melanogaster
678 species complex: a physiological, behavioral, and genetical analysis. *Proc. Natl. Acad. Sci.*
679 88:1835–1839.
680 Rahfeld P, Kirsch R, Kugel S, Wielsch N, Stock M, Groth M, Boland W, Burse A. 2014.
681 Independently recruited oxidases from the glucose-methanol-choline oxidoreductase
682 family enabled chemical defences in leaf beetle larvae (subtribe Chrysomelina) to evolve.
683 *Proc. R. Soc. B Biol. Sci.* 281.
684 Reilly PF. 2020. Comparative Genomics of the Drosophila yakuba group.
685 Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P, Thornton KR. 2014. Landscape of standing
686 variation for tandem duplications in drosophila yakuba and drosophila simulans. *Mol. Biol.*
687 *Evol.* 31:1750–1766.
688 Schie S, Chaudhary R, Debener T. 2014. Analysis of a Complex Polyploid Plant Genome using
689 Molecular Markers : Strong Evidence for Segmental Allooctoploidy in Garden Dahlias. *Plant*
690 *Genome* 7:1–6.
691 Schrider DR, Ayroles J, Matute DR, Kern AD. 2018. Supervised machine learning reveals
692 introgressed loci in the genomes of Drosophila simulans and D . sechellia. *Plos Genet.* 14:1–
693 29.
694 Shiao MS, Chang JM, Fan W-L, Lu MYJ, Notredame C, Fang S, Kondo R, Li W-H. 2015. Expression
695 Divergence of Chemosensory Genes between Drosophila sechellia and Its Sibling Species
696 and Its Implications for Host Shift. *Genome Biol. Evol.* 7:2843–2858.
697 Sodeland M, Jorde PE, Lien S, Jentoft S, Berg PR, Grove H, Kent MP, Arnyasi M, Olsen EM, Knutsen
698 H. 2016. "Islands of Divergence" in the atlantic cod genome represent polymorphic
699 chromosomal rearrangements. *Genome Biol. Evol.* 8:1012–1022.
700 Suvorov A, Kim BY, Wang J, Armstrong EE, Peede D, D'Agostino ERR, Price DK, Waddell P, Lang

701 M, Courtier-Orgogozo V, et al. 2022. Widespread introgression across a phylogeny of 155
702 Drosophila genomes. *Curr. Biol.* 32:111-123.e5.
703 Thompson MJ, Jiggins CD. 2014. Supergenes and their role in evolution. *Heredity (Edinb).*
704 [Internet] 113:1–8. Available from: https://pubmed.ncbi.nlm.nih.gov/24642887/
705 Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, Matthews BB, Millburn
706 G, Antonazzo G, Trovisco V, et al. 2019. FlyBase 2.0: The next generation. *Nucleic Acids Res.*
707 47:D759–D765.
708 Turner TL, Hahn MW, Nuzhdin S V. 2005. Genomic islands of speciation in Anopheles gambiae.
709 *PLoS Biol.* 3:1572–1578.
710 Verity R, Nichols RA. 2016. Estimating the number of subpopulations (K) in structured
711 populations.
712 Via S. 2009. Natural selection in action during speciation. *Proc. Natl. Acad. Sci.* 106:9939–9946.
713 Villoutreix R, Ayala D, Joron M, Gompert Z, Feder JL, Nosil P. 2021. Inversion breakpoints and the
714 evolution of supergenes. *Mol. Ecol.* [Internet] 30:2738–2755. Available from:
715 https://pubmed.ncbi.nlm.nih.gov/33786937/
716 Wolf JBW, Ellegren H. 2017. Making sense of genomic islands of differentiation in light of
717 speciation. *Nat. Rev. Genet.* [Internet] 18:87–100. Available from:
718 http://dx.doi.org/10.1038/nrg.2016.133
719 Yassin A. 2017. Drosophila yakuba mayottensis, a new model for the study of incipient ecological
720 speciation. *Fly (Austin).* 11:37–45.
721 Yassin A, Debat V, Bastide H, Gidaszewski N, David JR, Pool JE. 2016. Recurrent specialization on
722 a toxic fruit in an island Drosophila population. *Proc. Natl. Acad. Sci. U. S. A.* 113:4771–4776.
723 Yassin A, Delaney EK, Reddiex AJ, Seher TD, Bastide H, Appleton NC, Lack JB, David JR,
724 Chenoweth SF, Pool JE, et al. 2016. The pdm3 Locus Is a Hotspot for Recurrent Evolution of
725 Female-Limited Color Dimorphism in Drosophila. *Curr. Biol.* 26:2412–2422.
726 Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N,
727 Korneliussen TS, et al. 2010. Sequencing of Fifty Human Exomes Reveals Adaptation to

High Altitude. *Science (80-.).* 329:75–78.

Zhang C, Seong KM, Sun W, Mittapalli O, Qiu B, Clark JM, Pittendrigh BR. 2021. The insulin signaling pathway in Drosophila melanogaster: A nexus revealing an “Achilles’’ heel” in DDT resistance.” *Pestic. Biochem. Physiol.* [Internet] 171:104727. Available from: https://doi.org/10.1016/j.pestbp.2020.104727

Zimin A V., Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* 29:2669–2677.

737 **Figure legends**

738

739 Figure 1. A. Geographical distribution of Drosophila yakuba populations used in this study.
740 Orange and green dots refer to D. y. yakuba and D. y. mayottensis populations, respectively. B. A
741 Lemur eating a fruit of Morinda citrifolia on Mayotte Island, photo credit: Amir Yassin. D. y.
742 mayottensis flies were usually collected over fallen nibbled fruits. C. Pattern of gnetic structure
743 in 100-kb windows across the genome between the five populations (separated by red
744 horizontal lines). Light blue, navy blue and yellow refer to window attribution to the first,
745 second or third genetic cluster by Maverick. The lower rectangles in dark grey correspond to
746 regions of low recombination and the light grey ones to regions of high recombination based on
747 polymorphism patterns in CY (Yassin et al. 2016). The highlighted GIDs were identified only for
748 high recombination regions.

749

750 Figure 2. Alignment of Drosophila yakuba mayottensis genome on (A) Drosophila yakuba on a
751 new assemby of the reference genome correcting previous early missassemblies (David et al.
752 2022) and (B) D. y. yakuba from Kenya NY73PB (Assembly GCA_016746335.2). Each colour
753 represents a scaffold obtained during D. y. mayottensis reconstruction. The bottom rectangles
754 on graph (A) represents the low recombination regions in dark grey and the high recombination
755 regions in light grey.

756

757 Figure 3. Population Branch Excess (PBE) comparisons at orthologous genes ± 2kb between the
758 two noni specialists (A) Drosophila yakuba mayottensis and (B) Drosophila sechellia. Genomic
759 islands of divergence (GIDs) for D. y. mayottensis are indicated in yellow and the limits of
760 chromosomal inversions are indicated as colored horizontal bars. Below each graph the extent of
761 low recombining telomeric and centromeric regions as well as of high recombining regions are
762 given in black and grey, respectively, for each taxon. Genes that have been shown by functional
763 studies to be involved in noni specialization in D. sechellia, are given in blue.

Figure 1

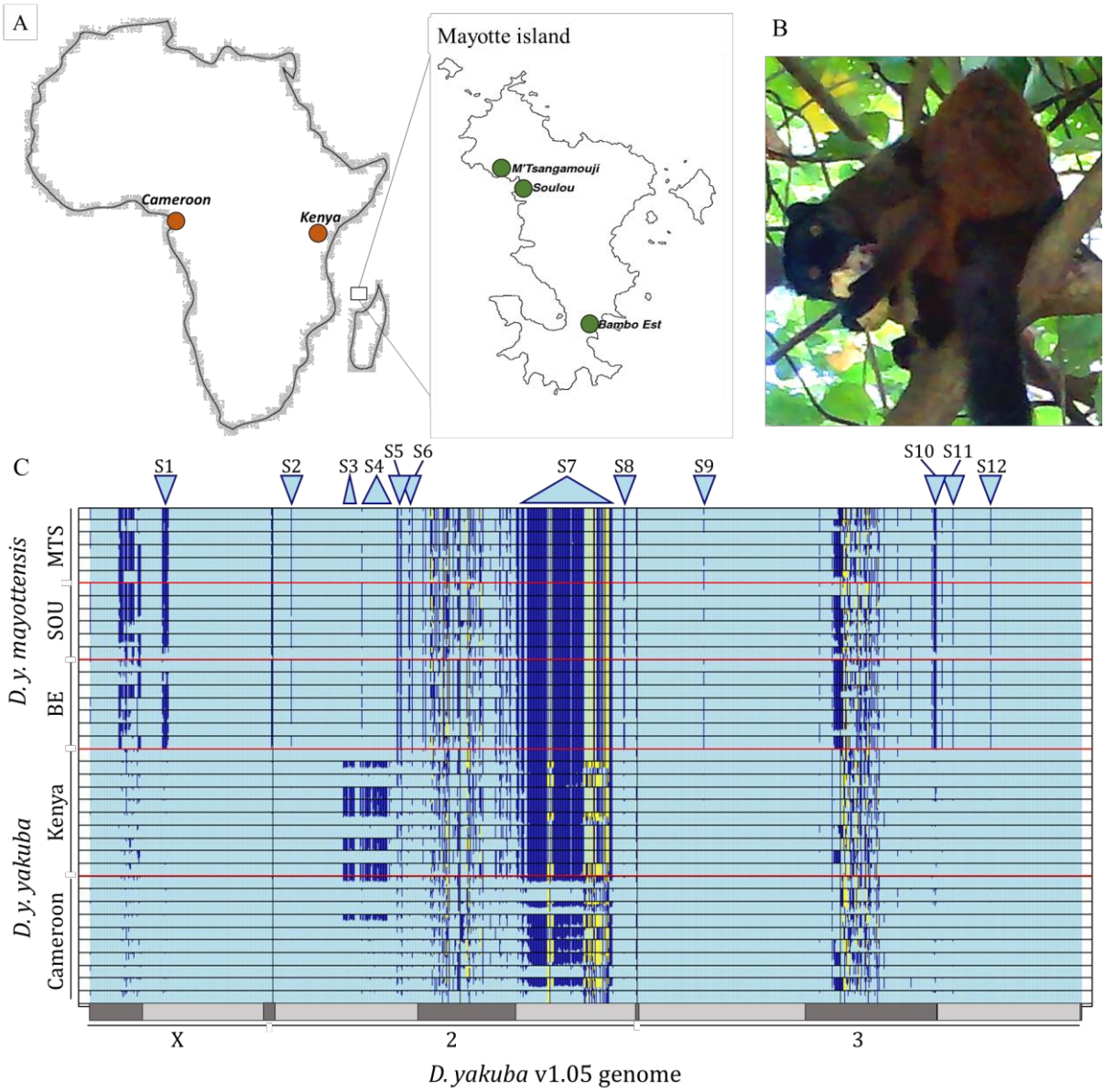

A
Cameroon
Kenya
Mayotte island
M'Tsangamouji
Soulou
Bambo Est
B
C
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
D. y. mayottensis
MTS
SOU
BE
D. y. yakuba
Kenya
Cameroon
X
2
3
D. yakuba v1.05 genome

Figure 2

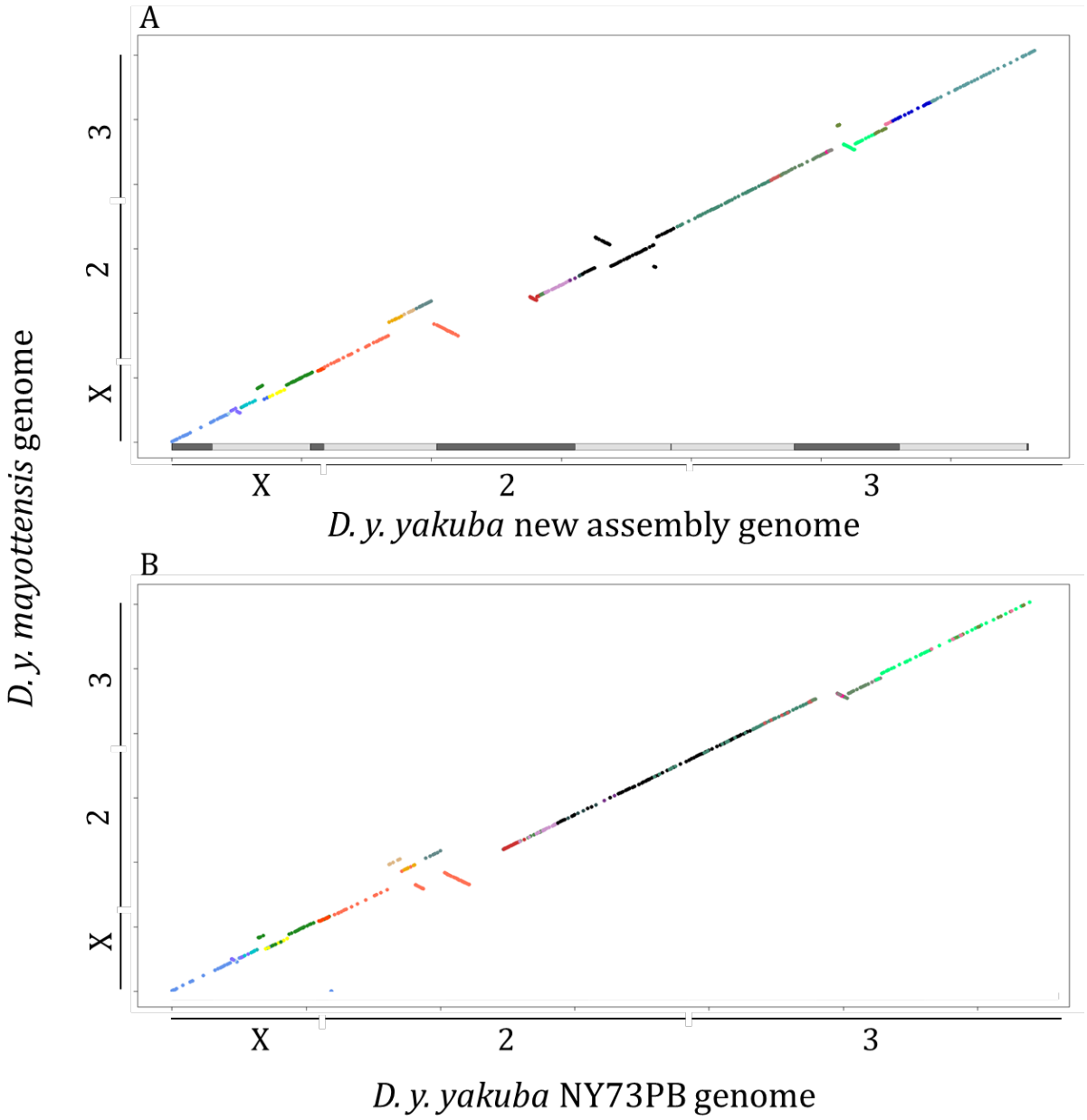
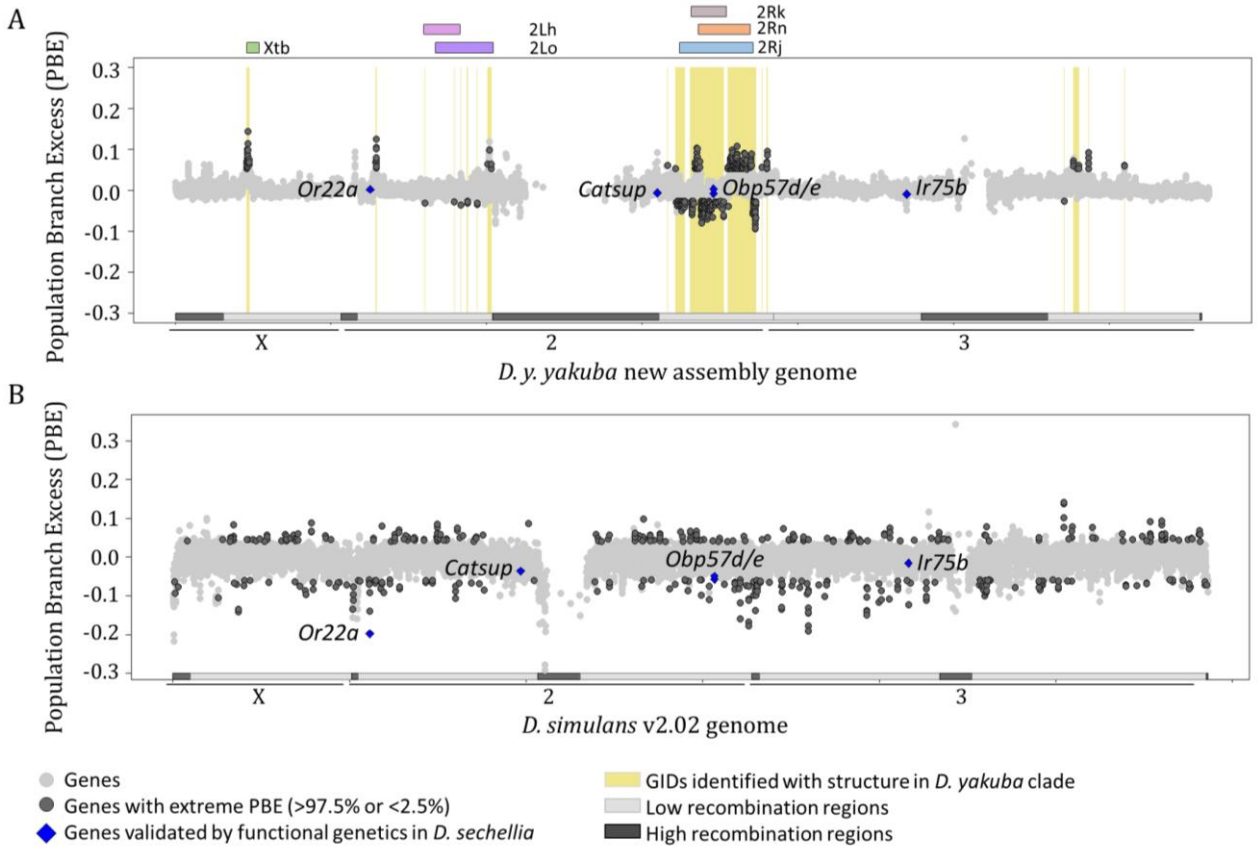

A
3
2
X
X
2
3
D. y. yakuba new assembly genome
D. y. mayottensis genome
B
3
2
X
X
2
3
D. y. yakuba NY73PB genome

Figure 3


A
Xtb
2Lh
2Lo
2Rk
2Rn
2Rj
Population Branch Excess (PBE)
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
Or22a
Catsup
Obp57d/e
Ir75b
X
2
3
D. y. yakuba new assembly genome
B
Population Branch Excess (PBE)
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
Catsup
Obp57d/e
Ir75b
Or22a
X
2
3
D. simulans v2.02 genome
Genes
Genes with extreme PBE (>97.5% or <2.5%)
Genes validated by functional genetics in D. sechellia
GIDs identified with structure in D. yakuba clade
Low recombination regions
High recombination regions