



Genomic islands of divergence between *Drosophila yakuba* subspecies are predominantly driven by chromosomal inversions and the recombination landscape

Erina Ferreira, Cathy Moore, David Ogereau, Arnaud Suwalski, Stéphane Prigent, Rebekah Rogers, Amir Yassin

► To cite this version:

Erina Ferreira, Cathy Moore, David Ogereau, Arnaud Suwalski, Stéphane Prigent, et al.. Genomic islands of divergence between *Drosophila yakuba* subspecies are predominantly driven by chromosomal inversions and the recombination landscape. *Molecular Ecology*, 2025, 34 (3), <10.1111/mec.17627>. <hal-04910597>

HAL Id: hal-04910597

<https://hal.science/hal-04910597v1>

Submitted on 24 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

Original Article

Genomic islands of divergence between *Drosophila yakuba* subspecies are predominantly driven by chromosomal inversions and the recombination landscape

Erina A. Ferreira^{1,2}, Cathy C. Moore³, David Ogereau¹, Arnaud Suwalski², Stéphane R. Prigent², Rebekah L. Rogers³, Amir Yassin^{1,2,*}

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

²Institut Systématique, Evolution, Biodiversité (ISYEB), CNRS, MNHN, Sorbonne Université, EPHE, Université des Antilles, 57 rue Cuvier, CP 50, 75005 Paris, France. ³Department of Bioinformatics and Genomics, University of North Carolina, Charlotte NC, USA.

* Corresponding author: amir.yassin@universite-paris-saclay.fr

Short title: Inversions and host shift in *Drosophila*

Abstract

During the early stages of local adaptation and 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*, and reconfirm using field collection its association with toxic noni (*Morinda citrifolia*) fruits on the Mayotte island. Population genomics revealed lack of genetic structure on the island and identified 23 GIDs distinguishing *D. y. mayottensis* from generalist mainland populations of *D. y. yakuba*. The GIDs were enriched with gene families involved in the metabolism of lipids, sugars, peptides and xenobiotics, suggesting a role in host shift. We assembled a new genome for *D. y. mayottensis* and identified five novel chromosomal inversions. Twenty one GIDs (~99% of outlier windows) fell in low recombining regions or subspecies-specific inversions. However, only two GIDs were in collinear, normally recombining regions suggesting a signal of hard selective sweeps. Unlike *D. y. mayottensis*, *D. sechellia*, the only other noni-specialist, is known to be homosequential with its generalist relatives. Thus, whereas structural variation may disproportionately shape GIDs in some species, striking parallel adaptations can occur between species despite distinct genomic architectures.

Keywords: ecological speciation; structural variation; population genomics; genetic parallelism; insect-plant interactions.

36 **Introduction**

37 The genomic landscape of inter-population differentiation is often heterogeneous, with
38 certain genomic regions, known as genomic islands of divergence (GIDs), emerging above the
39 average differentiation level (Feder and Nosil 2010; Wolf and Ellegren 2017). GIDs were
40 originally hailed as a result of divergent selection acting on genes that underlie adaptation to
41 distinct environment and/or reproductive isolation and counteracting the effects of gene flow
42 (Turner et al. 2005). Indeed, strong selection reduces neutral diversity around the adaptive
43 site(s), leading to the formation of fixed or nearly-fixed long tracts of homozygosity within the
44 population, a phenomenon known as ‘selective sweep’ (Garud et al. 2021). When compared to
45 other populations not subject to the same selection pressure, these localized low diversity
46 regions produce higher signals of differentiation, *i.e.* GIDs. Gene flow on the other hand reduces
47 differentiation. Despite initial enthusiasm, a plethora of adaptive and non-adaptive mechanisms
48 were later shown to also form localized long tracts of homozygosity within populations, and
49 consequently GIDs between populations, even in the absence of divergent selection or gene
50 flow (Quilodrán et al. 2020). Of these, variation in recombination rate, due to heterochromatin
51 content or structural variation, strongly affects the rate of admixture and the accumulation of
52 genetic differences in natural populations and statistically bias differentiation estimates (Feder
53 and Nosil 2010; Cruickshank and Hahn 2014; Booker et al. 2020; Quilodrán et al. 2020).
54 Therefore, proper understanding of the genomic architecture of the study model is essential to
55 untangle the mechanisms underlying the evolution of GIDs.

56 Heterochromatic effects on recombination are usually higher near the centromere and,
57 to a lesser extent, the telomeres (Brand et al. 2018; Hughes and Hawley 2020). They often lead
58 to a general pattern of low diversity at the extremes of chromosomal arms and intermediate to
59 high diversity in the middle of the arms (Vijay et al. 2017; Garud et al. 2021). The effect on
60 diversity is due to interaction with background selection purging deleterious mutations and their

physically linked neutral variation in low recombining regions and/or the mutagenic effect of recombination, which will be more pronounced in normally recombining regions. Consequently, drift can fix distinct neutral mutations in low-recombining regions more readily than in high-recombining regions due to the reduced effective population size, leading to rapid complete lineage sorting in the formers and the formation of GIDs (Pease and Hahn 2013). Similarly, positive selection acting on low recombining regions will have an exaggerated linkage effect. Indeed, large GIDs were found in centromeric regions in multiple species (White et al. 2010; Ellegren et al. 2012a). The extent of such centromeric and telomeric effects along the chromosomal arm differs between species (Brand et al. 2018; Chakraborty et al. 2021a; Pettie et al. 2022).

Chromosomal inversions are among the most common forms of structural variation, and a number of studies has unraveled an overlap between inversions and GIDs (Feder and Nosil 2009; Michel et al. 2010; Ellegren et al. 2012b; Sodeland et al. 2016; Huang et al. 2020). Inversions often suppress recombination in heterozygotes, reducing the rate at which slightly-deleterious mutations are purged from the population (Berdan et al. 2021; Jay et al. 2021, 2022). Tracts of neutral and slightly-deleterious mutations can be fixed if different karyotypes were randomly sampled and reached fixation in diverging populations, leading to the formation of GIDs. Conversely, the reduction of recombination could conserve one or more favorable allelic combinations for survival in contrasting environments, hence becoming targets of balancing or diverging selection depending on the heterogeneity of the environment (Faria et al. 2019; Kapun and Flatt 2019; Mérot et al. 2021; Hager et al. 2022; Haringmeyer and Hoekstra 2022; Sanchez-Donoso et al. 2022; Tepolt et al. 2022). The fate of the inversion and consequently the shape of the GID will mostly depend on the age and the distribution of beneficial alleles (Schaal et al. 2022). Young inversions usually do not form GIDs, but as time goes, and each karyotype accumulates mutations, divergence increases and the GID appears as a plateau or multi-peaked

(Guerrero et al. 2012). However, because recombination may still persist in the middle of the inversion, a phenomenon known as gene flux, divergence peaks mainly concentrate at the breakpoints in old inversions (Navarro et al. 2000; Guerrero et al. 2012; Schaal et al. 2022). Remarkably, inversions in *Drosophila* also suppress and reduce recombination in their 3 and up to 12 Mb long regions outside the breakpoints (Koury 2023), which should also lead to the formation of GIDs outside but close to the breakpoints. 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.

Host shifts or changes in the width of the ecological niche, e.g., specialization, are a major source of diversification in herbivorous insects (Wiens et al. 2015; 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 (Futuyma and Moreno 1988; Vertacnik and Linnen 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 citrifolia*) on the Seychelles islands in the Indian Ocean (Louis and David 1986; Auer et al. 2021). *Drosophila sechellia* is a member of the *simulans* species complex, which comprises two additional species, namely the cosmopolitan *Drosophila simulans* and *Drosophila mauritiana*, an endemic to Mauritius. Unlike *D. sechellia*, both species are generalists and are repulsed by the odor of and intolerant to noni main toxins, octanoic and hexanoic acids (R’Kha et al. 1991; Amlou et al. 1997, 1998; Drum et al. 2022). The three species represent a sister clade to *Drosophila melanogaster*. This facilitated the extension of the unique genetic and genomic tools from this paradigmatic species to the study of *D. sechellia* specialization (Matsuo et al. 2007; Lavista-Llanos et al. 2014; Prieto-Godino et al. 2017; Auer et al. 2020). Besides, the three species show partial reproductive isolation and are karyologically homosequential, making

them ideal for quantitative trait loci (QTLs) mapping through interspecific crosses (Earley and Jones 2011; Hungate et al. 2013; Huang and Erezyilmaz 2015; Chakraborty et al. 2021b). Despite these advantages that can help identifying the genetic basis of potential host-driven divergence, the divergence of *D. sechellia* from its generalist sister species *D. simulans* is estimated at around 89,000-242,000 years ago (Garrigan et al. 2012; Schrider et al. 2018); its entire genome has accumulated so many differences that traces of possible GIDs during the early stages of divergence may be difficult to observe. Besides, a particularly low level of genetic diversity has been detected in this species (Legrand et al. 2009; Schrider et al. 2018), further complicating population genomics analyses.

In January 2013, a second case of *Drosophila* species associated with toxic noni was discovered in Mayotte, another island in the Indian Ocean. This was a population of *Drosophila yakuba*, which had higher preference for and survival on noni unlike its continental conspecifics (Yassin et al. 2016). The noni-specialist population likely diverged from generalist continental populations nearly 28,000 years ago and specialization was accompanied by precopulatory isolation between Mayotte and continental flies, earning them subspecific status as *D. y. mayottensis* and *D. y. yakuba*, respectively (Yassin et al. 2016; Yassin 2017). Genomic comparisons between this single population of *D. y. mayottensis* and two *D. y. yakuba* continental populations identified multiple GIDs, widespread across the genome, consistent with theoretical expectations (Yassin et al. 2016). Demographic inference also revealed a substantial amount of persistent gene flow between the specialist and generalist subspecies indicating that the GIDs may correspond to regions resistant to admixture (Yassin et al. 2016). Indeed, many *D. y. mayottensis* GIDs overlap with large noni survival QTLs in *D. sechellia* (Yassin et al. 2016). Despite those similarities, *D. yakuba* is a species where, unlike the *simulans* species complex, multiple polymorphic chromosomal inversions have been identified

in continental populations (Lemeunier and Ashburner 1976). Several GIDs in *D. y. mayottensis* may therefore be the result of such inversions.

The recent divergence of *D. y. mayottensis* and the presence of multiple inversions in *D. yakuba* make this system ideal to investigate the impact of structural variation on GIDs formation. Large GIDs in collinear, normally-recombining regions present strong evidence for selective sweeps (Garud et al. 2021), and we aimed to quantify the extent of such GIDs in *D. y. mayottensis*. Hence, we report here the population genomics analysis of three additional populations of *D. y. mayottensis* all associated with noni in Mayotte. We identified 23 GIDs distinguishing the three populations from continental *D. y. yakuba*, inferred the width of the telomeric and centromeric effects on diversity and assembled a genome for *D. y. mayottensis*. Multiple inversions predominantly overlapped with the GIDs, hence allowing the discrimination between GIDs presenting genuine signal of selective sweep from those which may be mere results of structural variation or low recombination.

Materials and methods

Fly collection and establishment of isofemale lines from Mayotte island

Flies were collected in April 2017 from 14 localities on the Grande Terre and Petite Terre islands of Mayotte. We particularly searched fruiting noni trees but also collected drosophilids in places with no noni trees. Flies were collected by net sweeping and direct aspiration over fallen fruits or using standard baits with fermented banana hung from tree branches. Flies were also bred from ripe fruits brought back to the laboratory. *D. y. mayottensis* isofemale lines were established by placing one female and one male in a small vial with instant Carolina medium and the F₁ progeny was transferred to an axenic medium in the laboratory. Fly species that were or could not be bred in the laboratory were killed in absolute ethanol and

taxonomically sorted in the laboratory. Data on *D. yakuba* seasonal abundance and attraction to banana baits in West Africa was obtained from Voudibio (1985).

Pooled and individual lines genome sequencing

Drosophila yakuba mayottensis were sequenced both as pooled (Pool-Seq) as well as individual lines (Ind-Seq). For Pool-Seq analyses, we pooled single F₁ females per line in three pools according to the populations, namely Bambo Est (BE), M'tsangamouji (MTS) and Soulou (SOU), which is the same locality where the subspecies was first collected in 2013. Pools consisted of 26, 22 and 29 females, respectively, which were sent in ethanol to End2end Genomics LLC, Davis, California for DNA extraction, PE library preparation using NEBNext Protocol and double-paired Illumina NovoSeq sequencing at 75x coverage. For Ind-Seq analyses, DNA was extracted from single flies flash frozen in liquid nitrogen following QIAamp Mini Kit (Qiagen) protocol without using RNase A. These included 7, 6 and 6 lines from the three BE, MTS and SOU populations, respectively. The resulting DNA samples were quantified (Qubit dsDNA HS assay kit, ThermoFisher Scientific), assessed for quality (Nanodrop ND-2000), and stored at -20°C. Illumina TruSeq Nano DNA libraries were prepared manually following the manufacturer's protocol (TruSeq Nano DNA, RevD; Illumina). Briefly, samples were normalized to 100ng DNA and sheared by sonication with Covaris M220 (microTUBE 50; Sage Science). The samples were end repaired, purified with Ampure XP beads (Agencourt; Beckman Coulter), adaptors adenylated, and Unique Dual Indices ligated. Adaptor enrichment was performed using eight cycles of PCR. Following Ampure XP bead cleanup, fragment sizes for all libraries were measured using Agilent Bioanalyzer 2100 (HS DNA Assay; Applied Biosystems). The libraries were diluted 1:10,000 and 1:20,000 and quantified in triplicate using the KAPA Library Quantification Kit (Kapa Biosystems). Equimolar samples were pooled, and the libraries were size selected targeting 400-700bp range

to remove adaptor monomers and dimers using Pippin Prep DNA Size Selection system (1.5% Agarose Gel Cassette #CDF1510; Sage Sciences). Library pools were run on an Illumina HiSeq 4000 platform using the 150bp paired end (PE) Cluster Kit.

Reads alignment and variant calling

As in Yassin et al. (2016) and Ferreira et al. (2021), we used previously published genomic sequences from two continental *D. y. yakuba* populations, namely Cameroon (CY) and Kenya (NY) (Rogers et al. 2014), for comparisons with *D. y. mayottensis*. Each population consisted of 10 inbred lines that were sequenced individually. We also used previous Pool-Seq reads from a single *D. y. mayottensis* population collected from Soulou in 2013 for temporal comparisons within the subspecies (Yassin et al. 2016). For all comparisons, reads were mapped to the *D. y. yakuba* v.1.05 reference genome obtained from Flybase (<https://flybase.org/>, Thurmond et al. 2019) using Minimap2 software package (Li 2018). Minimap2-generated SAM files were converted to BAM format using samtools 1.9 software (Li et al. 2009) and after keeping with mapping quality of ≥ 20 . The BAM files were then cleaned and sorted using Picard v.2.0.1 (<http://broadinstitute.github.io/picard/>). Popoolation 2 software package v.1.20162 (Kofler et al. 2011) was used to generate a synchronized mpileup file, with the argument --min-quality 20 to retain only bases with quality scores ≥ 20 . Ind-Seq and Pool-Seq data were genotyped using customized perl scripts associated with this manuscript. In essence, Ind-Seq lines were genotyped after only retaining sites with a minimum and maximum depths 10 and 100, respectively, since the average depth was $\sim 40x$ for these libraries. The minimum and maximum depths for the Pool-Seq libraries were 20 and 200, respectively, because of their higher sequencing depth ($\sim 80x$). A variant was called if the minor allele frequency (MAF) exceeded 25% or $1/2N$ after depth correction for Ind-Seq or Pool-Seq sequences, respectively, with N being the number of individuals in a pool.

209

210 *Population structure analysis in D. yakuba*

211 We used the rmaverick v.1.10 package
212 (<https://bobverity.github.io/rmaverick/index.html>), derived from the Maverick software
213 (Verity and Nichols 2016), to investigate population structure among the three *D. y.*
214 *mayottensis* populations in addition to the two *D. y. yakuba* populations. We extracted bi-allelic
215 single nucleotide polymorphism (SNP) sites that were genotyped in all of the 19 Ind-Seq *D. y.*
216 *mayottensis* and 20 *D. y. yakuba* lines and with a minor allele frequency across all lines > 20%.
217 Ind-Seq data were used because they allow individual genotyping. The file was then parsed and
218 one site every 100 SNPs was sampled, leading to a total of 2,897 presumably independent sites.
219 The program estimates evidence for 1 to 5 subpopulations (k) and the level of admixture (α)
220 using the generalized thermodynamic integration (GTI) method. We conducted the MCMC
221 runs with 10,000 burn-in and 2000 sampling iterations on 50 rungs of the temperature ladder
222 for each k and tested for convergence every 100 iterations. MCMC performance and the GTI
223 path were checked for every k , and the log-evidence was compared between the different values
224 of k to identify the most likely number of subpopulations. The posterior allocations of the
225 inferred subpopulations across the five localities were then plotted.

226

227 *Delimitation of the GIDs in D. y. mayottensis*

228 To estimate the number of segregating sites (S) and between-populations genetic
229 differentiation (F_{ST}) from Ind-Seq and Pool-Seq synchronized files, we used a series of
230 customized perl scripts all available on the Figshare folder associated with this manuscript. Due
231 to convergence of results (see below), Pool-Seq data were presented for the GIDs analyses since
232 they have a larger sampling size. We conducted the Population Branch Excess (PBE) analysis
233 (Yassin et al. 2016) to identify GIDs specifically distinguishing *D. y. mayottensis* from the two

continental populations CY and NY. *PBE* is a standardized version of the Population Branch Statistic (*PBS*), which was proposed to identify population-specific allelic changes from F_{ST} pairwise comparisons between three populations (Yi et al. 2010). The standardization measures how *PBS* at a locus deviates from the chromosome-wide median *PBS* divided by the median F_{ST} estimate from the two non-focal populations. Therefore, *PBE* is less sensitive than *PBS* to genome-wide variation in polymorphism levels. Due to the lack of genetic structure within *D. y. mayottensis* (see below), Pool-Seq reads of the three populations were pooled into a single Pool of Pools (PoP). *PBS* and *PBE* were made assuming an (PoP,(CY,NY)) tree and were estimated for non-overlapping 10-kb windows. For a given window on a chromosomal arm:

$$PBS = \frac{-\log(1 - F_{ST_{PoP_CY}}) - \log(1 - F_{ST_{PoP_NY}}) + \log(1 - F_{ST_{CY_NY}})}{2}$$

And

$$PBE = PBS - \left[-\log(1 - F_{ST_{CY_NY}}) \times \left(\frac{PBS_{median}}{-\log(1 - F_{ST_{CY_NY}})_{median}} \right) \right]$$

We delimited GIDs following common approaches in peak calling based on F_{ST} outliers, *i.e.* by defining a chromosome-wide quantile threshold (σ) above which a window is considered candidate for selection, a radius (d) in bp (e.g., Schlamp et al. 2016). Windows with *PBE* values $\geq 97.5\%$ quantiles were considered outliers, and the radius of a genomic island of divergence was considered if two outlier windows did not fall more than 500-kb apart. Gene enrichment tests were conducted for each GID as well as using the whole list of genes from all GIDs using the online Database for Annotation, Visualization and Integrated Discovery (DAVID) (Sherman et al. 2022). We used the Functional Annotation Chart of DAVID which provides enrichment analysis using a modified Fisher Exact Test to identify the most overrepresented annotation terms associated with a gene list (here *D. melanogaster* orthologs) as compared to a background gene set, thereby giving insight into the main biological themes associated with a GID.

Defining low diversity and recombining regions and testing their effects on differentiation

For each 10-kb window, we counted the number of SNPs segregating in at least one of the three populations, *i.e.* CY, NY and PoP. We conducted a mixture analysis on the number of SNPs per window using the `normalmixEM` function in the `mixtools` package (Benaglia et al. 2010) in R v.4.3 (R Core Team 2023). This revealed two distributions with the lower almost ending around 500 SNPs (Supplementary Figure S1). We considered this value as a cutoff for low diversity and pooled windows of low diversity as long as they were not separated by more than 50-kb intervals, even if such short intervals had slightly more than 500 SNPs. To test the effect of low diversity on GIDs, we coded windows in the low- and high-diversity regions as 0 and 1, respectively, and conducted an asymptomatic two-sample Fisher-Pitman permutation test using the `oneway_test` function in the `coin` package (Hothorn et al. 2006) in R.

Yassin et al. (2016) estimated regions with low diversity in the CY population as a proxy of the recombination rate in *D. yakuba*. More recently, Pettie et al. (2022) estimated the recombination landscape in this species from crosses between three inbred lines from West and East Africa. The authors used a reference genome version, Dyak2.0, that is newer than the one used by Yassin et al. (2016) and us for population genomics analyses, *i.e.* Dyak1.05. To convert Pettie et al.'s (2022) estimates to the 1.05 version, we mapped *D. yakuba* translated protein sequences of *D. yakuba* v.1.05. reference genome (obtained from Flybase) to both the v.1.05 and v.2.1 reference genomes, using `miniprot2` software package (Li 2023). We identified low recombining regions as windows with crossover rates of ≤ 4 cM/Mb from Figure 3 in Pettie et al. (2022). To test the effect of low recombination on diversity and GIDs, we coded windows in the low- and high-recombination regions as 0 and 1, respectively, and conducted the asymptomatic two-sample Fisher-Pitman permutation test using the `oneway_test` function in the `coin` package (Hothorn et al. 2006) in R as stated above.

De novo assembly of D. y. mayottensis genome and identification of chromosomal rearrangements

Offspring of the isofemale lines of *D. y. mayottensis* collected in 2017 were pooled and maintained on a standard *Drosophila* medium as a single mass culture. Thirty females randomly sampled from this mass culture were used to assemble a reference genome for this subspecies. Genomic DNA was extracted using the Nucleobond AXG20 kit and buffer set IV from Macherey-Nagel (ref. 740544 and 740604, <https://www.mn-net.com>, Düren, Germany), and size selection was conducted using the SRE XS kit from Circulomics (<https://www.circulomics.com/>, Baltimore, Maryland, USA). Fragments less than 10 Kb were eliminated. The SQK-LSK109 kit from Oxford nanopore technology (Lu et al. 2016), <https://nanoporetech.com/>) was then used to prepare the samples for nanopore sequencing following manufacturer's protocol. The library was then loaded on a R9.4 flow cell for sequencing on MinION. Raw data were basecalled with Guppy v4.0.11.

Illumina short-read sequences from the SOU population that was used in the Pool-Seq analysis have also been used to reconstruct our genome, using a hybrid assembly strategy, combining those short accurate Illumina reads to our long Nanopore reads. Illumina sequencing was realized by Novogene Company Limited (<https://en.novogene.com>, Cambridge, UK) on our previously extracted DNA. DNA sequences, from Nanopore and Illumina technologies, were then assembled using MaSuRCA genome software v3.4.0 with Cabog assembler (Zimin et al. 2013). The program processes Illumina short reads first using an average pair-end (PE) read length of 150, a jelly fish hash size of 3200000000, and a kmer size of 99 for deBruijn graph. Cabog uses the Illumina superreads to correct the Nanopore long reads and then assembles the corrected long reads. Assembly contiguity was checked using the N50 statistic

and its completeness was assigned by evaluating the degree of completeness of single-copy-genes conserved across Diptera using the BUSCO software package (Simão et al. 2015).

To identify chromosomal rearrangements, we mapped *D. yakuba* translated protein sequences of *D. yakuba* v.1.05. reference genome (obtained from Flybase) to both our *D. y. mayottensis* assembly and the v.1.05 reference genome, using miniprot2 software package (Li 2023). Miniprot2 alignments were summarized in Pairwise mApping Format (PAF) files, which were then processed using a customized perl script: First, the longest match of each protein was retained and the PAF files were joined. *D. y. mayottensis* scaffolds were then ordered according to the average midpoint positions of the orthologous sequences on the reference genome, and oriented according to the frequency of matching or mismatching sense signs. A dotplot was then drawn using R by assigning for each gene a single dot corresponding to its orthologous position on the ordered scaffolds of *D. y. mayottensis* and the reference genome. We used the cytological position of *D. melanogaster* orthologs of *D. yakuba* genes to delimit *D. y. yakuba* rearrangements described by Lemeunier and Ashburner (1976). Similar to telomeric/centromeric regions, we tested the effect of inversion on GIDs by coding the absence or presence of inversions as 0 or 1, respectively, and then conducting a permutation test using the coin package in R. All scripts related to this article were deposited in Figshare.

Identifying GIDs potentially due to selective sweeps in D. y. mayottensis

To detect for traces of potential selective sweeps, we scanned the genome for long tracts of homozygosity. We used the H-scan software package (Schlamp et al. 2016) which estimates for each SNP the average length of pairwise homozygosity tracts, known as the H-statistic. We run the program on the 19 unphased Ind-Seq lines from Mayotte due to the lack of genetic structure in *D. y. mayottensis* (see below). Only SNPs that were present in all 19 individuals were sampled. The program was run using default parameters, *i.e.* measuring tract lengths in

terms of base pairs (bp) rather than SNPs, because of the presumably recent divergence of *D. y. mayottensis* (Yassin et al. 2016). SNP-based methods can miss the detection of recent complete sweeps. As with *PBE*, we defined our outlier threshold as SNPs having *H* statistics $\geq$ 97.5% quantile. We then defined GIDs that were potentially formed by selective sweeps as those with *H* statistics above this threshold but falling neither in a low-recombining telomeric or centromeric region, nor in an inversion specific to *D. y. mayottensis*.

Results

Drosophila yakuba mayottensis is strictly associated with noni in Mayotte

In April 2017, we investigated the association between *D. y. mayottensis* and noni across the two main islands of Mayotte, *i.e.* Grande Terre and Petite Terre. Out of 14 collection sites, we identified noni trees in 9 locations, mostly near the coastline. However, the trees were not always at the same fruiting stage, with fallen ripe fruits found at only three locations, namely Soulou (the same locality for the 2013 sample, hereafter SOU), M'tsangamouji (MTS) in the humid northwest and Bambo Est (BE) in the arid southeast (Figure 1A). *D. y. mayottensis* flies were collected through net sweeping or direct aspiration over fallen, rotting noni fruits only at those three locations. No single *D. y. mayottensis* fly was collected using fermenting banana traps at those sites or at any other site on the island, although the noni trees at Bambo Est were present amid a banana plantation (Figure 1B). In West Africa, such traps often collect tens or hundreds of *D. y. yakuba* (Rio et al. 1983; Voudibio 1985; Llopart et al. 2005; Prigent et al. 2013; Turissini and Matute 2017) (Figure 1B; $\chi^2 P < 0.001$ for comparison with data from Voudibio (1985)). The total number of collected *D. y. mayottensis* was low, ~200 flies throughout the 10-days duration of the expedition, compared to ~600 flies that were collected during the last 2 days of the January 2013 at a single location (David et al. 2014). This indicates

a strong seasonality in the abundance of the species, consistent with other observations on *D. y. yakuba* on the continent (Vouldibio 1985; Prigent et al. 2013), and highlights the necessity of year-round investigations. We also observed that flies were more abundant on noni fruits that have been nibbled and thrown to the ground by the brown lemur (*Eulemur fulvus*) than on intact fallen fruits (Figure 1C).

Isofemale line sequencing of the three Mayotte populations reveal low level of differentiation on the island

We established 29, 22 and 26 isofemale lines from wild-caught flies from the three collection sites, *i.e.*, MTS, SOU and BE, respectively. As in Yassin et al. (2016), we sequenced a single F₁ fly per line in three pools according to the populations (Pool-Seq). The level of genetic differentiation estimated in 10-kb windows in SOU between the 2013 and 2017 samples was low ($F_{ST} = 0.0504 \pm 0.0001$; Supplementary Figure S2A), as well as the level of differentiation between the three 2017 geographical samples ($F_{ST} = 0.0450 \pm 0.0001$; Supplementary Figure S2B). The genetic differentiation among *D. y. mayottensis* populations was significantly lower than that between the two *D. y. yakuba* populations from Cameroon (CY) and Kenya (NY) ($F_{ST} = 0.0718 \pm 0.0003$; Student's *t*-test $P < 2.2 \times 10^{-16}$ for both analyses). To gain further insights on the population genetic structure both within and between *D. yakuba* subspecies, we sequenced 6, 6, and 7 individual isofemale lines from MTS, SOU, and BE, respectively (Ind-Seq). In agreement with the F_{ST} results, genetic structure analysis using *rmavrick* provided strong evidence for three subpopulations, with one subpopulation predominating in the three localities of *D. y. mayottensis* and a second subpopulation predominating in *D. y. yakuba* (Figure 2A,B). The third subpopulation mainly corresponded to a single inbred line from Cameroon (CY), which was highly homozygous for alleles alternative to the reference genome. Most importantly, the structure analysis supported a substantial level

of admixture (α) with a mode approaching 0.07 (Figure 2C), in agreement with previous demographic models (Yassin et al. 2016). Reconducting the analysis with a lower MAF threshold of 5% produced similar results for *D. y. mayottensis*, although no population structure was found for *D. y. yakuba*, i.e. $k = 2$ rather than 3 (Supplementary Figure S3).

The distribution of segregating sites shows large centromeric and telomeric effects

The distribution of SNPs per 10-kb windows identified long tracts of low diversity extending from either the telomere or the centromere, with the tracts being longer towards the telomere on the X chromosome and the centromeres for the autosomes (Figure 3A, as in Yassin et al. 2016; Pettie et al. 2022). We also identified a short tract of low diversity in the middle of the X chromosome between 9.62-10.94. This tract likely colocalizes with a small heterochromatic block that was previously cytogenetically identified (Lemeunier et al. 1978; Roy et al. 2005). A fine-map of recombination rates in *D. y. yakuba* also identified a valley of low recombination in this area (Fig. 3 in Pettie et al. 2022). Therefore, we also considered this tract as a low-recombining region. Asymptomatic two-sample Fisher-Pitman permutation test permutation showed strong overlap between population genetic diversity estimates and recombination rates ($Z = -86.49$; $P < 2.2 \times 10^{-6}$) as expected from theory and postulated by Yassin et al. (2016).

*Differentiation between the two *D. yakuba* subspecies identified 23 GIDs specific to Mayotte*

To quantify the level of genetic differentiation between Mayotte and mainland populations, we estimated F_{ST} in 10-kb windows across the genome, from which the *PBE* statistics were derived to detect windows specific to Mayotte. Because of the lack of genetic structure among the three populations of *D. y. mayottensis*, we pooled either the Ind-Seq or the

Pool-Seq sequences of the three populations in a single population, referring to them as Pool of Individuals (PoI) and Pool of Pools (PoP), respectively. Results did not greatly differ between the two methods (Supplementary Figure S4), and we will show here analyses on the PoP data due to its larger sampling size and consequently reduced ascertainment bias. The *PBE* analysis showed 23 GIDs falling above the 97.5% quantiles in *D. y. mayottensis* (Figure 3B; Table 1). Eight GIDs were enriched for specific gene families (Table 1). Gene families that were enriched for the whole list of genes from all GIDs included LITAF (LPS-induced TNF-activating factor) endosome-associated membrane proteins, glucose-methanol-choline (GMC) oxidoreductases, cytochrome P450, UDP-glycosyltransferases, maltose-alpha-glucosidases, serine-type endopeptidases, and glutathione-S-transferases at $P < 0.05$ after False Discovery Rate (FDR) correction for multiple testing. All these gene families have major roles in the metabolism of ecdysteroids, lipids, sugars, proteins and xenobiotics, as it would be expected in cases of major dietary shifts. Asymptotic two-sample Fisher-Pitman permutation test also confirmed the strong correlation between *PBE* and diversity ($Z = 21.43$, $P < 2 \times 10^{-16}$) as well as recombination ($Z = 15.97$, $P < 2 \times 10^{-16}$). Indeed, 16 out of 23 GIDs fell within low-recombining telomeric or centromeric region, and one GID fell 0.43 Mb apart from the low recombining region in the middle of the X chromosome.

GIDs in D. y. mayottensis predominantly overlap with chromosomal inversions

Our hybrid assembly of the *D. y. mayottensis* genome yielded 563 fragments with an N50 of 6.9Mb, and a BUSCOv5 score of 99.5%. Comparing the assembly to the v.1.05 reference genome assembly of *D. y. yakuba* identified 10 large (>100 kb-long) structural variations with a characteristic contrary diagonal-shape (solid arrows in Figure 4A). The intersection of these contrary diagonals corresponded to the breakpoints of 6 inversions and

their numbers depend on the length of the inversion and the evolutionary trajectory separating the configurations of the two assemblies (Figure 4; Supplementary Figure S5).

The first contrary diagonal intersection was found on the X chromosome and indicated a ~1 Mb-long inversion between X:8,648,348..9,652,060, just adjacent to the heterochromatic block in the middle of this chromosome. Lemeunier and Ashburner (1976) did not detect any inversion on chromosome X in *D. y. yakuba*. Following their nomenclature, we call this new subspecies-specific inversion *Xtb*, and note that it overlaps with *GID2*. We also noted a collinear connection between X:13,477,771..17,070,823 on a single scaffold (grey) that could have resulted from double inversions, but no direct evidence for breakpoints was observed at this region (Figure 4A).

The second and third contrary diagonal intersections were found on chromosomal arm 2L and corresponded to a large, ~12 Mb-long inversion between 2L:8423810..20352247. The distal breakpoint of this inversion, *i.e.* the second contrary diagonal, fell in the middle of an artificial inversion that has long been recognized as a misassembled portion of this chromosomal arm in the reference genome v. 1.05 between 8,586,415 and 11,101,068 (Cande et al. 2012), and which have been corrected in newer versions of this genome (Miller et al. 2018) (Supplementary Figure S6). The breakpoints of this artificial inversion correspond to two scaffolds in the *D. y. mayottensis* assembly (indicated with blank arrows in Figure 4A) and are delimited in Figure 4B. Lemeunier and Ashburner (1976) found four polymorphic inversions in mainland *D. y. yakuba*, namely *2Lh*, *2Ll*, *2Lm* and *2Ln*, with *2Lh* being ancestral to *2Ll*, and *2Lm* and *2Ln* independently derived from *2Ll* (Figure 4B). We found the large inversion in *D. y. mayottensis* to be independently derived from *2Ll* too, and called it *2Lo* (Figure 4B). The proximal end of this inversion falls within the large subcentromeric region of 2L. Compared to the two most common inversions of the mainland, *i.e.* *2Lh* and *2Ll* (Lemeunier and Ashburner 1976), four GIDs (5 to 8) fall in an inverted order in *2Lo* (Table 1).

The fourth contrary diagonal intersection fell in the subcentromeric region of chromosomal arm 2R at 2R:182,816..436,892 but did not overlap with a major GID. It hence corresponded to a new inversion that we call *2Ro*. The fifth to the eight contrary diagonal intersections fell within the normally recombining region of 2R. Their cytogenetic positions corresponded to the breakpoints separating chromosomal inversions *2Rn* and *2Rj* (Lemeunier and Ashburner 1976). The reference genome has the *2Rn* order indicating that *2Rj* (2R:8,717,812..18,133,566), the most common inversion in *D. yakuba* (Lemeunier and Ashburner 1976), is present in *D. y. mayottensis*. Because we have used a mass culture of *D. y. mayottensis* following the pooling of multiple isofemale lines, and because of the absence of excess variability at the breakpoints of this inversion among the three populations of Mayotte, *2Rj* is likely fixed in this subspecies. Both *2Rn* and *2Rj* inversions independently derived from a hypothetical *2Ri* order (2R:11,157,330..18,066,346) that has never been observed in natural populations (Figure 4B; Lemeunier and Ashburner 1976). The fifth and the eight contrary diagonal intersections correspond to the breakpoints of the *2Rj* inversion relative to *2Ri*, whereas the sixth and the seventh contrary diagonal intersections correspond to the breakpoints of the *2Rn* inversion relative to *2Ri*. Two GIDs, 11 and 14, fell ~1.7 Mb apart from the breakpoints separating *2Rj* from *2Ri*, whereas GID12 and 13 overlapped with the breakpoints separating *2Rn* from *2Ri*. Remarkably, the large GID13 is inverted in inversion *2Rk* (2R:10,365,887..14,704,259), another common 2R inversion in *D. y. yakuba* that was not detected in *D. y. mayottensis* (Figure 4B; Table 2).

The ninth and tenth contrary diagonal intersections were found in the subcentromeric regions of chromosomal arms 3L and 3R, respectively. Each corresponded to ~0.2 and 0.8 Mb-long inversions and overlapped with GIDs 16 and 17, respectively. These small subcentromeric inversions were not detected by Lemeunier and Ashburner (1976) and we denoted them *3Lk* (3L:23,737,761..23,929,648) and *3Ri* (3R:1..818,622), respectively (Figure 4A; Table 2).

Permutation analysis showed a significant overlap between GIDs and inversions ($Z = -22.17$, $P < 2.2 \times 10^{-16}$). Summary of the eight inversions associated with GIDs, *i.e.* corresponding to the six inversions between *D. y. mayottensis* and *D. y. yakuba* assemblies as well as the hypothetical *2Ri* and the *D. y. yakuba*-specific *2Rk* not present in the reference genome, along with their geographical distribution, is given in Table 2.

Evidence for selective sweeps within GIDs in normally-recombining, collinear regions

To gain further insights on the extent of potential selective sweeps within the GIDs distinguishing *D. y. mayottensis* from *D. y. yakuba*, we conducted the H-scan analysis which estimates the length of homozygosity tracts for each SNP within *D. y. mayottensis*. Both selective sweeps and low recombination are expected to lead to high H values. As expected, H was highly correlated with low recombining regions (Figure 5A). The highest H value outside the low-recombining regions was found on the GID overlapping with the subspecies-specific *Xtb* inversion, a region that is itself adjacent to the low-recombining heterochromatic block in the middle of the X chromosome. The highest H values on the autosomes coincided with the breakpoints of the subspecies-specific *2Lo* and the ancestral *2Ri* inversion. All GIDs except one, GID10, had H values above the 97.5% quantile (Figure 5A). These included the two GIDs falling in normally-recombining regions and not overlapping with inversions, *i.e.* GID4 and GID23, both had high H values, supporting the action of a hard selective sweep. To further identify the target of such potential sweeps, we investigated H profile at each SNP in the two GIDs. For GID4, the highest H fell adjacent to a gene orthologous to *D. melanogaster* *CG17242* (Figure 5B). This gene is a serine protease homolog of the S1A family. It is a seminal fluid protein that affects male reproductive success (Patlar and Civetta 2022), and whose rate of sequence evolution provided evidence for positive selection in other *Drosophila* species (Wong et al. 2012). For GID23, the highest H fell in the mid of a cluster of *Osiris* paralogs (Figure 5C),

whose implication in underlying resistance to octanoic acid, noni's main toxin, has been demonstrated in *D. melanogaster* and *D. sechellia* (Hungate et al. 2013; Andrade López et al. 2017; Lanno et al. 2017, 2019), hence providing further support for the action of a selective sweep.

Discussion

Identifying the loci underlying population divergence at the early stages of speciation has long been a central topic in evolutionary genetics. The observation that the genomic landscape of divergence at these stages usually forms few islands provided an initial promise, but the enthusiasm has rapidly faded by the realization that such islands can also “accidentally” arise by a plethora of neutral processes with no obligatory impact on reproductive isolation or habitat specialization. Therefore, the ability to distinguish islands whose divergence signal can be genuinely attributed to selection from those whose signals may have been driven by other confounding, mainly structural effects, is an important challenge. In this study, we combined field observation with population and structural genomics, to investigate the potential mechanisms underlying the evolution of GIDs in the noni-specialist subspecies *Drosophila yakuba mayottensis*. First, we showed that *D. y. mayottensis* GIDs were subspecies-specific. There was no different pattern of divergence whether F_{ST} was inferred by comparing mainland populations to the Soulou population collected in January 2013 or April 2017, or to any other noni-associated *D. y. mayottensis* population collected in 2017 across the island. Second, we found that the two subspecies had distinct inversions. Of the 23 GIDs distinguishing *D. y. mayottensis* from two mainland *D. y. yakuba* populations, 21 fell either in low-recombining telomeric or centromeric regions, or close to or within subspecies-specific chromosomal

inversions, corresponding to 36 and 63% of the outlier windows in the *PBE* analysis, respectively.

Chromosomal inversions can play an important role in promoting local adaptation, either through reducing gene flow at selected genomic regions or through bringing to a new habitat an advantageous allelic combination segregating at the ancestral range (Faria et al. 2019; Mérot et al. 2020; Schaal et al. 2022). Depending on the age of the inversion, selection can lead to the formation of GIDs within the inversions or near their breakpoints (Guerrero et al. 2012; Fuller et al. 2017; Kapun et al. 2023). GIDs falling in the subspecies-specific inversions *Xtb* and *2Lo* include several ecologically-relevant candidates that are involved in the metabolism of ecdysone, lipids and carbohydrates (GMC oxidoreductases (Iida et al. 2007; Glaser-Schmitt and Parsch 2018); maltases (Inomata et al. 2019)), ethanol tolerance (*Adh* (Siddiq and Thornton 2019)), and hexanoic acid sensing (*pdm3* (Arguello et al. 2021)). The *2Rj* inversion that is fixed in *D. y. mayottensis* is, on the other hand, widespread in *D. y. yakuba*, and it is ancestral to three 2R inversions that are specific to the mainland subspecies (Lemeunier and Ashburner 1976). GIDs near the breakpoints of this ancestral inversion contain important clusters of detoxification genes of cytochrome P450 oxidases and glutathione-S-transferases, which are involved in basal resistance to noni toxins in *D. sechellia* and its close relatives (Peyser et al. 2017; Lanno and Coolon 2019). Alternatively, both derived and shared inversions might have been fixed by drift, although demographic models do not suggest major bottlenecks during *D. y. mayottensis* colonization of Mayotte (Yassin et al. 2016). Testing the potential role of derived and shared inversions would require experimental isolation of each karyotype and comparison of the effect of each on different noni use traits, but the large breadth of these inversions would complicate fine-scale QTL mapping or genome-wide association studies (GWAS) in *D. yakuba*.

High levels of differentiation in collinear, high recombining regions constitute strong signals of hard selective sweeps. Only 2 GIDs (4 and 23) satisfied these conditions. Most

notably, GID23 consisted of syntenic members of the *Osiris* gene family. The role of these genes in conferring resistance against noni's main toxin, octanoic acid, has been demonstrated through genetic mapping (Hungate et al. 2013) and transcriptomic analyses in *D. sechellia* (Lanno et al. 2017), as well as functional gene silencing in *D. melanogaster* (Andrade López et al. 2017; Lanno et al. 2019). The other GID also involves genes with a potential role in resistance to plant chemical defenses in other insects, such serine-type endopeptidases (Hoang et al. 2015; Alyokhin and Chen 2017; Lanno and Coolon 2019). These genes therefore are best candidates for future functional genetics analyses. However, hard selective sweeps indicate a particular mode of selection wherein the selected alleles were absent or had low frequency at the onset of selection. Alternatively, incomplete selective sweeps wherein the selected alleles may have not reached fixation yet or soft selective sweeps on alleles segregating at intermediate frequencies in the ancestral range could also be important factors, although loci under these patterns of sweeps will hardly be detected by GID analyses. A signal of soft selective sweep on the odorant receptor *Or22a*, a gene that controls *D. sechellia* attraction to methyl hexanoate, the main noni volatile (Auer et al. 2020), was recently demonstrated in *D. y. mayottensis* (Ferreira et al. 2021). However, the efficiency of this approach will be limited to collinear, normally recombining regions, which constitute a relatively small proportion of the *D. yakuba* genome.

Identifying the type of selection pressures responsible of GIDs formation may be difficult. Whereas we focused here on the adaptations to the highly toxic noni fruit, we cannot rule out that other environmental factors have also forged the divergence landscape in *D. yakuba*. For example, GID2 which is associated with the *Xtb* inversion encompasses *Flotillin2* (*Flo2*), a gene that underlies the defensive melanic response against parasitoids (Howell et al. 2012) and indeed collected *D. y. mayottensis* flies had high incidences of melanic capsules sequestering dead larvae of *Leptopilina victoriae* (David et al. 2014). Similarly, the drosophilid

community, the level of urbanization, and the alternations of the rain and dry seasons greatly differ between the island and the mainland (Vouldibio 1985; Prigent et al. 2013; David et al. 2014). Besides, some GIDs may underlie the asymmetric precopulatory reproductive isolation between the two subspecies (Yassin et al. 2016) or other sources of reproductive isolation that have not yet been identified. For example, the serine-endopeptidases of GID4 are enriched in the female spermatheca and some, such as *Send*, are involved in the female postcopulatory sexual selection (McDonough-Goldstein et al. 2022). All these factors can constitute non-mutually exclusive selective pressures acting on the *D. y. mayottensis* genome.

Although it is highly recommended to infer demographic model to properly identify the signals of selective sweeps, the inference of such models is often complex for most organisms (Garud et al. 2021). In the case of *D. yakuba*, high depth genome sequences from mainland populations are available from two populations, each represented by only 10 inbred lines (Rogers et al. 2014). Low sampling and inbreeding can therefore influence the model, in particular through neglecting the effects of “ghost populations” on the rate of admixture and divergence time estimates (Hancock and Blackmon 2020; Tricou et al. 2022; Marchi et al. 2024). Further sampling of wild *D. y. yakuba* closer to the Comoros archipelago, such as in Mozambique and Madagascar, is highly needed. In addition, the timing of the arrival of the noni host plant, which is of Oriental origin, to the islands of the Western Indian Ocean, has long been debated (Lachaise and Silvain 2004). This information is crucial to infer and properly model the timing of the sweeps on loci potentially involved in the adaptation to the toxic host, but our knowledge of the phylogeography of the host plant in these islands is still limited.

Will early stages of divergence always lead to the formation of GIDs? Our conclusion suggests that this will predominantly depend on the genomic architecture, since ~99% of the outlier 10-kb windows in the GIDs distinguishing the two *D. yakuba* subspecies were found in low recombining telomeric or centromeric regions or large chromosomal inversions. The extent

of both phenomena largely depends on the species, even in cases of parallel responses to the same selective pressures. Indeed, shared polymorphic inversions had promoted parallel adaptations in different geographical populations in many cases (Mérot et al. 2018; Lucek et al. 2019; Morales et al. 2019; Matschiner et al. 2022). However, this situation cannot explain the recurrent specialization on toxic noni fruits in distinct islands of the Indian Ocean between *D. yakuba* and *D. sechellia*. Despite their striking host plant convergence (Yassin et al. 2016; Ferreira et al. 2021), the divergence between the two species is deep and their genomic architectures are quite different. Nearly 35.3% of the genome in *D. yakuba* has low number of segregating sites extending from the telomeres and centromeres (Yassin et al. 2016; Pettie et al. 2022). In *D. simulans* and *D. mauritiana*, only 10.5% of the genome has low number of segregating sites (Brand et al. 2018). The three species of the *simulans* clade are homosequential and chromosomal inversions were very rarely reported in *D. simulans* (Chakraborty et al. 2021b). As functional genetics analyses will continue to be undergone in both species, the degree of genic parallelism underlying the recurrent specialization on a toxic resource despite distinct genomic architectures will be clarified.

Advances in third-generation sequencing currently allow the construction of highly contiguous assemblies, hence providing better understanding of the role of structural variation in GIDs formation. Although our assembly for *D. y. mayottensis* based on the hybrid approach of both short and long reads did not lead to the inference of whole chromosomes, all chromosomal arms were represented by few, long scaffolds hence giving us confidence with the identification of the breakpoints of the new inversions. Indeed, our assembly well identified on a single scaffold the breakpoints of large inversions previously characterized by classical cytogenetic analyses. As new genomic approaches to produce chromosome-level genome assemblies are being produced in multiple non-model organisms, our understanding of the extent of the impact of the telomeres, centromeres and structural variation on GIDs formation

629 will continue to improve. Our quantification of this effect to reach ~99% of outlier windows in
630 *D. yakuba* is a significant step towards this goal, but further similar studies in other organisms
631 with distinct architectures will show how much our finding can be generalized.

632

633

Authors contributions

E.A.F. sequenced and assembled new genome, conducted population genomics and structural variation analyses, wrote the primary draft of the paper. C.C.M. sequenced individual lines. D.O. sequenced and assembled new genome. A.S. collected flies. S.R.P. taxonomically sorted flies. R.L.R. sequenced individual lines, obtained funds. A.Y. conceptualized the work, collected flies, taxonomically sorted flies, supervised population genomics and structural variation analyses, obtained funds, wrote the final version of the paper with inputs from all authors.

Acknowledgments

The authors are grateful to Claire Goll  ty for help identifying noni plantation locations in Mayotte. Our late colleague, Jean R. David (1931-2021), who initially discovered *D. y. mayottensis* in 2013, participated to the 2017 expedition.

Data availability statement

Scripts used and data generated in this study are publicly available on Figshare (<https://doi.org/10.6084/m9.figshare.c.7070498.v1>). Genome sequences generated for this study are available on NCBI BioProject: PRJNA972991, PRJNA973002 and PRJNA973271.

Funding statement

This work was funded by the Agence Nationale de la Recherche grant no. ANR-18-CE2-0008, a Richard Lounsbery Foundation grant and an ATM bursary from the Mus  um National d'Histoire Naturelle to AY, and a National Institutes of Health grant NIH NIGMS R35 GM133376 to RLR.

659 **Benefits generated**

660 Benefits from this research accrue from the sharing of our data and results on public databases
661 as described above.

662

663 **Conflict of interest**

664 The authors declare no conflict of interest.

665

666 **References**

- 667 Alyokhin, A., and Y. H. Chen. 2017. Adaptation to toxic hosts as a factor in the evolution of
668 insecticide resistance. *Current Opinion in Insect Science* 21:33–38.
- 669 Amlou, M., B. Moreteau, and J. R. David. 1998. Genetic Analysis of *Drosophila sechellia*
670 Specialization: Oviposition Behavior Toward the Major Aliphatic Acids of Its Host Plant.
671 *Behav Genet* 28:455–464.
- 672 Amlou, M., E. Pla, B. Moreteau, and J. R. David. 1997. Genetic analysis by interspecific crosses
673 of the tolerance of *Drosophila sechellia* to major aliphatic acids of its host plant. *Genet Sel Evol*
674 29:511.
- 675 Andrade López, J. M., S. M. Lanno, J. M. Auerbach, E. C. Moskowitz, L. A. Sligar, P. J.
676 Wittkopp, and J. D. Coolon. 2017. Genetic basis of octanoic acid resistance in *Drosophila*
677 *sechellia*: functional analysis of a fine-mapped region. *Mol Ecol* 26:1148–1160.
- 678 Arguello, J. R., L. Abuin, J. Armida, K. Mika, P. C. Chai, and R. Benton. 2021. Targeted
679 molecular profiling of rare olfactory sensory neurons identifies fate, wiring, and functional
680 determinants. *eLife* 10:e63036. eLife Sciences Publications, Ltd.
- 681 Auer, T. O., M. A. Khallaf, A. F. Silbering, G. Zappia, K. Ellis, R. Álvarez-Ocaña, J. R.
682 Arguello, B. S. Hansson, G. S. X. E. Jefferis, S. J. C. Caron, M. Knaden, and R. Benton. 2020.
683 Olfactory receptor and circuit evolution promote host specialization. *Nature* 579:402–408.
- 684 Auer, T. O., M. P. Shahandeh, and R. Benton. 2021. *Drosophila sechellia*: A Genetic Model for
685 Behavioral Evolution and Neuroecology. *Annual Review of Genetics* 55:527–554.
- 686 Benaglia, T., D. Chauveau, D. R. Hunter, and D. S. Young. 2010. mixtools: An R Package for
687 Analyzing Mixture Models. *Journal of Statistical Software* 32:1–29.
- 688 Berdan, E. L., A. Blanckaert, R. K. Butlin, and C. Bank. 2021. Deleterious mutation
689 accumulation and the long-term fate of chromosomal inversions. *PLOS Genetics* 17:e1009411.
690 Public Library of Science.

691 Booker, T. R., S. Yeaman, and M. C. Whitlock. 2020. Variation in recombination rate affects
 692 detection of outliers in genome scans under neutrality. *Molecular Ecology* 29:4274–4279.

693 Brand, C. L., M. V. Cattani, S. B. Kingan, E. L. Landeen, and D. C. Presgraves. 2018. Molecular
 694 Evolution at a Meiosis Gene Mediates Species Differences in the Rate and Patterning of
 695 Recombination. *Current Biology* 28:1289-1295.e4.

696 Cande, J., P. Andolfatto, B. Prud'homme, D. L. Stern, and N. Gompel. 2012. Evolution of
 697 Multiple Additive Loci Caused Divergence between *Drosophila yakuba* and *D. santomea* in
 698 Wing Rowing during Male Courtship. *PLOS ONE* 7:e43888. Public Library of Science.

699 Chakraborty, M., C.-H. Chang, D. E. Khost, J. Vedanayagam, J. R. Adrion, Y. Liao, K. L.
 700 Montooth, C. D. Meiklejohn, A. M. Larracuente, and J. J. Emerson. 2021a. Evolution of
 701 genome structure in the *Drosophila simulans* species complex. *Genome Res.* 31:380–396.

702 Chakraborty, M., C.-H. Chang, D. E. Khost, J. Vedanayagam, J. R. Adrion, Y. Liao, K. L.
 703 Montooth, C. D. Meiklejohn, A. M. Larracuente, and J. J. Emerson. 2021b. Evolution of
 704 genome structure in the *Drosophila simulans* species complex. *Genome Res.* 31:380–396.

705 Cruickshank, T. E., and M. W. Hahn. 2014. Reanalysis suggests that genomic islands of
 706 speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology* 23:3133–
 707 3157.

708 David, J. R., A. Yassin, N. Gidaszewski, and V. Debat. 2014. *Drosophilids* (Diptera) from
 709 Mayotte island: an annotated list of species collected in 2013 and comments on the colonisation
 710 of Indian Ocean Islands. *Annales de la Société entomologique de France* (N.S.) 50:336–342.

711 Drum, Z., S. Lanno, S. M. Gregory, S. Shimshak, W. Barr, A. Gatesman, M. Schadt, J. Sanford,
 712 A. Arkin, B. Assignon, S. Colorado, C. Dalgarno, T. Devanny, T. Ghandour, R. Griffin, M.
 713 Hogan, E. Horowitz, E. McGhie, J. Multer, H. O'Halloran, K. Ofori-Darko, D. Pokushalov, N.
 714 Richards, K. Sagarin, N. Taylor, A. Thielking, P. Towle, and J. Coolon. 2022. Genomics
 715 analysis of *Drosophila sechellia* response to *Morinda citrifolia* fruit diet. *G3*

716 Genes|Genomes|Genetics 12:jkac153.

717 Earley, E. J., and C. D. Jones. 2011. Next-Generation Mapping of Complex Traits with
 718 Phenotype-Based Selection and Introgression. *Genetics* 189:1203–1209.

719 Ellegren, H., L. Smeds, R. Burri, P. I. Olason, N. Backström, T. Kawakami, A. Künstner, H.
 720 Mäkinen, K. Nadachowska-Brzyska, A. Qvarnström, S. Uebbing, and J. B. W. Wolf. 2012a.
 721 The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* 491:756–760.
 722 Nature Publishing Group.

723 Ellegren, H., L. Smeds, R. Burri, P. I. Olason, N. Backström, T. Kawakami, A. Künstner, H.
 724 Mäkinen, K. Nadachowska-Brzyska, A. Qvarnström, S. Uebbing, and J. B. W. Wolf. 2012b.
 725 The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* 491:756–760.
 726 Nature Publishing Group.

727 Faria, R., K. Johannesson, R. K. Butlin, and A. M. Westram. 2019. Evolving Inversions. *Trends*
 728 in *Ecology & Evolution* 34:239–248.

729 Feder, J. L., and P. Nosil. 2009. Chromosomal inversions and species differences: when are
 730 gene affecting adaptive divergence and reproductive isolation expected to reside within
 731 inversions? *Evolution* 63:3061–3075.

732 Feder, J. L., and P. Nosil. 2010. The efficacy of divergence hitchhiking during ecological
 733 speciation. *Evolution* 64:1729–1747.

734 Ferreira, E. A., S. Lambert, T. Verrier, F. Marion-Poll, and A. Yassin. 2021. Soft Selective
 735 Sweep on Chemosensory Genes Correlates with Ancestral Preference for Toxic Noni in a
 736 Specialist *Drosophila* Population. *Genes* 12:32. Multidisciplinary Digital Publishing Institute.

737 Forbes, A. A., S. N. Devine, A. C. Hippee, E. S. Tvedte, A. K. G. Ward, H. A. Widmayer, and
 738 C. J. Wilson. 2017. Revisiting the particular role of host shifts in initiating insect speciation.
 739 *Evolution* 71:1126–1137.

740 Fuller, Z. L., G. D. Haynes, S. Richards, and S. W. Schaeffer. 2017. Genomics of natural

741 populations: Evolutionary forces that establish and maintain gene arrangements in *Drosophila*
 742 *pseudoobscura*. *Molecular Ecology* 26:6539–6562.
 743 Futuyma, D. J., and G. Moreno. 1988. The Evolution of Ecological Specialization. *Annual*
 744 *Review of Ecology and Systematics* 19:207–233.
 745 Garrigan, D., S. B. Kingan, A. J. Geneva, P. Andolfatto, A. G. Clark, K. R. Thornton, and D.
 746 C. Presgraves. 2012. Genome sequencing reveals complex speciation in the *Drosophila*
 747 *simulans* clade. *Genome Res.* 22:1499–1511.
 748 Garud, N. R., P. W. Messer, and D. A. Petrov. 2021. Detection of hard and soft selective sweeps
 749 from *Drosophila melanogaster* population genomic data. *PLOS Genetics* 17:e1009373. Public
 750 Library of Science.
 751 Glaser-Schmitt, A., and J. Parsch. 2018. Functional characterization of adaptive variation
 752 within a cis-regulatory element influencing *Drosophila melanogaster* growth. *PLOS Biology*
 753 16:e2004538. Public Library of Science.
 754 Guerrero, R. F., F. Rousset, and M. Kirkpatrick. 2012. Coalescent patterns for chromosomal
 755 inversions in divergent populations. *Philosophical Transactions of the Royal Society B:*
 756 *Biological Sciences* 367:430–438. Royal Society.
 757 Hager, E. R., O. S. Harringmeyer, T. B. Wooldridge, S. Theingi, J. T. Gable, S. McFadden, B.
 758 Neugeboren, K. M. Turner, J. D. Jensen, and H. E. Hoekstra. 2022. A chromosomal inversion
 759 contributes to divergence in multiple traits between deer mouse ecotypes. *Science* 377:399–
 760 405. American Association for the Advancement of Science.
 761 Hancock, Z. B., and H. Blackmon. 2020. Ghosts of a Structured Past: Impacts of Ancestral
 762 Patterns of Isolation-by-Distance on Divergence-Time Estimation. *Journal of Heredity*
 763 111:573–582.
 764 Harringmeyer, O. S., and H. E. Hoekstra. 2022. Chromosomal inversion polymorphisms shape
 765 the genomic landscape of deer mice. *Nat Ecol Evol* 6:1965–1979. Nature Publishing Group.

766 Hoang, K., L. M. Matzkin, and J. M. Bono. 2015. Transcriptional variation associated with
 767 cactus host plant adaptation in *Drosophila mettleri* populations. *Molecular Ecology* 24:5186–
 768 5199.

769 Hothorn, T., K. Hornik, M. A. van de Wiel, and A. Zeileis. 2006. A Lego System for
 770 Conditional Inference. *The American Statistician* 60:257–263. Taylor & Francis.

771 Howell, L., C. J. Sampson, M. J. Xavier, E. Bolukbasi, M. M. S. Heck, and M. J. Williams.
 772 2012. A directed miniscreen for genes involved in the *Drosophila* anti-parasitoid immune
 773 response. *Immunogenetics* 64:155–161.

774 Huang, K., R. L. Andrew, G. L. Owens, K. L. Ostevik, and L. H. Rieseberg. 2020. Multiple
 775 chromosomal inversions contribute to adaptive divergence of a dune sunflower ecotype.
 776 *Molecular Ecology* 29:2535–2549.

777 Huang, Y., and D. Erezyilmaz. 2015. The Genetics of Resistance to Morinda Fruit Toxin During
 778 the Postembryonic Stages in *Drosophila sechellia*. *G3: Genes, Genomes, Genetics* 5:1973–
 779 1981.

780 Hughes, S. E., and R. S. Hawley. 2020. Meiosis: Location, Location, Location, How Crossovers
 781 Ensure Segregation. *Current Biology* 30:R311–R313.

782 Hungate, E. A., E. J. Earley, I. A. Boussy, D. A. Turissini, C.-T. Ting, J. R. Moran, M.-L. Wu,
 783 C.-I. Wu, and C. D. Jones. 2013. A Locus in *Drosophila sechellia* Affecting Tolerance of a Host
 784 Plant Toxin. *Genetics* 195:1063–1075.

785 Iida, K., D. L. Cox-Foster, X. Yang, W.-Y. Ko, and D. R. Cavener. 2007. Expansion and
 786 evolution of insect GMC oxidoreductases. *BMC Evolutionary Biology* 7:75.

787 Inomata, N., K. R. Takahasi, and N. Koga. 2019. Association between duplicated maltase genes
 788 and the transcriptional regulation for the carbohydrate changes in *Drosophila melanogaster*.
 789 *Gene* 686:141–145.

790 Jay, P., M. Chouteau, A. Whibley, H. Bastide, H. Parrinello, V. Llaurens, and M. Joron. 2021.

791 Mutation load at a mimicry supergene sheds new light on the evolution of inversion
792 polymorphisms. *Nat Genet* 53:288–293. Nature Publishing Group.

793 Jay, P., E. Tezenas, A. Véber, and T. Giraud. 2022. Sheltering of deleterious mutations explains
794 the stepwise extension of recombination suppression on sex chromosomes and other
795 supergenes. *PLOS Biology* 20:e3001698. Public Library of Science.

796 Kapun, M., and T. Flatt. 2019. The adaptive significance of chromosomal inversion
797 polymorphisms in *Drosophila melanogaster*. *Molecular Ecology* 28:1263–1282.

798 Kapun, M., E. D. Mitchell, T. J. Kawecki, P. Schmidt, and T. Flatt. 2023. An Ancestral
799 Balanced Inversion Polymorphism Confers Global Adaptation. *Molecular Biology and*
800 *Evolution* 40:msad118.

801 Kofler, R., R. V. Pandey, and C. Schlötterer. 2011. PoPoolation2: identifying differentiation
802 between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*
803 27:3435–3436.

804 Koury, S. A. 2023. Predicting recombination suppression outside chromosomal inversions in
805 *Drosophila melanogaster* using crossover interference theory. *Heredity* 130:196–208. Nature
806 Publishing Group.

807 Lachaise, D., and J.-F. Silvain. 2004. How two Afrotropical endemics made two cosmopolitan
808 human commensals: the *Drosophila melanogaster*-*D. simulans* palaeogeographic riddle. Pp.
809 17–39 in P. Cappy, P. Gibert, and I. Boussy, eds. *Drosophila melanogaster, Drosophila simulans:*
810 *So Similar, So Different*. Springer Netherlands, Dordrecht.

811 Lanno, S. M., and J. D. Coolon. 2019. Derived esterase activity in *Drosophila sechellia*
812 contributes to evolved octanoic acid resistance. *Insect Molecular Biology* 28:798–806.

813 Lanno, S. M., S. M. Gregory, S. J. Shimshak, M. K. Alverson, K. Chiu, A. L. Feil, M. G.
814 Findley, T. E. Forman, J. T. Gordon, J. Ho, J. L. Krupp, I. Lam, J. Lane, S. C. Linde, A. E.
815 Morse, S. Rusk, R. Ryan, A. Saniee, R. B. Sheth, J. J. Siranosian, L. Sirichantaropart, S. R.

816 Sternlieb, C. M. Zaccardi, and J. D. Coolon. 2017. Transcriptomic Analysis of Octanoic Acid
817 Response in *Drosophila sechellia* Using RNA-Sequencing. *G3: Genes, Genomes, Genetics*
818 7:3867–3873.

819 Lanno, S. M., S. J. Shimshak, R. D. Peyser, S. C. Linde, and J. D. Coolon. 2019. Investigating
820 the role of Osiris genes in *Drosophila sechellia* larval resistance to a host plant toxin. *Ecology*
821 and Evolution 9:1922–1933.

822 Lavista-Llanos, S., A. Svatoš, M. Kai, T. Riemensperger, S. Birman, M. C. Stensmyr, and B.
823 S. Hansson. 2014. Dopamine drives *Drosophila sechellia* adaptation to its toxic host. *eLife*
824 3:e03785.

825 Legrand, D., M. I. Tenaillon, P. Matyot, J. Gerlach, D. Lachaise, and M.-L. Cariou. 2009.
826 Species-Wide Genetic Variation and Demographic History of *Drosophila sechellia*, a Species
827 Lacking Population Structure. *Genetics* 182:1197–1206.

828 Lemeunier, F., and M. Ashburner. 1976. Relationships within the melanogaster species
829 subgroup of the genus *Drosophila* (Sophophora) - II. Phylogenetic relationships between six
830 species based upon polytene chromosome banding sequences. *Proceedings of the Royal Society*
831 of London. Series B. Biological Sciences 193:275–294. Royal Society.

832 Lemeunier, F., B. Dutrillaux, and M. Ashburner. 1978. Relationships within the Melanogaster
833 subgroup species of the genus *Drosophila* (Sophophora). III. The mitotic chromosomes and
834 quinacrine fluorescent patterns of the polytene chromosomes. *Chromosoma* 69:349–361.

835 Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–
836 3100.

837 Li, H. 2023. Protein-to-genome alignment with miniprot. *Bioinformatics* 39:btad014.

838 Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and
839 R. Durbin. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*
840 25:2078–2079.

841 Llopart, A., D. Lachaise, and J. A. Coyne. 2005. An Anomalous Hybrid Zone in *Drosophila*.
 842 *Evolution* 59:2602–2607.

843 Louis, J., and J. R. David. 1986. Ecological specialization in the *Drosophila melanogaster*
 844 species subgroup: a case study of *D. sechellia*. *Acta oecologica. Oecologia generalis* 7:215–
 845 229.

846 Lu, H., F. Giordano, and Z. Ning. 2016. Oxford Nanopore MinION Sequencing and Genome
 847 Assembly. *Genomics, Proteomics & Bioinformatics* 14:265–279.

848 Lucek, K., Z. Gompert, and P. Nosil. 2019. The role of structural genomic variants in population
 849 differentiation and ecotype formation in *Timema cristinae* walking sticks. *Molecular Ecology*
 850 28:1224–1237.

851 Marchi, N., A. Kapopoulou, and L. Excoffier. 2024. Demogenomic inference from spatially
 852 and temporally heterogeneous samples. *Molecular Ecology Resources* 24:e13877.

853 Matschiner, M., J. M. I. Barth, O. K. Tørresen, B. Star, H. T. Baalsrud, M. S. O. Briec, C.
 854 Pampoulie, I. Bradbury, K. S. Jakobsen, and S. Jentoft. 2022. Supergene origin and
 855 maintenance in Atlantic cod. *Nat Ecol Evol* 6:469–481. Nature Publishing Group.

856 Matsuo, T., S. Sugaya, J. Yasukawa, T. Aigaki, and Y. Fuyama. 2007. Odorant-Binding
 857 Proteins OBP57d and OBP57e Affect Taste Perception and Host-Plant Preference in
 858 *Drosophila sechellia*. *PLOS Biology* 5:e118.

859 McDonough-Goldstein, C. E., S. Pitnick, and S. Dorus. 2022. *Drosophila* female reproductive
 860 glands contribute to mating plug composition and the timing of sperm ejection. *Proceedings of*
 861 *the Royal Society B: Biological Sciences* 289:20212213. Royal Society.

862 Mérot, C., E. L. Berdan, C. Babin, E. Normandeau, M. Wellenreuther, and L. Bernatchez. 2018.
 863 Intercontinental karyotype–environment parallelism supports a role for a chromosomal
 864 inversion in local adaptation in a seaweed fly. *Proceedings of the Royal Society B: Biological*
 865 *Sciences* 285:20180519. Royal Society.

866 Mérot, C., E. L. Berdan, H. Cayuela, H. Djambazian, A.-L. Ferchaud, M. Laporte, E.
 867 Normandeau, J. Ragoussis, M. Wellenreuther, and L. Bernatchez. 2021. Locally Adaptive
 868 Inversions Modulate Genetic Variation at Different Geographic Scales in a Seaweed Fly.
 869 *Molecular Biology and Evolution* 38:3953–3971.

870 Mérot, C., R. A. Oomen, A. Tigano, and M. Wellenreuther. 2020. A Roadmap for
 871 Understanding the Evolutionary Significance of Structural Genomic Variation. *Trends in*
 872 *Ecology & Evolution* 35:561–572.

873 Michel, A. P., S. Sim, T. H. Q. Powell, M. S. Taylor, P. Nosil, and J. L. Feder. 2010. Widespread
 874 genomic divergence during sympatric speciation. *Proceedings of the National Academy of*
 875 *Sciences* 107:9724–9729. *Proceedings of the National Academy of Sciences*.

876 Miller, D. E., C. Staber, J. Zeitlinger, and R. S. Hawley. 2018. Highly Contiguous Genome
 877 Assemblies of 15 *Drosophila* Species Generated Using Nanopore Sequencing. *G3*
 878 *Genes|Genomes|Genetics* 8:3131–3141.

879 Morales, H. E., R. Faria, K. Johannesson, T. Larsson, M. Panova, A. M. Westram, and R. K.
 880 Butlin. 2019. Genomic architecture of parallel ecological divergence: Beyond a single
 881 environmental contrast. *Science Advances* 5:eaav9963. American Association for the
 882 Advancement of Science.

883 Navarro, A., A. Barbadilla, and A. Ruiz. 2000. Effect of Inversion Polymorphism on the Neutral
 884 Nucleotide Variability of Linked Chromosomal Regions in *Drosophila*. *Genetics* 155:685–698.

885 Patlar, B., and A. Civetta. 2022. Seminal fluid gene expression and reproductive fitness in
 886 *Drosophila melanogaster*. *BMC Ecology and Evolution* 22:20.

887 Pease, J. B., and M. W. Hahn. 2013. More accurate phylogenies inferred from low-
 888 recombination regions in the presence of incomplete lineage sorting. *Evolution* 67:2376–2384.

889 Pettie, N., A. Llopart, and J. M. Comeron. 2022. Meiotic, genomic and evolutionary properties
 890 of crossover distribution in *Drosophila yakuba*. *PLOS Genetics* 18:e1010087. Public Library

891 of Science.

892 Peyser, R. D., S. M. Lanno, S. J. Shimshak, and J. D. Coolon. 2017. Analysis of cytochrome
893 P450 contribution to evolved plant toxin resistance in *Drosophila sechellia*. *Insect Molecular*
894 *Biology* 26:715–720.

895 Prieto-Godino, L. L., R. Rytz, S. Cruchet, B. Bargeton, L. Abuin, A. F. Silbering, V. Ruta, M.
896 D. Peraro, and R. Benton. 2017. Evolution of Acid-Sensing Olfactory Circuits in *Drosophilids*.
897 *Neuron* 93:661-676.e6.

898 Prigent, S. R., P. Le Gall, S. W. Mbunda, and M. Veuille. 2013. Seasonal and altitudinal
899 structure of drosophilid communities on Mt Oku (Cameroon volcanic line). *Comptes Rendus*
900 *Geoscience* 345:316–326.

901 Quilodrán, C. S., K. Ruegg, A. T. Sendell-Price, E. C. Anderson, T. Coulson, and S. M. Clegg.
902 2020. The multiple population genetic and demographic routes to islands of genomic
903 divergence. *Methods in Ecology and Evolution* 11:6–21.

904 R Core Team. 2023. R: A Language and Environment for Statistical Computing. R Foundation
905 for Statistical Computing, Vienna, Austria.

906 Rio, B., G. Couturier, F. Lemeunier, and D. Lachaise. 1983. Evolution d’une spécialisation
907 saisonnière chez *Drosophila erecta* (Dipt., Drosophilidae). *Ann. Soc. entomol. Fr* 19:235–248.

908 R’Kha, S., P. Capy, and J. R. David. 1991. Host-plant specialization in the *Drosophila*
909 *melanogaster* species complex: a physiological, behavioral, and genetical analysis. *PNAS*
910 88:1835–1839.

911 Rogers, R. L., J. M. Cridland, L. Shao, T. T. Hu, P. Andolfatto, and K. R. Thornton. 2014.
912 Landscape of Standing Variation for Tandem Duplications in *Drosophila yakuba* and
913 *Drosophila simulans*. *Mol Biol Evol* 31:1750–1766.

914 Roy, V., L. Monti-Dedieu, N. Chaminade, S. Siljak-Yakovlev, S. Aulard, F. Lemeunier, and C.
915 Montchamp-Moreau. 2005. Evolution of the chromosomal location of rDNA genes in two

916 *Drosophila* species subgroups: *ananassae* and *melanogaster*. *Heredity* 94:388–395. Nature
917 Publishing Group.

918 Sanchez-Donoso, I., S. Ravagni, J. D. Rodríguez-Teijeiro, M. J. Christmas, Y. Huang, A.
919 Maldonado-Linares, M. Puigcerver, I. Jiménez-Blasco, P. Andrade, D. Gonçalves, G. Friis, I.
920 Roig, M. T. Webster, J. A. Leonard, and C. Vilà. 2022. Massive genome inversion drives
921 coexistence of divergent morphs in common quails. *Current Biology* 32:462-469.e6.

922 Schaal, S. M., B. C. Haller, and K. E. Lotterhos. 2022. Inversion invasions: when the genetic
923 basis of local adaptation is concentrated within inversions in the face of gene flow.
924 *Philosophical Transactions of the Royal Society B: Biological Sciences* 377:20210200. Royal
925 Society.

926 Schlamp, F., J. van der Made, R. Stambler, L. Chesebrough, A. R. Boyko, and P. W. Messer.
927 2016. Evaluating the performance of selection scans to detect selective sweeps in domestic
928 dogs. *Molecular Ecology* 25:342–356.

929 Schrider, D. R., J. Ayroles, D. R. Matute, and A. D. Kern. 2018. Supervised machine learning
930 reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLOS*
931 *Genetics* 14:e1007341. Public Library of Science.

932 Sherman, B. T., M. Hao, J. Qiu, X. Jiao, M. W. Baseler, H. C. Lane, T. Imamichi, and W.
933 Chang. 2022. DAVID: a web server for functional enrichment analysis and functional
934 annotation of gene lists (2021 update). *Nucleic Acids Research* 50:W216–W221.

935 Siddiq, M. A., and J. W. Thornton. 2019. Fitness effects but no temperature-mediated balancing
936 selection at the polymorphic *Adh* gene of *Drosophila melanogaster*. *Proceedings of the National*
937 *Academy of Sciences* 116:21634–21640. *Proceedings of the National Academy of Sciences*.

938 Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. 2015.
939 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.
940 *Bioinformatics* 31:3210–3212.

941 Sodeland, M., P. E. Jorde, S. Lien, S. Jentoft, P. R. Berg, H. Grove, M. P. Kent, M. Arnyasi, E.
 942 M. Olsen, and H. Knutsen. 2016. “Islands of Divergence” in the Atlantic Cod Genome
 943 Represent Polymorphic Chromosomal Rearrangements. *Genome Biology and Evolution*
 944 8:1012–1022.

945 Tepolt, C. K., E. D. Grosholz, C. E. de Rivera, and G. M. Ruiz. 2022. Balanced polymorphism
 946 fuels rapid selection in an invasive crab despite high gene flow and low genetic diversity.
 947 *Molecular Ecology* 31:55–69.

948 Thurmond, J., J. L. Goodman, V. B. Strelets, H. Attrill, L. S. Gramates, S. J. Marygold, B. B.
 949 Matthews, G. Millburn, G. Antonazzo, V. Trovisco, T. C. Kaufman, B. R. Calvi, N. Perrimon,
 950 S. R. Gelbart, J. Agapite, K. Broll, L. Crosby, G. dos Santos, D. Emmert, L. S. Gramates, K.
 951 Falls, V. Jenkins, B. Matthews, C. Sutherland, C. Tabone, P. Zhou, M. Zytkevich, N. Brown,
 952 G. Antonazzo, H. Attrill, P. Garapati, A. Holmes, A. Larkin, S. Marygold, G. Millburn, C.
 953 Pilgrim, V. Trovisco, P. Urbano, T. Kaufman, B. Calvi, B. Czoch, J. Goodman, V. Strelets, J.
 954 Thurmond, R. Cripps, and P. Baker. 2019. FlyBase 2.0: the next generation. *Nucleic Acids Res*
 955 47:D759–D765.

956 Tricou, T., E. Tannier, and D. M. de Vienne. 2022. Ghost Lineages Highly Influence the
 957 Interpretation of Introgression Tests. *Systematic Biology* 71:1147–1158.

958 Turissini, D. A., and D. R. Matute. 2017. Fine scale mapping of genomic introgressions within
 959 the *Drosophila yakuba* clade. *PLOS Genetics* 13:e1006971.

960 Turner, T. L., M. W. Hahn, and S. V. Nuzhdin. 2005. Genomic Islands of Speciation in
 961 *Anopheles gambiae*. *PLOS Biology* 3:e285.

962 Verity, R., and R. A. Nichols. 2016. Estimating the Number of Subpopulations (K) in Structured
 963 Populations. *Genetics* 203:1827–1839.

964 Vertacnik, K. L., and C. R. Linnen. 2017. Evolutionary genetics of host shifts in herbivorous
 965 insects: insights from the age of genomics. *Ann. N.Y. Acad. Sci.* 1389:186–212.

966 Vijay, N., M. Weissensteiner, R. Burri, T. Kawakami, H. Ellegren, and J. B. W. Wolf. 2017.
 967 Genomewide patterns of variation in genetic diversity are shared among populations, species
 968 and higher-order taxa. *Molecular Ecology* 26:4284–4295.

969 Voudibio, J. 1985. *Ecologie des populations et biologie évolutive des drosophiles en Afrique*
 970 *équatoriale (Congo)*. Université Paris XII Val-de-Marne, France.

971 White, B. J., C. Cheng, F. Simard, C. Costantini, and N. J. Besansky. 2010. Genetic association
 972 of physically unlinked islands of genomic divergence in incipient species of *Anopheles*
 973 *gambiae*. *Molecular Ecology* 19:925–939.

974 Wiens, J. J., R. T. Lapoint, and N. K. Whiteman. 2015. Herbivory increases diversification
 975 across insect clades. *Nature Communications* 6:8370.

976 Wolf, J. B. W., and H. Ellegren. 2017. Making sense of genomic islands of differentiation in
 977 light of speciation. *Nat Rev Genet* 18:87–100.

978 Wong, A., M. C. Turchin, M. F. Wolfner, and C. F. Aquadro. 2012. Temporally Variable
 979 Selection on Proteolysis-Related Reproductive Tract Proteins in *Drosophila*. *Molecular*
 980 *Biology and Evolution* 29:229–238.

981 Yassin, A. 2017. *Drosophila yakuba mayottensis*, a new model for the study of incipient
 982 ecological speciation. *Fly* 11:37–45.

983 Yassin, A., V. Debat, H. Bastide, N. Gidaszewski, J. R. David, and J. E. Pool. 2016. Recurrent
 984 specialization on a toxic fruit in an island *Drosophila* population. *PNAS* 113:4771–4776.

985 Yi, X., Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, J. E. Pool, X. Xu, H. Jiang, N.
 986 Vinckenbosch, T. S. Korneliussen, H. Zheng, T. Liu, W. He, K. Li, R. Luo, X. Nie, H. Wu, M.
 987 Zhao, H. Cao, J. Zou, Y. Shan, S. Li, Q. Yang, Asan, P. Ni, G. Tian, J. Xu, X. Liu, T. Jiang, R.
 988 Wu, G. Zhou, M. Tang, J. Qin, T. Wang, S. Feng, G. Li, Huasang, J. Luosang, W. Wang, F.
 989 Chen, Y. Wang, X. Zheng, Z. Li, Z. Bianba, G. Yang, X. Wang, S. Tang, G. Gao, Y. Chen, Z.
 990 Luo, L. Gusang, Z. Cao, Q. Zhang, W. Ouyang, X. Ren, H. Liang, H. Zheng, Y. Huang, J. Li,

991 L. Bolund, K. Kristiansen, Y. Li, Y. Zhang, X. Zhang, R. Li, S. Li, H. Yang, R. Nielsen, J.
992 Wang, and J. Wang. 2010. Sequencing of 50 Human Exomes Reveals Adaptation to High
993 Altitude. *Science* 329:75–78.
994 Zimin, A. V., G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg, and J. A. Yorke. 2013. The
995 MaSuRCA genome assembler. *Bioinformatics* 29:2669–2677.
996

997 **Table 1** – Position, structural features and candidate genes or families of the 23 genomic islands
 998 of divergence (GIDs) distinguishing noni-specialist *Drosophila yakuba mayottensis* from
 999 generalist mainland species. Correspondence with recombination rate (R), inversions (I) and
 1000 selective sweeps (H) are given. For R, low (L) and high (H) rates were obtained from Pettie et
 1001 al. (2022) who used collinear parents. For H, whether a signal was detected (Y) or not detected
 1002 (N) from the H-scan analysis is given. *P* levels from DAVID gene ontology enrichment
 1003 analyses for each GID.

GID	Coordinates (Mb)	R	I	H	Candidate genes
1	X:3.63-4.30	L		Y	
2	X:8.68-9.30	L	<i>Xtb</i>	Y	<i>Flo2</i> ; Glucose-methanol-choline oxidoreductase (ecdysteroid metabolic process, $P < 7 \times 10^{-22}$)
3	X:21.22-21.77	L		Y	
4	2L:2.21-2.27	H		Y	Serine-type endopeptidase activity ($P < 1 \times 10^{-3}$), endonuclease ($P < 1 \times 10^{-3}$)
5	2L:14.51-14.83	H	<i>2Lo</i>	Y	<i>Adh</i>
6	2L:16.20-17.15	H	<i>2Lo</i>	Y	<i>pdm3</i> ; maltose alpha-glucosidase activity ($P < 2 \times 10^{-10}$), sugar metabolism and transport, $P = 0.022$)
7	2L:18.23-18.25	L	<i>2Lo</i>	Y	
8	2L:21.11-21.57	L	<i>2Lo</i>	Y	
9	2R:2.09-2.23	L		Y	
10	2R:4.62-5.57	L		N	<i>Catsup</i> ; UDP-glycosyltransferases (metabolism of xenobiotics by cytochrome P450, $P < 0.05$)
11	2R:6.72-6.93	H	<i>2Rj</i>	Y	
12	2R:10.43-10.76	H	<i>2Rj</i> , <i>2Rk</i>	Y	Cytochrome P450 ($P < 1 \times 10^{-5}$)
13	2R:14.79-17.57	H	<i>2Rj</i> , <i>2Rk</i>	Y	LITAF ($P < 2 \times 10^{-15}$); glutathione S-transferase ($P < 0.05$)
14	2R:19.54-19.63	H	<i>2Rj</i>	Y	
15	2R:21.11-21.13	L		Y	
16	3L:21.76-22.40	L		Y	
17	3L:23.47-24.20	L	<i>3Lk</i>	Y	

18	3R:0.00-0.33	L	<i>3Ri</i>	Y	
19	3R:3.31-4.71	L		Y	
20	3R:6.81-6.83	L		Y	
21	3R:10.82-12.12	L		Y	UDP-glucuronosyl/UDP-glucosyltransferase ($P < 5 \times 10^{-7}$)
22	3R:13.38-13.42	L		Y	<i>Ace, Osi22</i>
23	3R:17.96-17.98	H		Y	<i>Osi2, Osi3, Osi4, Osi5</i> (Proteins of unknown function DUF1676, $P < 1 \times 10^{-8}$)

1004

1005

1006

1007

Table 2 – Inversions delimitations according to the intergenic regions on the reference genome. Approximate positions of the left and right breakpoints of each inversion relative to the reference genome are given in terms of intergenic regions (e.g., A/B and C/D genes in reference genome will lead to A/C and B/D configuration in inversion). Gene names and their corresponding cytological positions are given according to *D. melanogaster* orthologs. Inversion names, breakpoints and geographical distribution into West Africa (W), East Africa (E) and Mayotte (M) are given following Lemeunier and Ashburner (1976) and this study.

Inversion name	Coordinates (length)	Left breakpoint (A/B)	Right breakpoint (C/D)	Geographic distribution	GIDs
<i>Xtb</i>	X :8,648,348-9,652,060 (1.0 Mb)	<i>CG12540</i> (12E11)/ <i>CG14414</i> (12F1)	<i>Ac13E</i> (13C7)/ <i>CG7860</i> (13D3)	M	2
<i>2Lo</i>	2L:8,423,810-20,352,247 (11.9 Mb)	<i>Hacd1</i> (33B5)/ <i>Rh5</i> (33B8)	<i>bin3</i> (42A12)/ <i>tombay40</i> (42A14)	M	5, 6, 7, 8
<i>2Ro</i>	2R:182,816-436,892 (0.3 Mb)	<i>CG3262</i> (40F6)/ <i>Ir40a</i> (40F4)	<i>CG3635</i> (40F2)/ <i>Clamp</i> (40F2)	M	
<i>2Rj</i>	2R:8,717,812-18,133,566 (9.4 Mb)	<i>CG30047</i> (49A1)/ <i>ana3</i> (48F11)	<i>mth</i> (61C1)/ <i>EDTP</i> (54B7)	W, E, M	11, 12, 13, 14
<i>2Ri</i>	2R:11,157,330-18,066,346 (6.9 Mb)	<i>Hr51</i> (51F7)/ <i>CG34205</i> (58C1)	<i>ppk12</i> (58E1)/ <i>CG4847</i> (54C1)		12, 13, 14
<i>2Rk</i>	2R:10,365,887-14,704,259 (4.3 Mb)	<i>CG30076</i> (51A8)/ <i>CG43919</i> (51B1)	<i>CG42697</i> (55E1)/ <i>GEFmeso</i> (55D3)	W, E	12, 13
<i>3Lk</i>	3L:23,737,761-23,929,648 (0.2 Mb)	<i>Ckl2alpha</i> (80D1)/ <i>nAChRalpha4</i> (80D5)	<i>ND-MLRQ</i> (80F2)/ <i>CG17454</i> (80F9)	M	17
<i>3Ri</i>	3R:1-818,622 (0.8 Mb)	<i>5-HT2A</i> (82C4)/ <i>CG9775</i> (82C5)	<i>Spartin</i> (82C5)/ <i>CG33927</i> (82D1)	M	18

Figure legends

Figure 1 – Association of *Drosophila yakuba mayottensis* with noni (*Morinda citrifolia*) on the island of Mayotte. A) Frequency of *D. yakuba* (blue) among drosophilids (red) collected on the island of Mayotte in January 2013 (after David et al. (2014)) and April 2017. The size of the pie charts reflects the total number of collected flies per site. Grey shades refer to vegetation density. B) Abundance of *D. yakuba* flies among drosophilids collected using standard fermenting banana traps in Congo (after Voudibio (1985)) and in Mayotte. C) A brown lemur (*Eulemur fulvus*) eating a noni fruit in Mayotte (© Amir Yassin).

Figure 2 – Population structure analysis within and between *D. yakuba* subspecies as analyzed by rMaverick supports lack of structure in *D. y. mayottensis* despite substantial admixture. A) Structure plots for different values of subpopulations (k) for five *D. yakuba* populations (CY = Cameroon, NY = Kenya, MTS = M'Tsangamouji, SOU = Soulou, BE = Bambo Est). Each bar represents one individual (from Ind-Seq data), with colors referring to the posterior probability of belonging to each of the k subpopulations. B) Full posterior distribution of k after transforming and normalizing log evidence values showing $k = 3$. C) Posterior probability (density) plot of the admixture parameter (α) from the MCMC run of the $k = 3$ model supports a substantial degree of admixture (~ 0.07).

Figure 3 – High genetic differentiation among the two *D. yakuba* subspecies from Pool-Seq data. A) Distribution of segregating sites (SNPs) in 10-kb windows showing the extent of telomeric and centromeric effects as well as the heterochromatic block on diversity at the mid of the X chromosome (gray areas). These areas correspond to regions with low recombination

rate (Pettie et al. 2022). Each chromosomal arm according to the *D. y. yakuba* v.1.05 reference genome is represented by different color. B) Population Branch Excess (*PBE*) identification of the genomic islands of divergence (GIDs), enumerated in red, distinguishing *D. y. mayottensis* from the two *D. y. yakuba* populations. Horizontal dashed line corresponds to the 0.975 quantile threshold.

Figure 4 – Genomic islands of divergence (GIDs) predominantly overlap with major chromosomal inversions. A) Dot plot comparing *D. y. mayottensis* hybrid assembly with the *D. y. yakuba* reference genome v1.05. Each *D. y. mayottensis* scaffold is given in different color with the position of the GIDs indicated by their number. Shaded grey and pink areas indicate low recombining regions and GID spans, respectively. Solid colored arrows indicate inversion breakpoints, blank arrows indicate the breakpoints of an inversion misassembly on 2L. B) Schematic presentation of major inversions on chromosomal arms 2L (left) and 2R (right). Each chromosomal arm is presented as a set of colored arrows with the head of the arrow pointing to the orientation of the corresponding segment in the *D. y. yakuba* reference genome (first row), except for the two misassembled segments of 2L (whose orientation is corrected in the second row). Inversion names, breakpoints and geographical distribution into West Africa (W), East Africa (E) and Mayotte (M) are given following Lemeunier and Ashburner (1976) and this study. GIDs names and positions on the reference genome are given in red on the first row. Crossing lines connect the breakpoints between the configurations. Note that the three 2L inversions all independently evolved from *2Ll*, whereas the 2R inversions had two independent origins from *2Rj*, with one origin leading to a hypothetical *2Ri* from which the reference *2Rn* was derived, and another origin leading to *2Rk* from which *2Rm* was derived. The grey shaded area surrounding the centromere (black circle) refer to the low recombining region.

1068 **Figure 5** – Evidence for selective sweeps in regions in normally-recombining regions falling
1069 more than 3 Mb outside inversion breakpoints. A) Distribution of H statistics, *i.e.* average
1070 length of homozygosity tracts for each SNP in kilobases (kb) as estimated by the H-scan
1071 program from unphased Ind-Seq data. Color code and horizontal dashed line are as in Figure
1072 3. Red vertical dashed lines refer to the position of the inversion breakpoints. Horizontal
1073 dashed line refer to the 97.5% quantile value. B) Distribution of H statistics in the only two
1074 GIDs falling in normally-recombining region, distant from chromosomal inversion
1075 breakpoints, GID4 and GID23, showing the likely targets of selective sweeps. Gene names
1076 refer to the *D. melanogaster* orthologs.

1077 **Supplementary figures legends**

1078

1079 **Figure S1** – Bimodal distribution of the number of segregating sites (SNPs) per 10-kb windows
1080 from Pool-Seq data.

1081

1082 **Figure S2** – Low temporal and spatial genetic differentiation among *D. y. mayottensis*
1083 populations on the island of Mayotte. A) Pairwise F_{ST} values for 10-kb windows between Pool-
1084 Seq data from Soulou collected in January 2013 and April 2017. B) Average pairwise F_{ST} values
1085 for 10-kb windows between Pool-Seq data from the three populations collected in Mayotte in
1086 April 2017. Gray areas reflect subtelomeric and subcentromeric regions with low recombination
1087 rate (Yassin et al. 2016; Pettie et al. 2022). Each chromosomal arm according to the v.1.05
1088 reference genome is represented by different color.

1089

1090 **Figure S3** – Population structure analysis within and between *D. yakuba* subspecies using a
1091 minor allele frequency of 5% rather than 25% given in Figure 2. A) Structure plots for different
1092 values of subpopulations (k) for five *D. yakuba* populations (CY = Cameroon, NY = Kenya,
1093 MTS = M'Tsangamouji, SOU = Soulou, BE = Bambo Est). Each bar represents one individual
1094 (from Ind-Seq data), with colors referring to the posterior probability of belonging to each of
1095 the k subpopulations. B) Full posterior distribution of k after transforming and normalizing log
1096 evidence values showing $k = 2$. C) Posterior probability (density) plot of the admixture
1097 parameter (α) from the MCMC run of the $k = 2$ model supports a substantial degree of
1098 admixture (~ 0.05).

1099

1100 **Figure S4** – High genetic differentiation among the two *D. yakuba* subspecies from Ind-Seq
1101 data as shown from the Population Branch Excess (*PBE*) analysis. Color code and horizontal
1102 dashed line are as in Figure 3.

1103

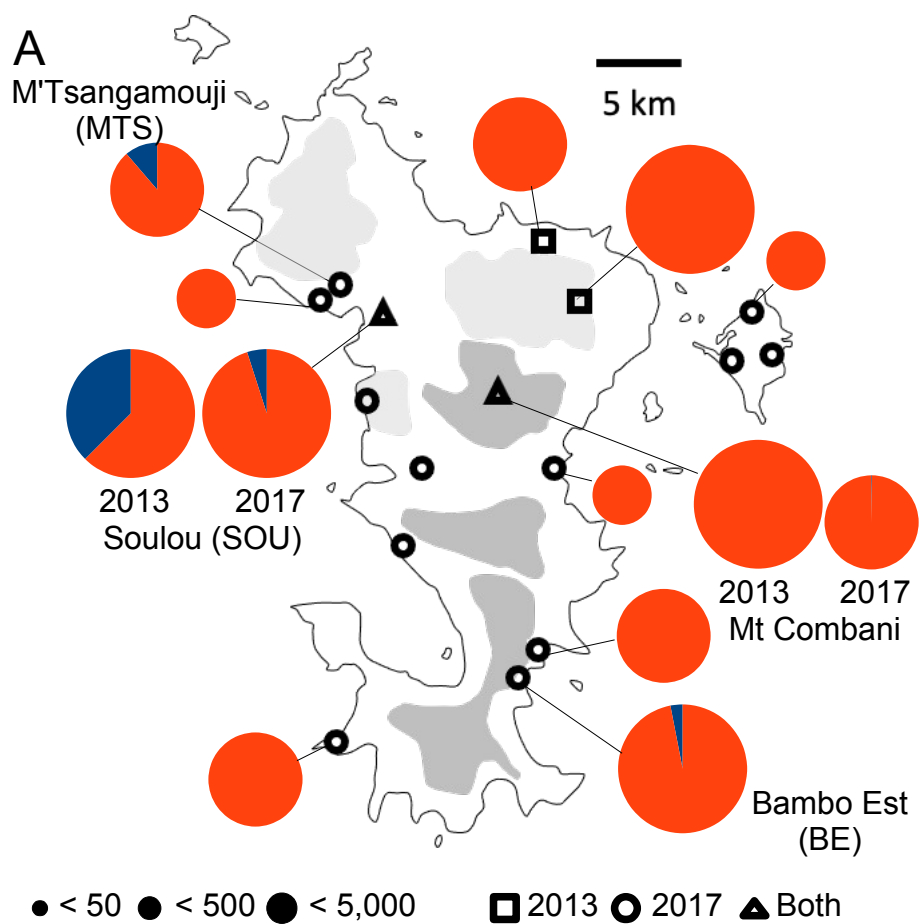
1104 **Figure S5** – Dot plots showing chromosomal inversions breakpoints for A) *Xtb*, B) *2Lo*, C)
1105 *2Ro*, D) *2Rj*, E) *3Lk*, and F) *3Ri*. Each dot corresponds to an orthologous gene. For each plot,
1106 the names of the scaffold of the *D. y. mayottensis* genome and of the chromosomal arm of the
1107 *D. y. yakuba* genome are given.

1108

1109 **Figure S6** – Dot plot comparing *D. y. yakuba* reference genome v2.1 with the *D. y. yakuba*
1110 reference genome v1.05. Each v2.1 chromosome/scaffold is given in different color.

1111

1112



B

Banana bait	Not yak	yak
Mayotte	210	0
Congo	6531	383

C



Figure 1

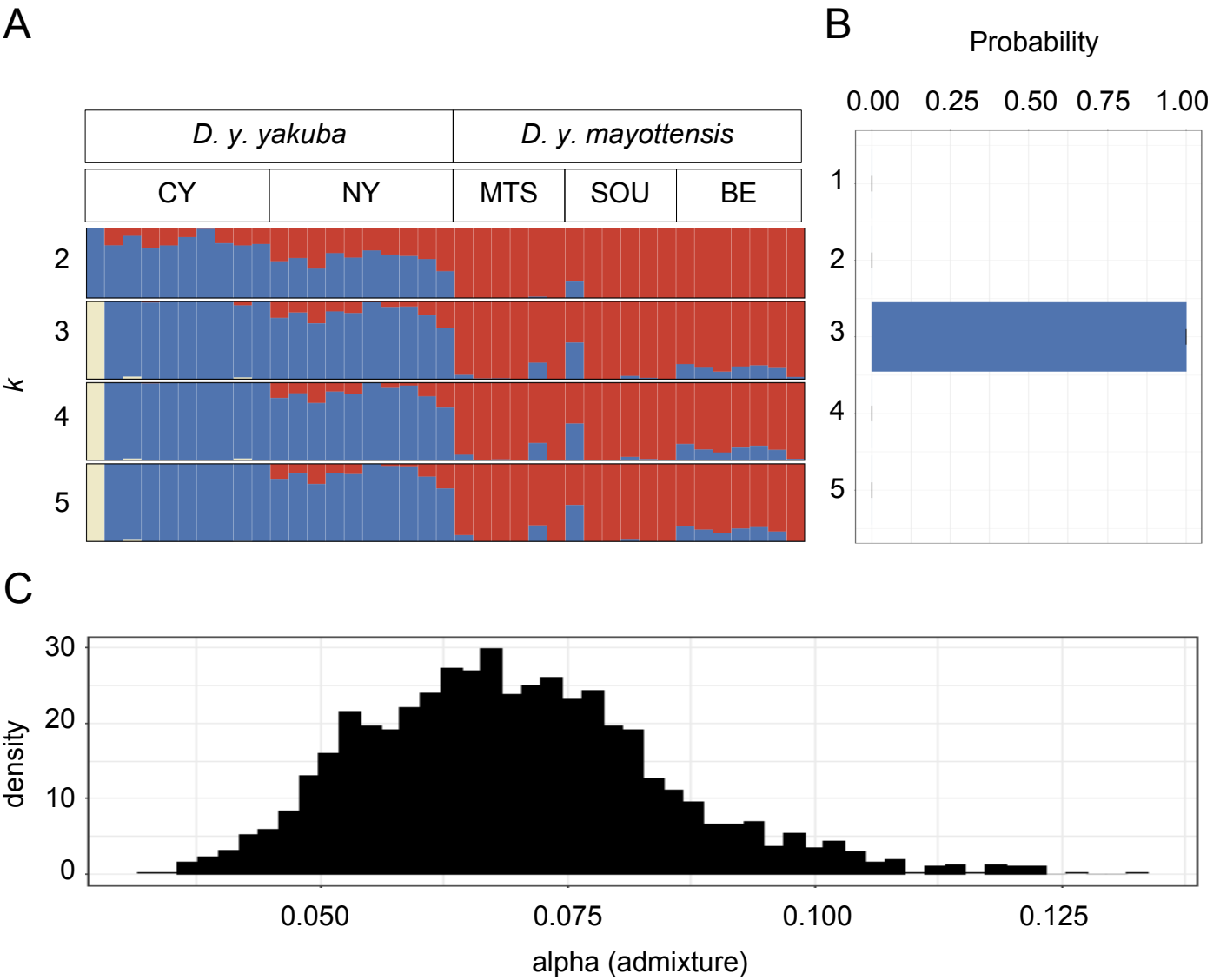
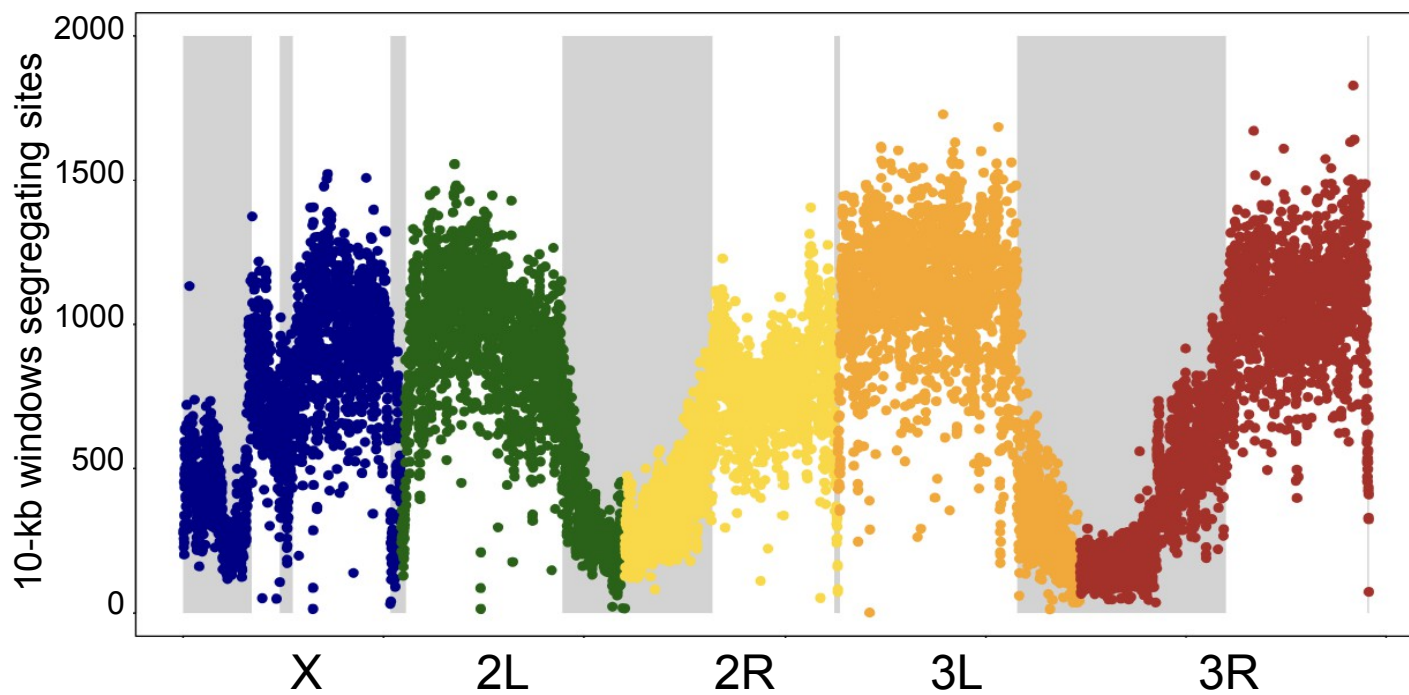


Figure 2

A



B

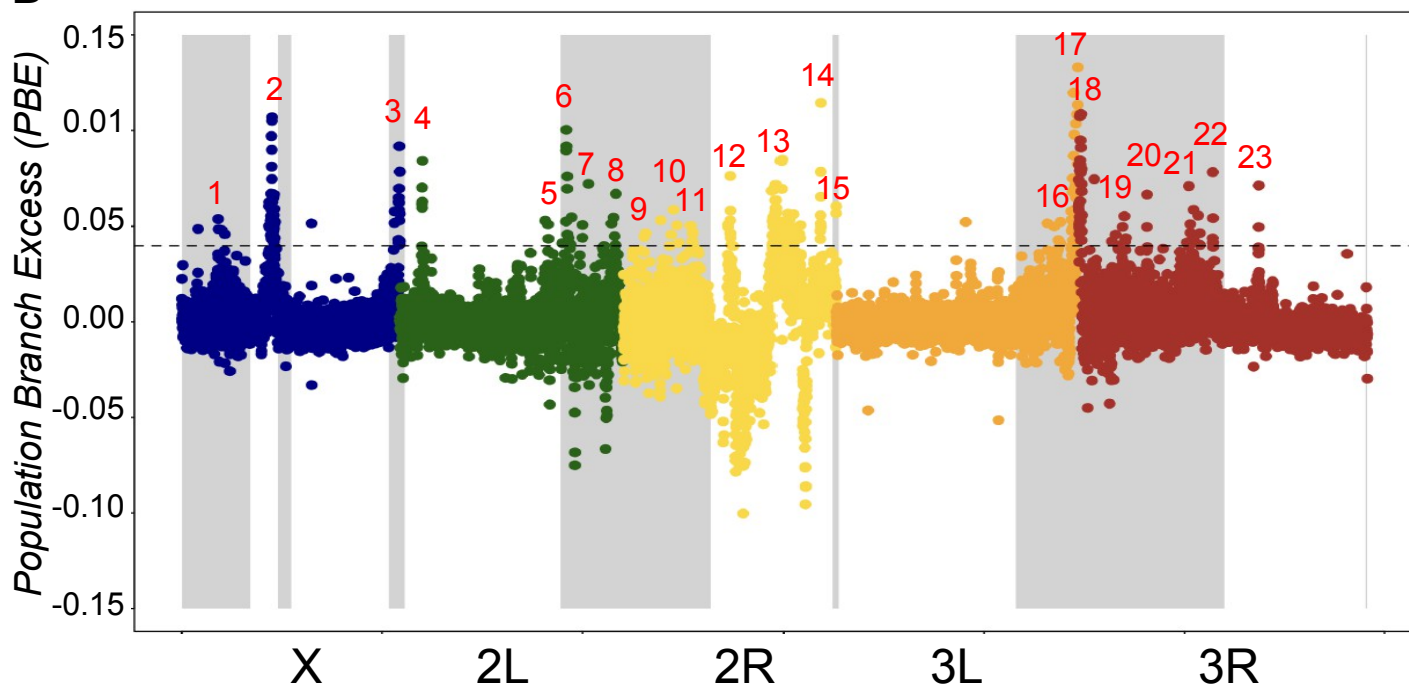


Figure 3

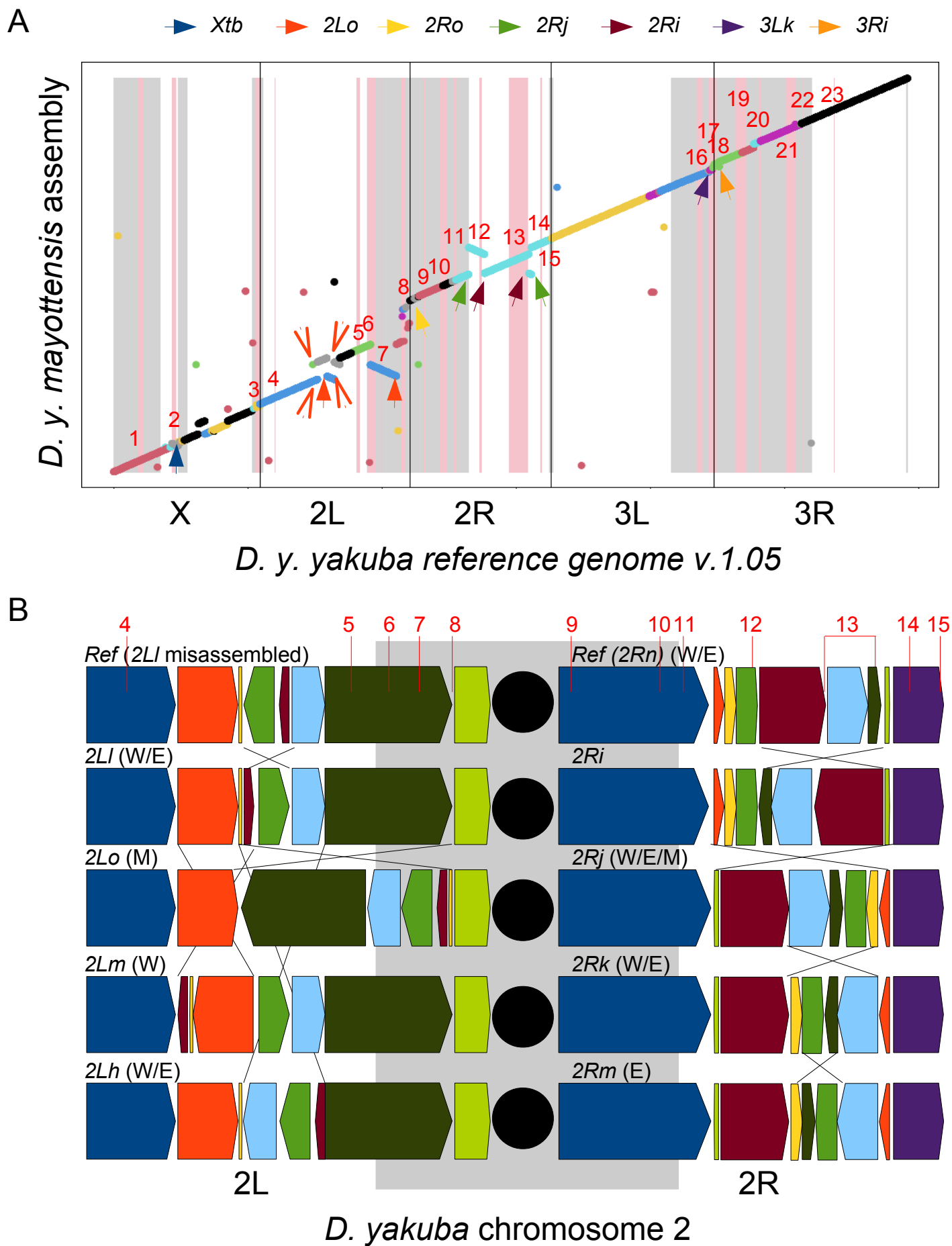


Figure 4

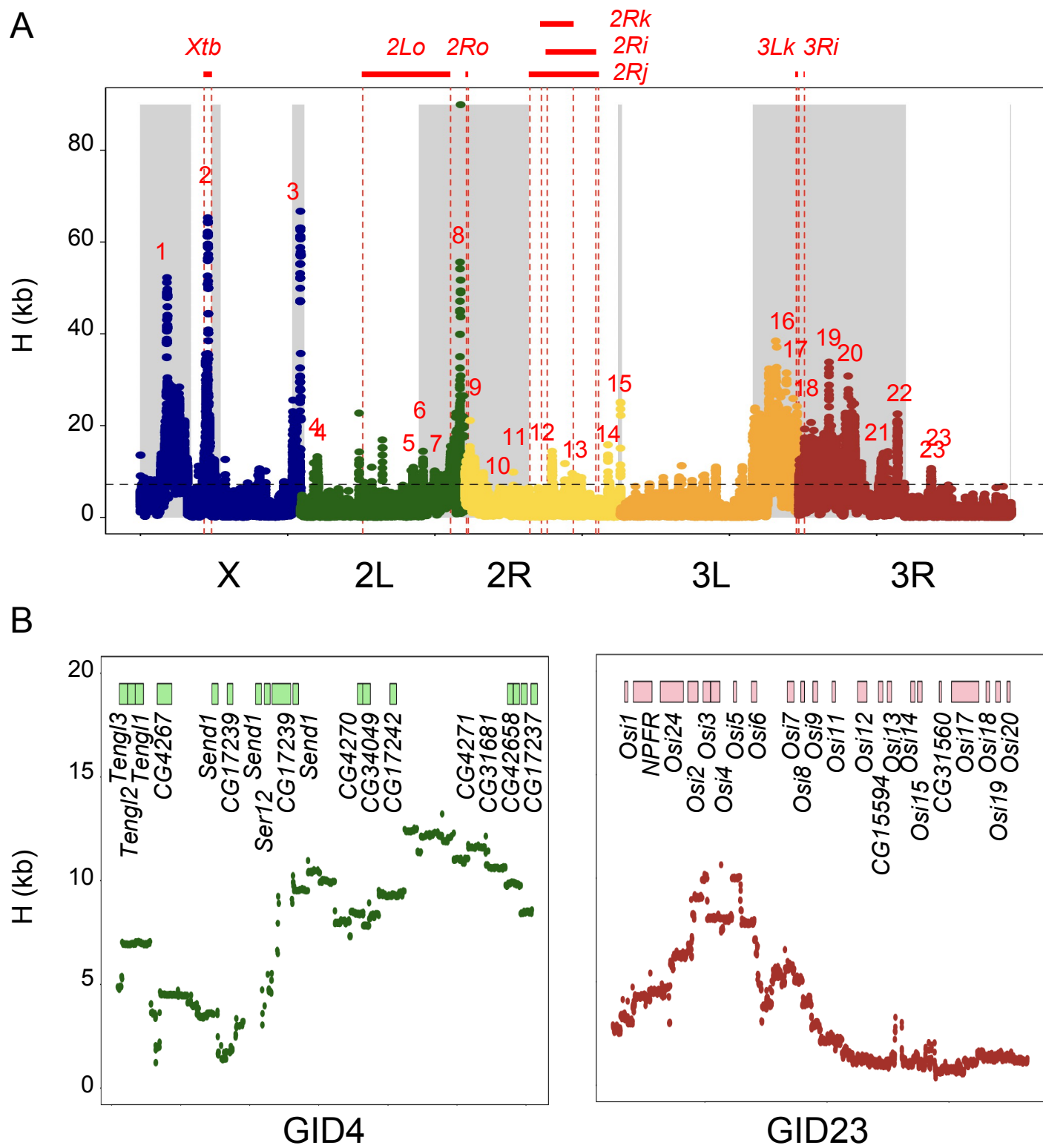


Figure 5

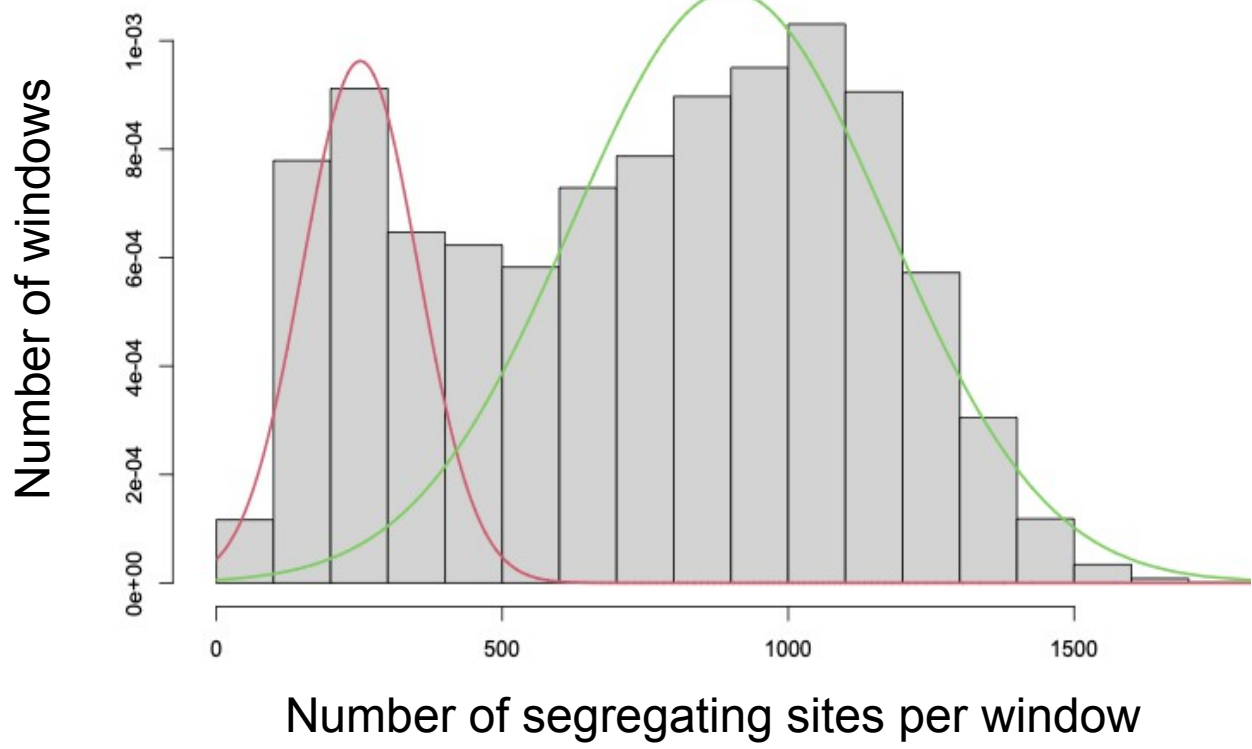


Figure S1

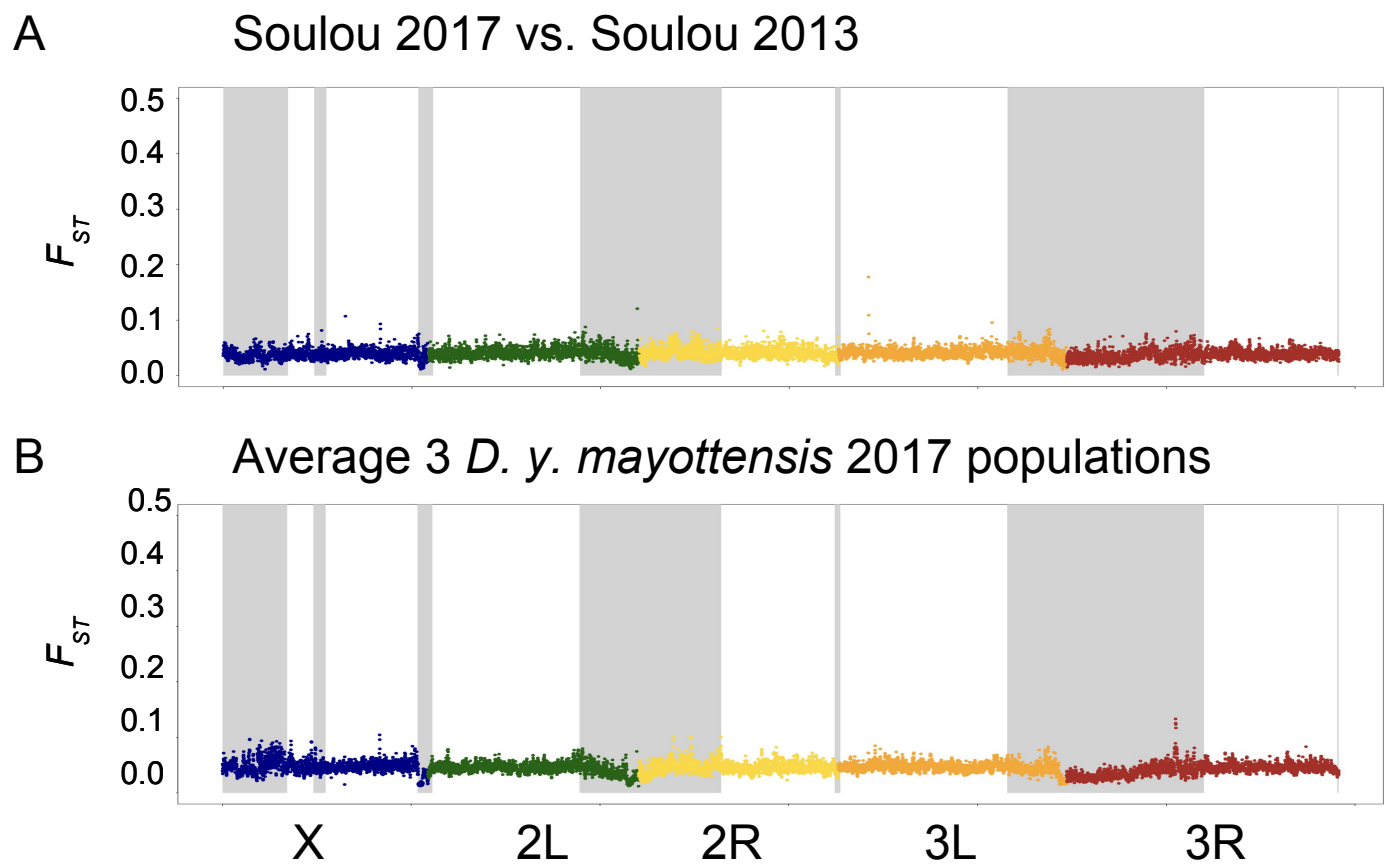
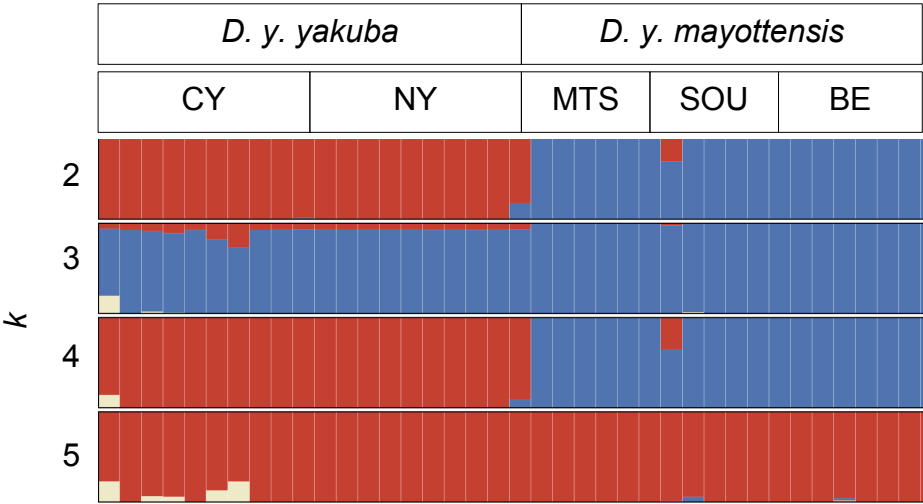
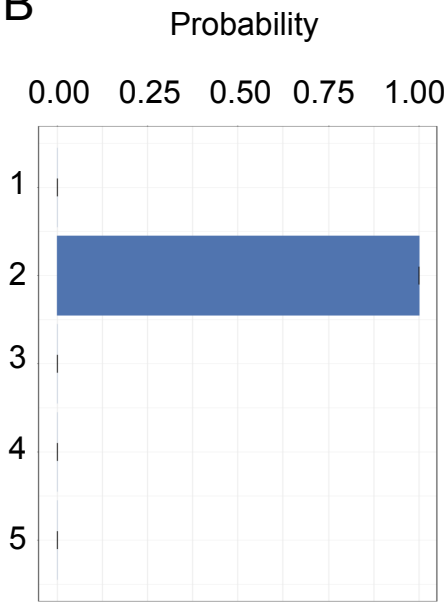


Figure S2

A



B



C

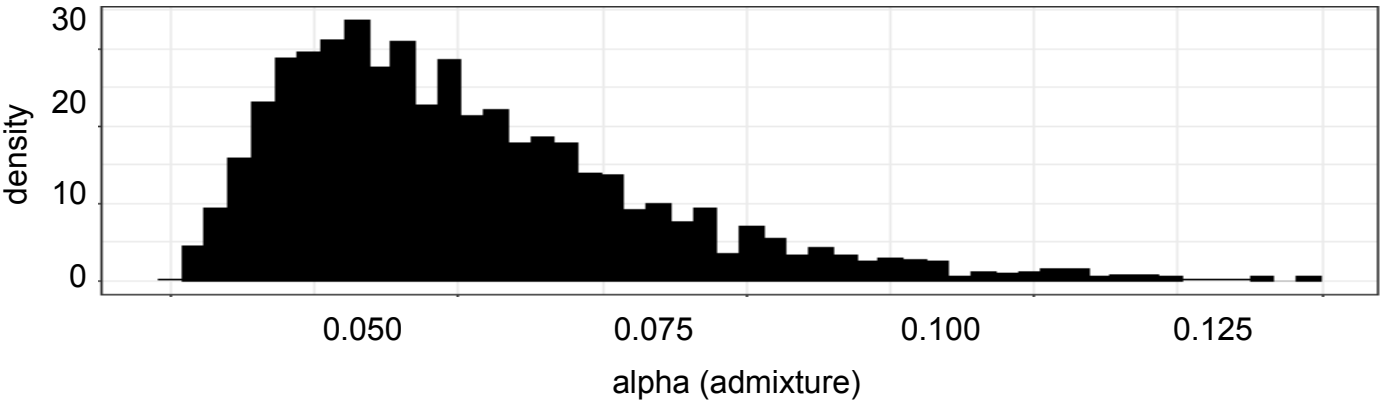


Figure S3

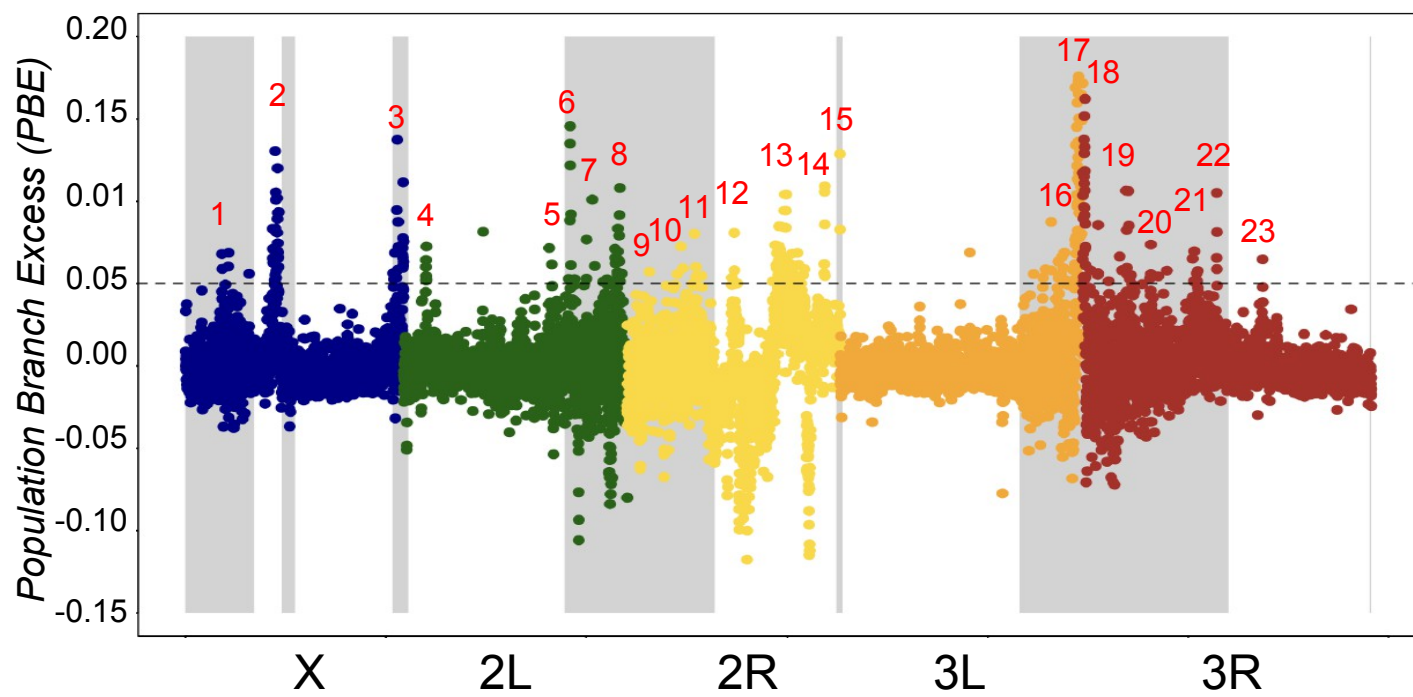
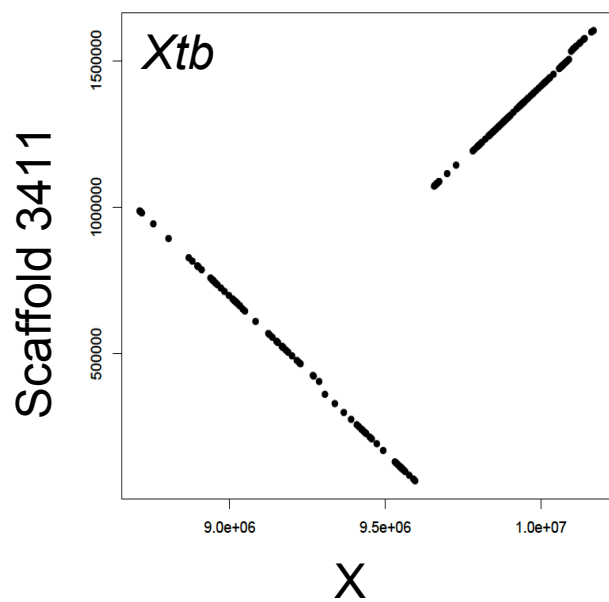
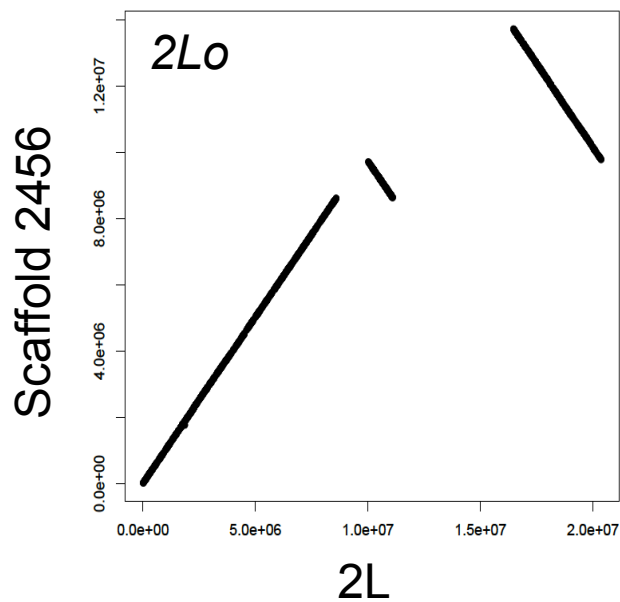


Figure S4

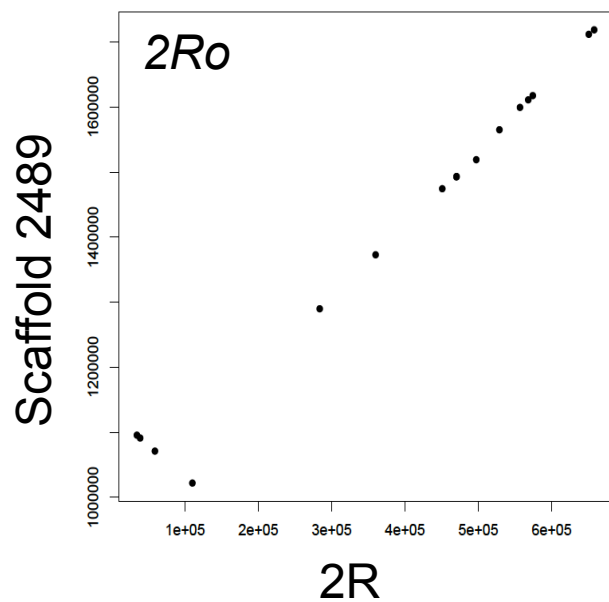
A



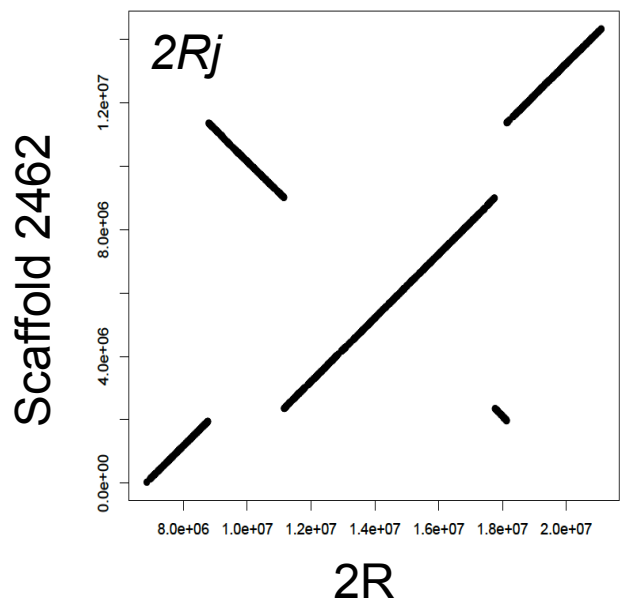
B



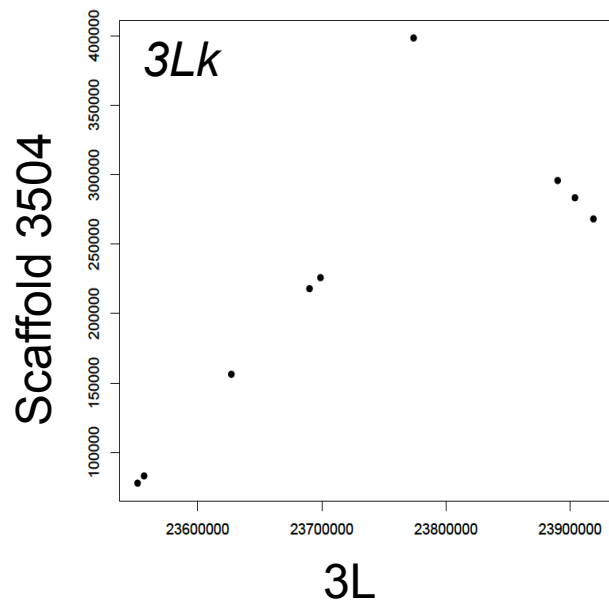
C



D



E



F

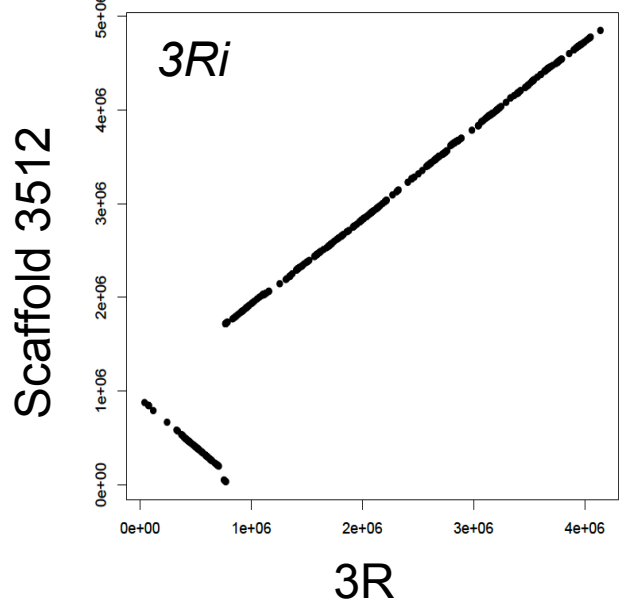


Figure S5

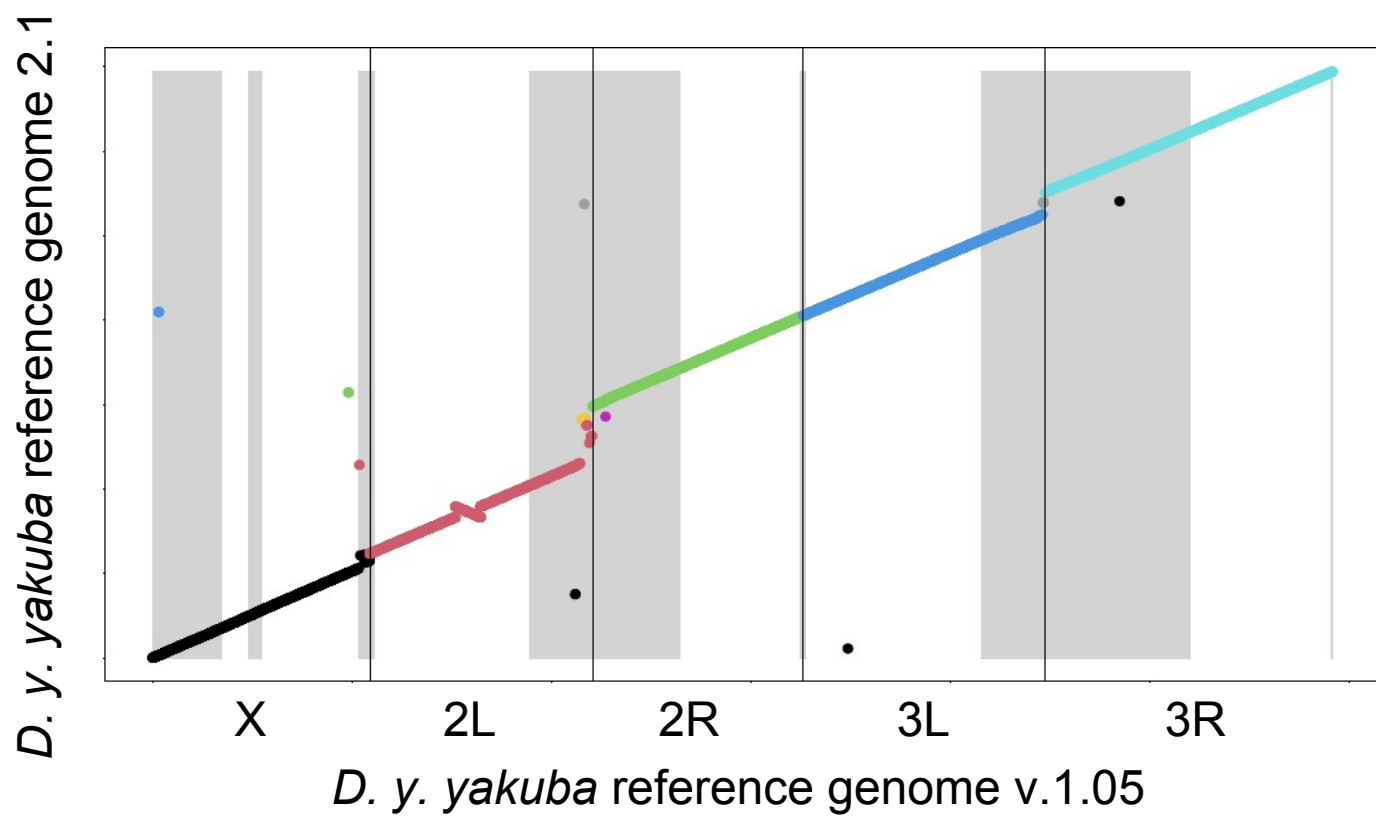


Figure S6