MOLECULAR ECOLOGY

Supplemental Information for:

Transposable elements and introgression introduce genetic variation in the invasive ant *Cardiocondyla obscurior*

Running title: Genome dynamics in *Cardiocondyla obscurior*

Key words: rapid adaptation, invasive species, introgression, transposable elements, population genomics, *Cardiocondyla obscurior*

Mohammed Errbii[1], Jens Keilwagen[2], Katharina J. Hoff[3,4], Raphael Steffen[1], Janine Altmüller[5], Jan Oettler[6], Lukas Schrader[1]

[1]Institute for Evolution and Biodiversity, University of Münster, 48149 Münster, Germany

[2]Julius Kühn Institute (JKI) – Federal Research Centre for Cultivated Plants, Institute for Biosafety in Plant Biotechnology, Erwin-Baur-Str. 27, 06484 Quedlinburg, Germany

[3]Institute of Mathematics and Computer Science, University of Greifswald, Walther-Rathenau-Str. 47, 17489 Greifswald, Germany

[4]Center for Functional Genomics of Microbes, University of Greifswald, Felix-Haussdorff-Str. 8, 17489 Greifswald, Germany

[5]Cologne Center for Genomics, Institute of Human Genetics, University of Cologne, 50931 Cologne, Germany

[6]Lehrstuhl für Zoologie/Evolutionsbiologie, University Regensburg, 93053 Regensburg, Germany

Correspondence: Lukas Schrader (lukas.schrader@uni-muenster.de) and Jan Oettler (joettler@gmail.com)

MOLECULAR ECOLOGY

Table of Contents:

Table S1: Quality control metrics for the single individual and pool data used in this study.	Page 3
Table S2: List of published ant genomes used to generate *de novo* repeat libraries with RepeatScout.	Page 4
Table S3: TE-related functional annotation terms used for identifying TE-encoded proteins.	Page 5
Table S4: QUAST based statistics calculated for the Cobs2.1 genome assembly.	Page 7
Table S5: Location and length of each of the 34 TE islands identified in the Cobs2.1 genome assembly.	Page 8
Table S6: Results of Gene Ontology Enrichment analyses with topGO of genes contained in TE islands in the Cobs2.1 genome assembly.	Page 9
Table S7: Most abundant repeat families identified in populations of *C. obscurior* from Tenerife and Itabuna, using dnaPipeTE with 0.1× coverage per samples.	Page 10
Table S8: Contingency table showing the number of low frequency, common and fixed TE insertions in TE islands as well as LDRs in the Itabuna and Tenerife population.	Page 11
Figure S1: IGV screenshot showing the relative TE content across the genome of *C. obscurior* recovered using different repeat annotation approaches.	Page 12
Figure S2: Close-up view of TE annotations across the TE island on scaffold 1.	Page 13
Figure S3: Mapping quality across the 127 scaffolds of the Cobs2.1 assembly.	Page 14
Figure S4: Insert size distribution in pool-seq data of Tenerife and Itabuna, estimated from the alignment files.	Page 15
Figure S5: Admixture patterns among *C. obscurior* populations by using ADMIXTURE at K = 2, K = 3, k = 4, k = 5 and k = 6.	Page 16
Figure S6: Population history in *C. obscurior* estimated using MSMC2 with eight phased haplotypes, representing the New and Old World lineages respectively.	Page 17
Figure S7: Principal Component Analysis based on 115,334 SNPs of samples belonging to the Old and New World lineages.	Page 18
Figure S8: Density plot showing the distribution Tajima's *D* values in *C. obscurior*, estimated in 100-kb non-overlapping windows for the Tenerife and Itabuna population.	Page 19
Figure S9: Genome-wide distribution of genetic diversity (π) of the 30 largest *C. obscurior* genome scaffolds.	Page 20
Figure S10: Genome-wide distribution of Tajima's *D* of the 30 largest *C. obscurior* genome scaffolds.	Page 21
Figure S11: Genome-wide distribution of genetic differentiation (Fst) of the 30 largest *C. obscurior* genome scaffolds.	Page 22
Figure S12: Genetic diversity and differentiation in introgressed regions compared to the remainder of LDRs.	Page 23
Figure S13: Genetic diversity and differentiation within LDRs and TE islands in each population after excluding introgressed regions.	Page 24
Figure S14: Association plots showing the signed contribution to Pearson's χ^2 (left panels) and the Pearson's χ^2 standardized residuals (right panels) calculated for low frequency (f < 0.25), common (0.25 ≤ f ≤ 0.95) and fixed (f > 0.95) TE insertions in each genomic region in (A) Tenerife and (B) Itabuna.	Page 25
Figure S15: Phylogenetic analysis of the 79 copies of the LTR/Gypsy *CobsR.176* element in the *C. obscurior* genome.	Page 26

MOLECULAR ECOLOGY

Supporting Tables

Table S1: Quality control metrics for the single individual and pool data used in this study.

Population	Sample name	GPS	No. reads	Mapping percentage	Mean coverage	Mean mapping quality
Leiden, NL	leiden6	52°09'24.5"N 4°28'59.2"E	22.708.357	99,42%	16	45
Leiden, NL	leiden3	52°09'24.5"N 4°28'59.2"E	20.896.837	98,68%	15	45
Taipei, TW	taiwan1	-	20.694.396	91,99%	13	45
Taipei, TW	taiwan2	-	20.147.723	97,38%	14	45
Itabuna, BR	itabuna1	14°45'22.6"S 39°13'50.6"W	20.311.547	99,18%	14	47
Itabuna, BR	itabuna2	14°45'43.1"S 39°13'57.1"W	16.762.591	99,05%	12	47
Itabuna, BR	itabuna3	14°47'29.2"S 39°11'14.1"W	17.851.827	98,15%	12	47
Itabuna, BR	itabuna4	14°46'49.2"S 39°12'54.6"W	21.950.560	98,34%	15	47
Una, BR	una1	15°15'48.5"S 39°05'00.1"W	19.555.998	99,05%	14	47
Una, BR	una3	15°13'14.8"S 39°02'08.6"W	18.914.093	98,77%	13	47
Una, BR	una29	15°17'12.8"S 39°03'49.1"W	19.637.806	98,79%	14	47
Una, BR	una2	15°17'12.8"S 39°03'49.1"W	20.623.229	98,57%	14	47
Guaruja, BR	guaruja2	23°59'19.3"S 46°14'31.3"W	21.092.709	96,61%	15	46
Guaruja, BR	guaruja1	23°59'18.8"S 46°14'31.8"W	22.771.440	92,68%	15	46
Guaruja, BR	guarujaes	23°59'24.9"S 46°14'33.9"W	22.880.849	96,69%	16	47
Guaruja, BR	guaruja3	23°59'23.4"S 46°14'27.5"W	16.760.982	99,22%	12	45
Tenerife, ES	Pool_Tenerife	16 colonies; Figure 2	100.970.500	99,35%	75	45
Itabuna, ES	Pool_Itabuna	30 colonies; Figure 2	99.159.021	98,92%	73	47

MOLECULAR ECOLOGY

Table S2: List of published ant genomes used to generate *de novo* repeat libraries with RepeatScout.

Species	Common name	GenBank assembly accession	Assembly name
Harpegnathos saltator	Jerdon's jumping ant	GCA_003227715.1	Hsal_v8.5
Dinoponera quadriceps		GCA_001313825.1	ASM131382v1
Ooceraea biroi	Clonal raider ant	GCA_003672135.1	Obir_v5.4
Pseudomyrmex gracilis	Graceful twig ant	GCA_002006095.1	ASM200609v1
Linepithema humile	Argentine ant	GCA_000217595.1	Lhum_UMD_V04
Formica exsecta	Wood ant	GCA_003651465.1	ASM365146v1
Camponotus floridanus	Florida carpenter ant	GCA_003227725.1	Cflo_v7.5
Pogonomyrmex barbatus	Red harvester ant	GCA_000187915.1	Pbar_UMD_V03
Solenopsis invicta	Red fire ant	GCA_000188075.2	Si_gnH
Monomorium pharaonis	Pharaoh ant	GCA_003260585.1	UPENN_Mphar_2.0
Temnothorax curvispinosus		GCA_003070985.1	ASM307098v1
Vollenhovia emeryi		GCA_000949405.1	V.emery_V1.0
Wasmannia auropunctata	Little fire ant	GCA_000956235.1	wasmannia.A_1.0
Cyphomyrmex costatus	Fungus-growing ant	GCA_001594065.1	Ccosl1.0
Trachymyrmex zeteki		GCA_001594055.1	Tzet1.0
Trachymyrmex cornetzi		GCA_001594075.1	Tcor1.0
Trachymyrmex septentrionalis		GCA_001594115.1	Tsep1.0
Acromyrmex echinatior	Leafcutter ants	GCA_000204515.1	Aech_3.9
Atta cephalotes		GCA_000143395.2	Attacep1.0
Atta colombica		GCA_001594045.1	Acol1.0

MOLECULAR ECOLOGY

Table S3: TE-related functional annotation terms used for identifying TE-encoded proteins.

ID	Description
G3DSA:3.30.70.270	Reverse transcriptase/Diguanylate cyclase domain
SSF53098	Ribonuclease H-like superfamily
PF00078	Reverse transcriptase (RNA-dependent DNA polymerase)
G3DSA:3.30.420.10	Ribonuclease H-like superfamily/Ribonuclease H
PS50878	Reverse transcriptase (RT) catalytic domain profile.
PS50994	Integrase catalytic domain profile.
G3DSA:3.10.10.10	HIV Type 1 Reverse Transcriptase, subunit A, domain 1
PF00665	Integrase core domain
G3DSA:3.10.20.370	retrotransposable element/transposon Tf2-type
PF17921	Integrase zinc binding domain
cd01647	RT_LTR
SSF57756	Retrovirus zinc finger-like domains superfamily
cd09274	RNase_HI_RT_Ty3
G3DSA:1.10.340.70	Ribonuclease H superfamily
PF17919	RNase H-like domain found in reverse transcriptase
cd00303	retropepsin_like
G3DSA:3.30.420.470	transposase type 1
cd01650	RT_nLTR_like
PF14529	Endonuclease-reverse transcriptase
PF05380	Pao retrotransposon peptidase
PF17917	RNase H-like domain found in reverse transcriptase
PF12259	Baculovirus F protein
PS50175	Aspartyl protease, retroviral-type family profile.
PF00077	Retroviral aspartyl protease
cd09077	R1-I-EN
PF07727	Reverse transcriptase (RNA-dependent DNA polymerase)
PF14223	gag-polypeptide of LTR copia-type
PF03732	Retrotransposon gag protein
PF04665	Poxvirus A32 protein
cd09272	RNase_HI_RT_Ty1
PF13359	DDE superfamily endonuclease
PF01498	Transposase
PF13976	GAG-pre-integrase domain
PF13975	gag-polyprotein putative aspartyl protease

PF14214	Helitron helicase-like domain at N-terminus
PF05699	hAT family C-terminal dimerisation region
cd09275	RNase_HI_RT_DIRS1
PF12017	Transposase protein
PF14787	GAG-polyprotein viral zinc-finger
cd09276	Rnase_HI_RT_non_LTR
PF01359	Transposase (partial DDE domain)
IPR000477	Reverse transcriptase domain
IPR012337	Ribonuclease H-like superfamily
IPR036397	Ribonuclease H superfamily
IPR041577	Reverse transcriptase/retrotransposon-derived protein, RNase H-like domain
IPR008042	Retrotransposon, Pao
IPR041373	Reverse transcriptase, RNase H-like domain
IPR022048	Envelope fusion protein-like
IPR001995	Peptidase A2A, retrovirus, catalytic
IPR013103	Reverse transcriptase, RNA-dependent DNA polymerase
IPR005162	Retrotransposon gag domain
IPR038717	Tc1-like transposase, DDE domain
IPR027806	Harbinger transposase-derived nuclease domain
IPR002492	Transposase, Tc1-like
IPR002156	Ribonuclease H domain
IPR025724	GAG-pre-integrase domain
IPR025476	Helitron helicase-like domain
IPR008906	HAT, C-terminal dimerisation domain
IPR021896	Transposase protein
IPR004211	Recombination endonuclease VII
IPR010998	Integrase/recombinase, N-terminal
IPR041426	Mos1 transposase, HTH domain
IPR034132	Retropepsin Saci-like domain
IPR026103	Harbinger transposase-derived nuclease, animal
IPR024445	ISXO2-like transposase domain
IPR029526	PiggyBac transposable element-derived protein
IPR018289	MULE transposase domain
IPR029472	Retrotransposon Copia-like, N-terminal
IPR025898	Tc3 transposase, DNA binding domain

MOLECULAR ECOLOGY

Table S4: QUAST based statistics calculated for the Cobs2.1 genome assembly.

Assembly	**Cobs.alpha.2.1**
# contigs	127
# contigs (>= 1000 bp)	127
# contigs (>= 5000 bp)	125
# contigs (>= 10000 bp)	121
# contigs (>= 25000 bp)	105
# contigs (>= 50000 bp)	91
Total length (>= 0 bp)	193051228
Total length (>= 1000 bp)	193051228
Total length (>= 5000 bp)	193047645
Total length (>= 10000 bp)	193025568
Total length (>= 25000 bp)	192755434
Total length (>= 50000 bp)	192237042
Largest contig	13148674
Total length	193051228
GC (%)	41,02
N50	6290588
N75	4487289
L50	11
L75	21
# N's per 100 kbp	94,76

MOLECULAR ECOLOGY

Table S5: Location and length of each of the 34 TE islands identified in the Cobs2.1 genome assembly.

Scaffold	Start	End	Length (bp)
1	6250000	8300000	2050000
2	0	1500000	1500000
2	12000000	12360777	360777
3	0	500000	500000
4	0	1000000	1000000
4	8400000	9000000	600000
5	7500000	8562622	1062622
6	0	900000	900000
7	0	800000	800000
8	6700000	7505125	805125
9	5400000	6290588	890588
10	5700000	6032196	332196
12	5200000	5755492	555492
13	7000000	7284943	284943
14	4400000	5477151	1077151
15	3650000	5070358	1420358
16	0	1000000	1000000
17	0	1200000	1200000
17	4400000	4616616	216616
18	4370000	4606763	236763
19	0	1000000	1000000
20	0	1100000	1100000
20	4400000	4487289	87289
21	3700000	4384715	684715
21	0	300000	300000
22	0	500000	500000
22	4200000	4241709	41709
23	3700000	4200000	500000
24	3200000	3831917	631917
25	5800000	6706168	906168
25	0	500000	500000
26	3200000	3629807	429807
27	2600000	3200000	600000
30	0	500000	500000

MOLECULAR ECOLOGY

Table S6: Results of Gene Ontology Enrichment analyses with topGO of genes contained in TE islands in the Cobs2.1 genome assembly.

GO.ID	Term	Annotated	Significant	Expected	parentChild	Ontology
GO:0004984	olfactory receptor activity	232	30	16,26	0,00012	MF
GO:0050660	flavin adenine dinucleotide binding	55	10	3,85	0,00056	MF
GO:0005549	odorant binding	245	30	17,17	0,00261	MF
GO:0003676	nucleic acid binding	982	99	68,82	0,00270	MF
GO:0031177	phosphopantetheine binding	7	3	0,49	0,00647	MF
GO:1901363	heterocyclic compound binding	1852	155	129,79	0,00669	MF
GO:0008173	RNA methyltransferase activity	19	5	1,33	0,00689	MF
GO:0097159	organic cyclic compound binding	1855	155	130,00	0,00721	MF
GO:0033218	amide binding	18	5	1,26	0,00753	MF
GO:0004312	fatty acid synthase activity	7	3	0,49	0,01416	MF
GO:0048037	cofactor binding	267	29	18,71	0,01622	MF
GO:0043169	cation binding	710	64	49,76	0,01830	MF
GO:0016491	oxidoreductase activity	386	36	27,05	0,01884	MF
GO:0072341	modified amino acid binding	10	3	0,70	0,03081	MF
GO:0008442	3-hydroxyisobutyrate dehydrogenase activ...	1	1	0,07	0,03226	MF
GO:0016614	oxidoreductase activity, acting on CH-OH...	51	9	3,57	0,03327	MF
GO:0016830	carbon-carbon lyase activity	22	3	1,54	0,03526	MF
GO:0000062	fatty-acyl-CoA binding	6	2	0,42	0,04163	MF
GO:0004516	nicotinate phosphoribosyltransferase act...	1	1	0,07	0,04762	MF
GO:0019725	cellular homeostasis	26	5	1,57	0,00710	BP
GO:0003008	system process	279	38	16,87	0,00940	BP
GO:0045454	cell redox homeostasis	15	3	0,91	0,01680	BP
GO:0000393	spliceosomal conformational changes to g...	1	1	0,06	0,02940	BP
GO:0007062	sister chromatid cohesion	3	1	0,18	0,03900	BP
GO:0006396	RNA processing	171	12	10,34	0,04370	BP
GO:0042592	homeostatic process	37	5	2,24	0,04510	BP
GO:0016042	lipid catabolic process	10	2	0,60	0,04690	BP
GO:0010256	endomembrane system organization	10	2	0,60	0,04700	BP
GO:0044423	virion part	3	3	0,17	0,00018	CC
GO:0019012	virion	3	3	0,17	0,00018	CC
GO:0016020	membrane	1254	82	71,18	0,00760	CC
GO:0072546	ER membrane protein complex	1	1	0,06	0,04167	CC

MOLECULAR ECOLOGY

Table S7: Most abundant repeat families identified in populations of *C. obscurior* from Tenerife and Itabuna, using dnaPipeTE with 0.1× coverage per samples.

Family	Tenerife (%)	Itabuna (%)
Simple_repeat	8.76	7.33
Unclassified	8.1	10.99
Gypsy	2.01	1.39
R1	0.79	0.83
R2	0.03	0.12
Helitron	0.17	0.17
TcMar-Tc	0	0.03
Low_complexity	0.22	0.04
Copia	0.04	0.1
Total	20.12	21

MOLECULAR ECOLOGY

Table S8: Contingency table showing the number of low frequency, common and fixed TE insertions in TE islands as well as LDRs in the Itabuna and Tenerife population.

	Low frequency		Common		Fixed	
	TE islands	**LDRs**	**TE islands**	**LDRs**	**TE islands**	**LDRs**
Itabuna	182	2587	1091	4236	1218	66
Tenerife	188	2350	999	2506	914	31

MOLECULAR ECOLOGY

Supporting figures

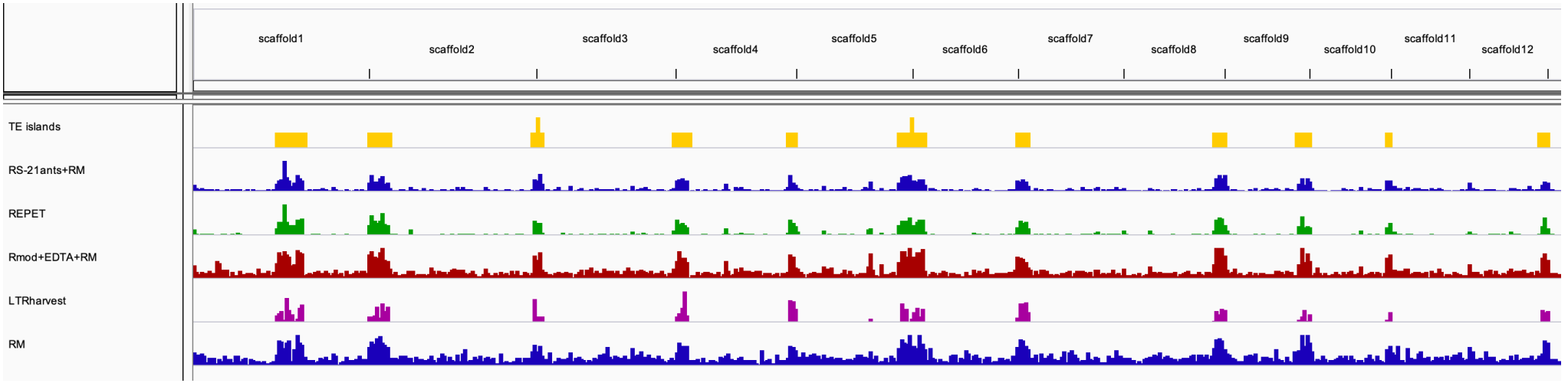

Figure S1: IGV screenshot showing the relative TE content across the genome of *C. obscurior* recovered using different repeat annotation approaches. The different approaches produced the same genome-wide patterns of TE distribution. RS-21ants+RM refers to the primary TE annotation, which combined *de novo* repeat annotations from 21 ant genomes with arthropod-specific repeats from RepBase and Hymenoptera-specific repeats from ArTEdb. TE islands are visible as prominently TE-enriched regions on all scaffolds.

MOLECULAR ECOLOGY

Figure S2: Close-up view of TE annotations across the TE island on scaffold 1. The results obtained by our primary approach (RS-21ants+RM) outperform the other strategies. The REPET-based annotation tends to combine independent loci to chimeric TEs, while the Repeat-Modeler/EDTA-based approach tends to miss loci consistently identified by the three other approaches.

MOLECULAR ECOLOGY

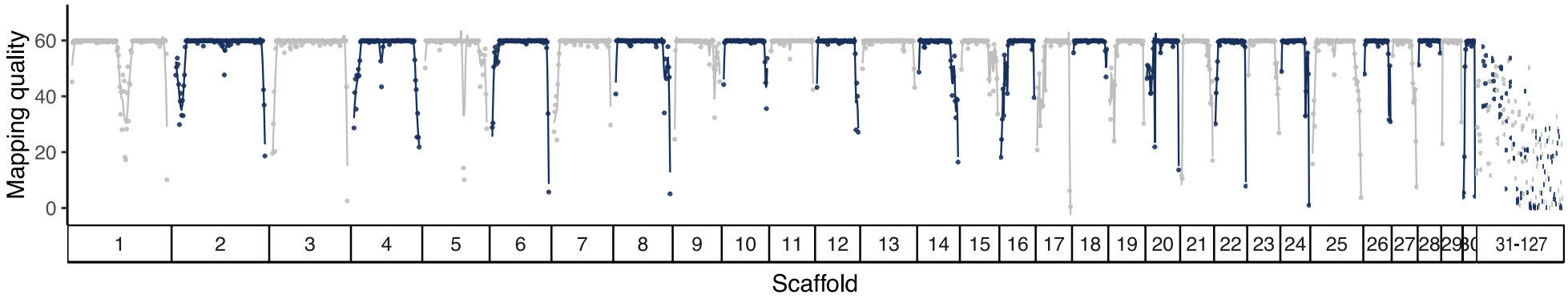

Figure S3: Mapping quality across the 127 scaffolds of the Cobs2.1 assembly. Each dot represents average mapping quality in 100 kb window, generated using the pooled data. Scaffolds 31-127 are excluded from all analyses due to fragmentation and low mapping quality.

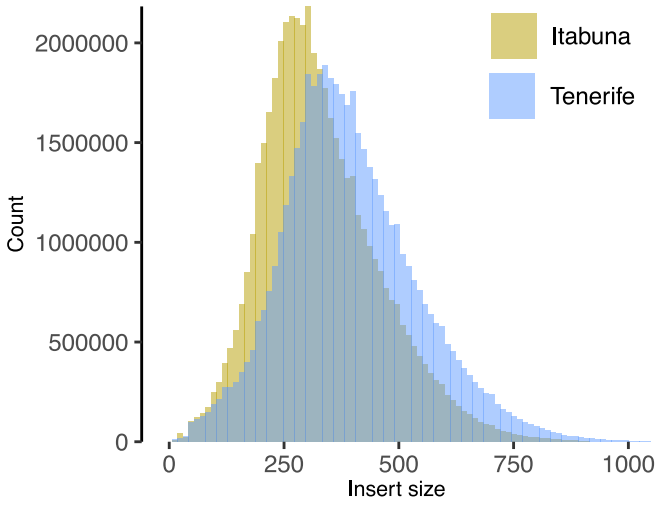

Figure S4: Insert size distribution in pool-seq data of Tenerife and Itabuna, estimated from the alignment files.

MOLECULAR ECOLOGY

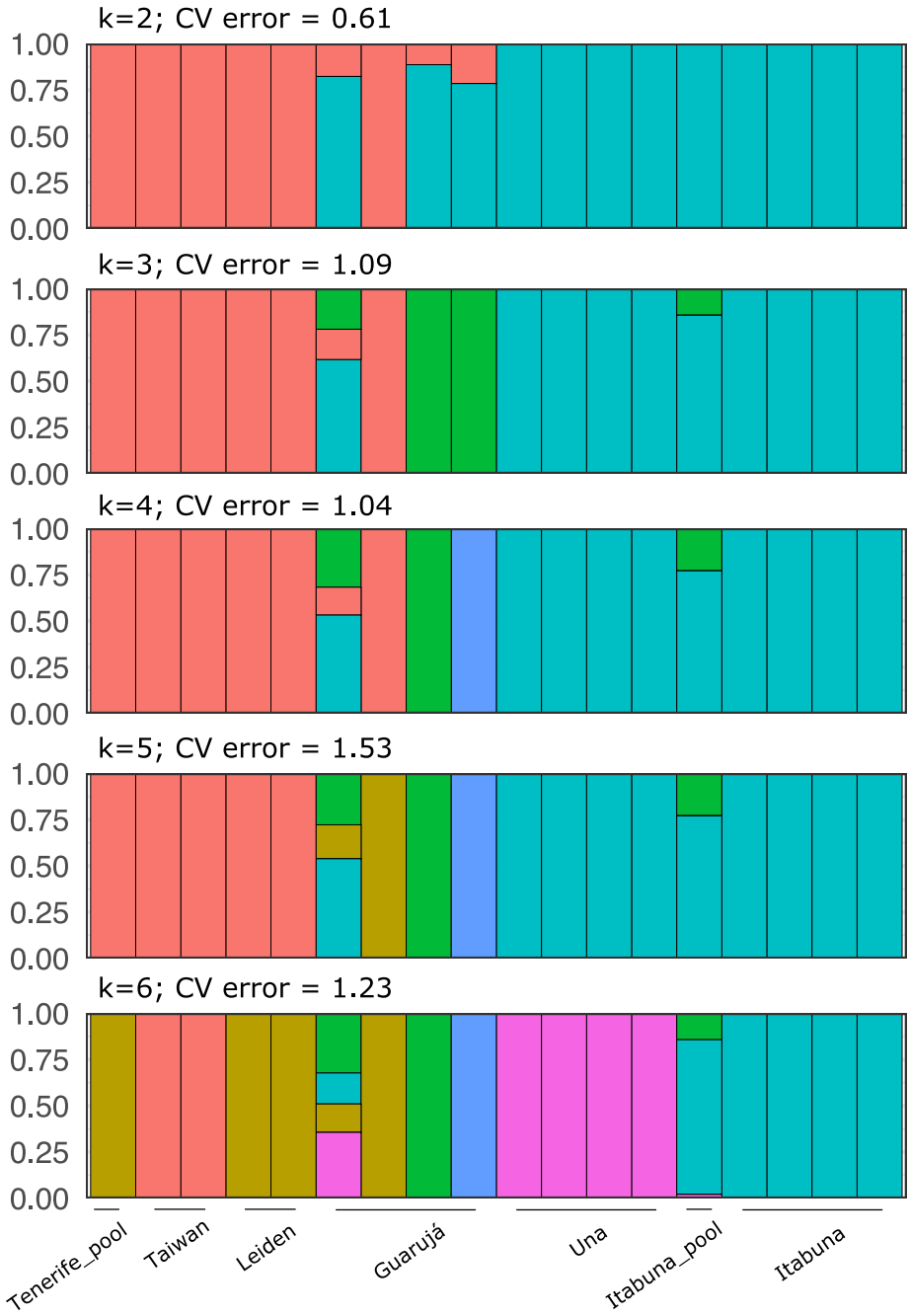

Figure S5: Admixture patterns among *C. obscurior* populations by using ADMIXTURE at K = 2, K = 3, k = 4, k = 5 and k = 6. For each value of k, the cross-validation error rate is presented. Note that K = 2 that exhibits a low CV error (0.61) is the best modeling choice.

MOLECULAR ECOLOGY

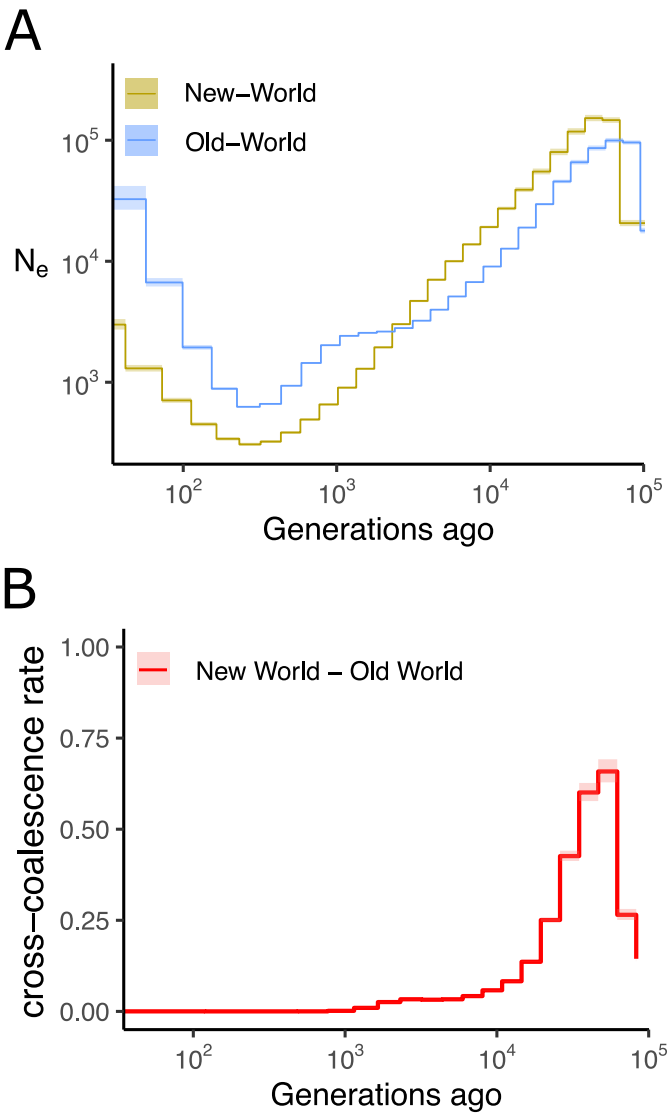

Figure S6: Population history in *C. obscurior* estimated using MSMC2 with eight phased haplotypes, representing the New and Old World lineages respectively. As representative of the New World lineage, we used four individuals from Itabuna. Lines and shaded areas are means and 95% confidence intervals, respectively. The inferred population history (A) and rCCR (B) matched the estimates obtained when using four individuals from Itabuna and Una, Brazil (Figure 2D) as representatives of the New World lineage.

MOLECULAR ECOLOGY

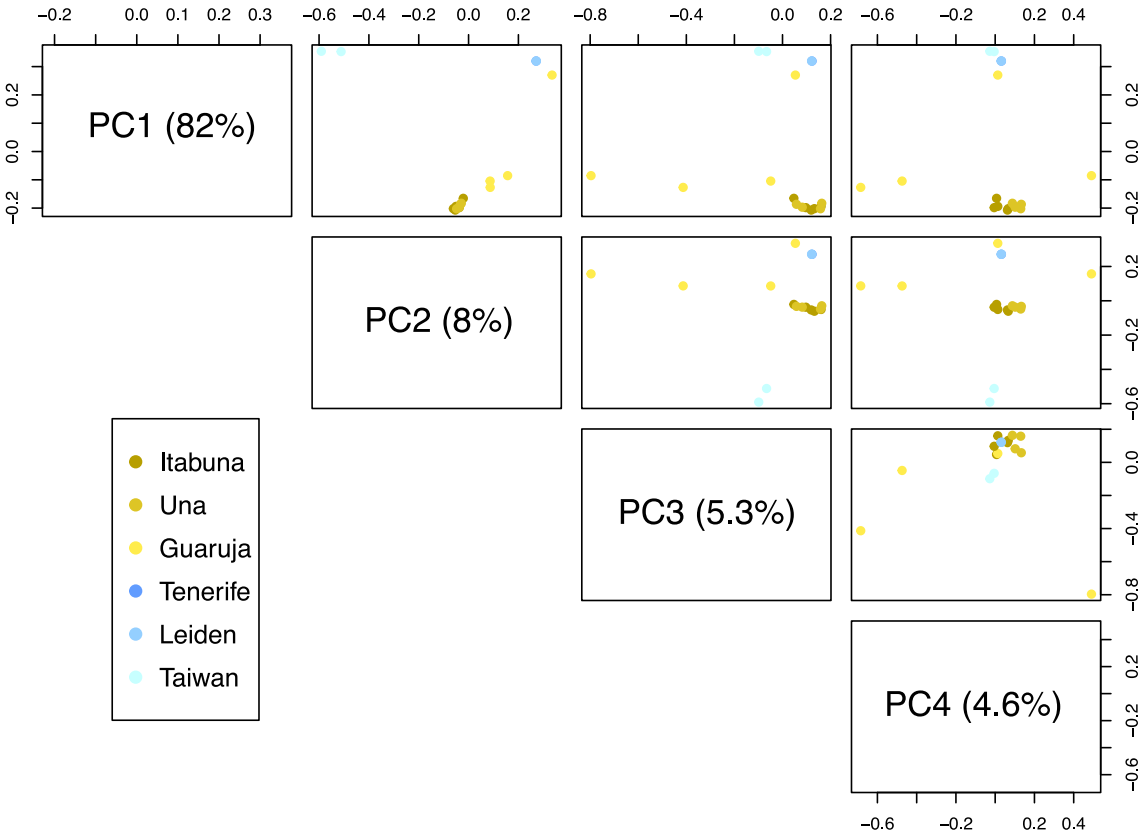

Figure S7: Principal Component Analysis based on 115,334 SNPs of samples belonging to the Old and New World lineages. The scatterplot matrix shows the four principal components. Note that dots representing the samples of Leiden and Tenerife are overlapping.

MOLECULAR ECOLOGY

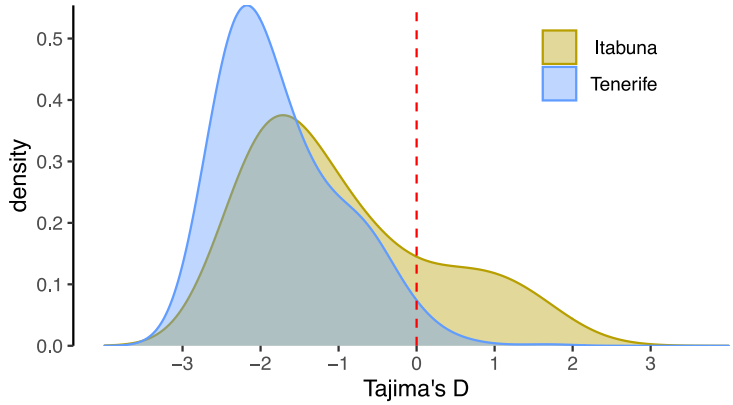

Figure S8: Density plot showing the distribution Tajima's *D* values in *C. obscurior*, estimated in 100-kb non-overlapping windows for the Tenerife and Itabuna population. The number of genomic windows evolving neutrally (Tajima's $D \simeq 0$) or under balancing selection (Tajima's $D > 0$) is much higher in Itabuna than in Tenerife (Fisher's exact test, $p < 2.2e\text{-}16$).

MOLECULAR ECOLOGY

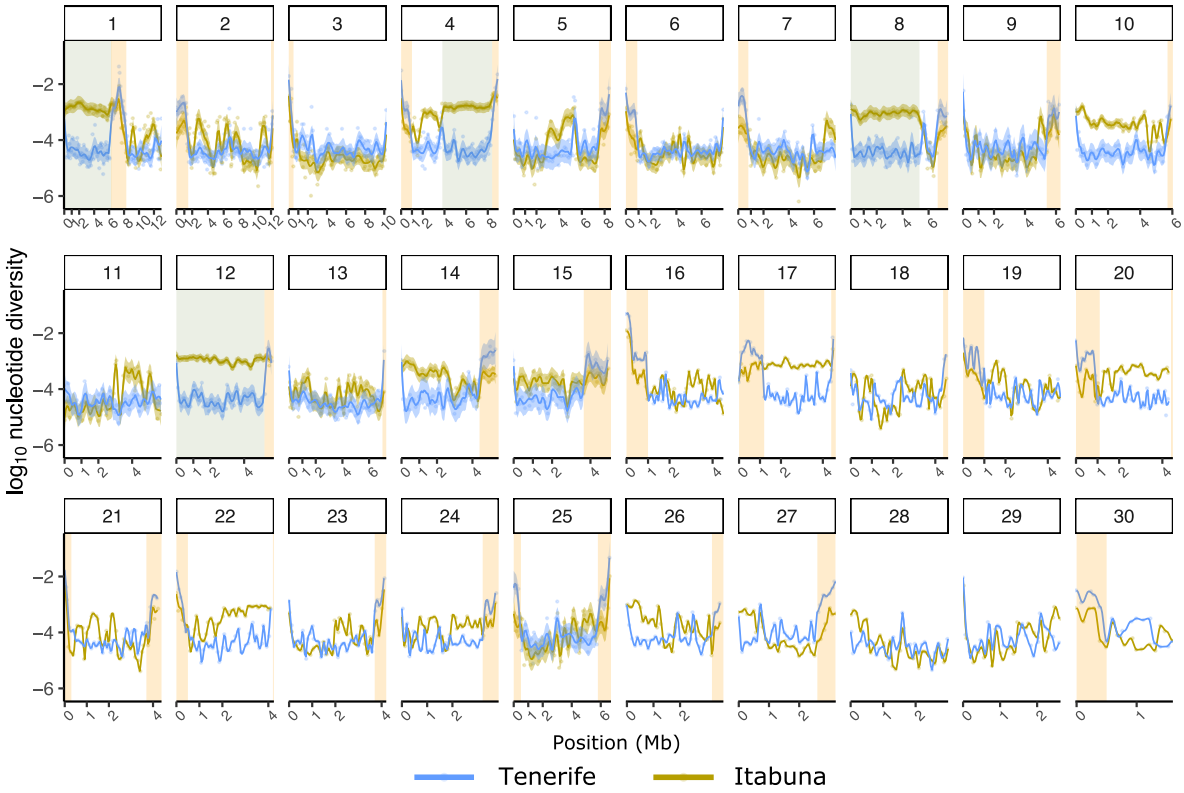

Figure S9: Genome-wide distribution of genetic diversity (π) of the 30 largest *C. obscurior* genome scaffolds. All estimates were calculated in 100-kb non-overlapping windows. Genomic position (Mb) is presented on the x-axis. Regions highlighted in orange are TE islands and in green are potentially introgressed regions. Lines and shaded areas are means and 95 % confidence intervals, respectively.

MOLECULAR ECOLOGY

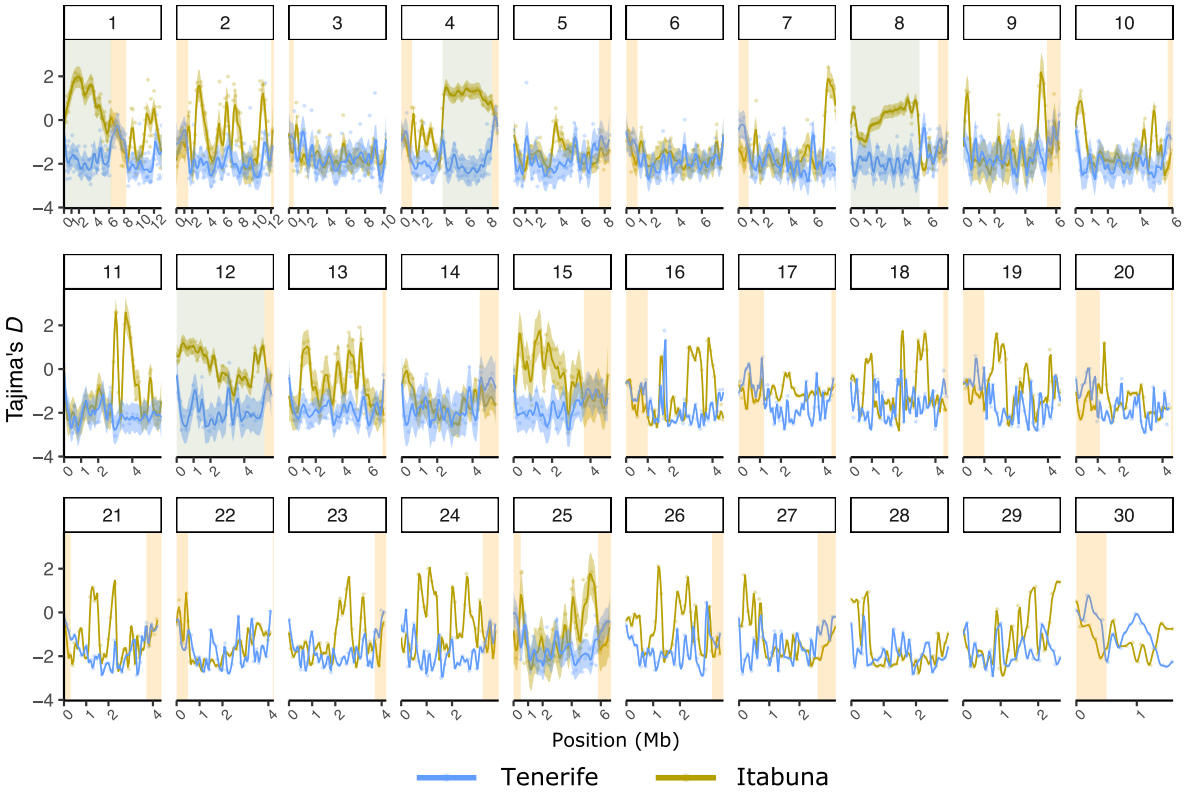

Figure S10: Genome-wide distribution of Tajima's *D* of the 30 largest *C. obscurior* genome scaffolds. All estimates were calculated in 100-kb non-overlapping windows. Genomic position (Mb) is presented on the x-axis. Regions highlighted in orange are TE islands and in green are potentially introgressed regions. Lines and shaded areas are means and 95 % confidence intervals, respectively.

MOLECULAR ECOLOGY

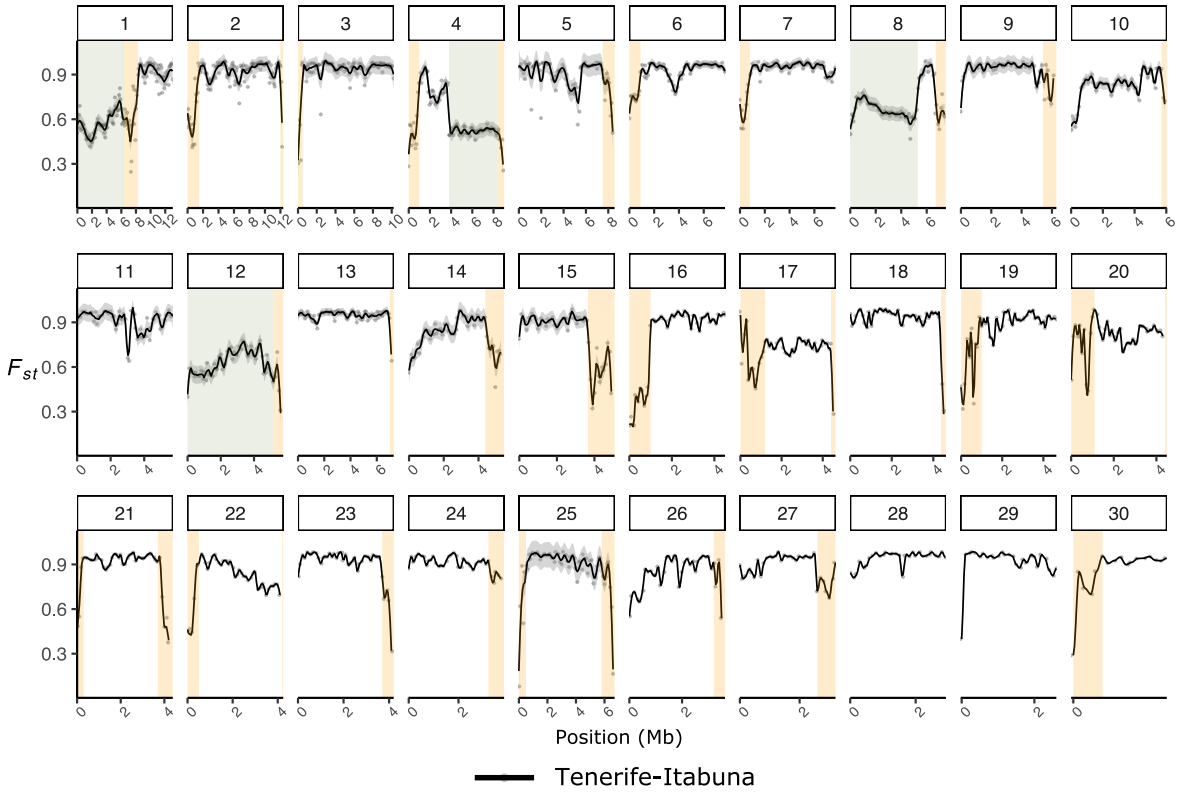

Figure S11: Genome-wide distribution of genetic differentiation (F_{st}) of the 30 largest *C. obscurior* genome scaffolds. All estimates were calculated in 100-kb non-overlapping windows. Genomic position (Mb) is presented on the x-axis. Regions highlighted in pink are TE islands and in green are potentially introgressed regions. Lines and shaded areas are means and 95 % confidence intervals, respectively.

MOLECULAR ECOLOGY

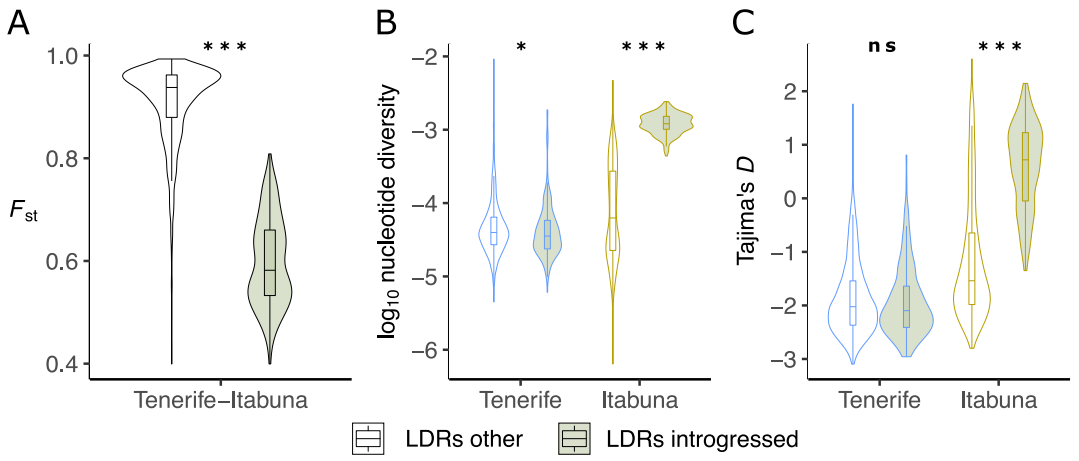

Figure S12: Genetic diversity and differentiation in introgressed regions compared to the remainder of LDRs. (A) Genetic differentiation (B) nucleotide diversity (π) and (C) Tajima's *D* in two populations of *C. obscurior*, from Tenerife and Itabuna. *p<0.05, ***p < 0.0001.

MOLECULAR ECOLOGY

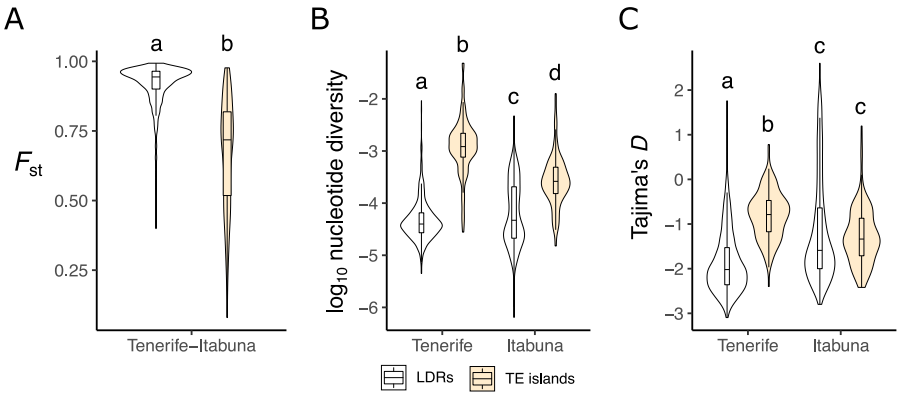

Figure S13: Genetic diversity and differentiation within LDRs and TE islands in each population after excluding introgressed regions. (A) Genetic differentiation, (B) nucleotide diversity (π) and (D) Tajima's *D* in two populations of *C. obscurior*, from Tenerife and Itabuna. For F_{st}, we performed a Wilcoxon rank sum test (W = 178680, p < 0.0001). We performed a Kruskal-Wallis rank sum test for π (χ^2 = 599.79, df = 3, p < 0.0001) and Tajima's *D* (χ^2 = 450.72, df = 3, p < 0.0001), followed by pairwise Wilcoxon rank sum *post hoc* tests. Different letters represent significant differences (p < 0.001).

MOLECULAR ECOLOGY

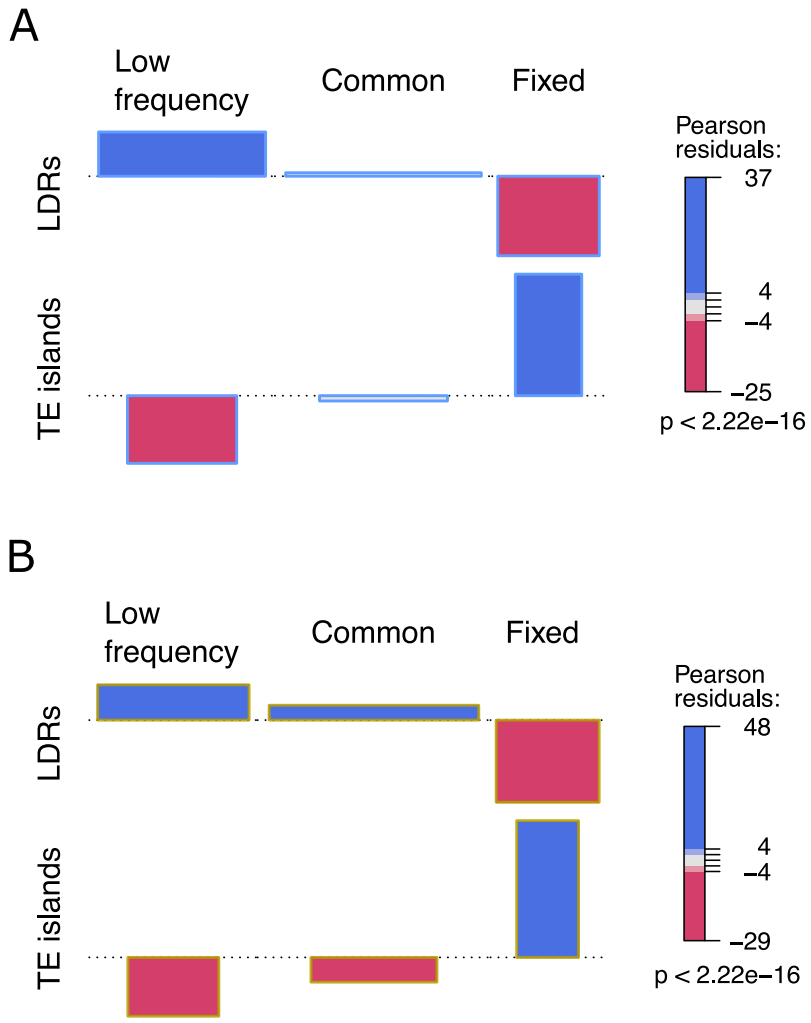

Figure S14: Association plots showing the signed contribution to Pearson's χ^2 (left panels) and the Pearson's χ^2 standardized residuals (right panels) calculated for low frequency (f < 0.25), common (0.25 ≤ f ≤ 0.95) and fixed (f > 0.95) TE insertions in each genomic region in (A) Tenerife and (B) Itabuna. Blue color of the rectangles indicates a positive correlation, while dark pink indicates a negative correlation between variables on the y and x-axis. The area of the rectangles is proportional to the difference between observed and expected frequencies of each cell in Table S6.

MOLECULAR ECOLOGY

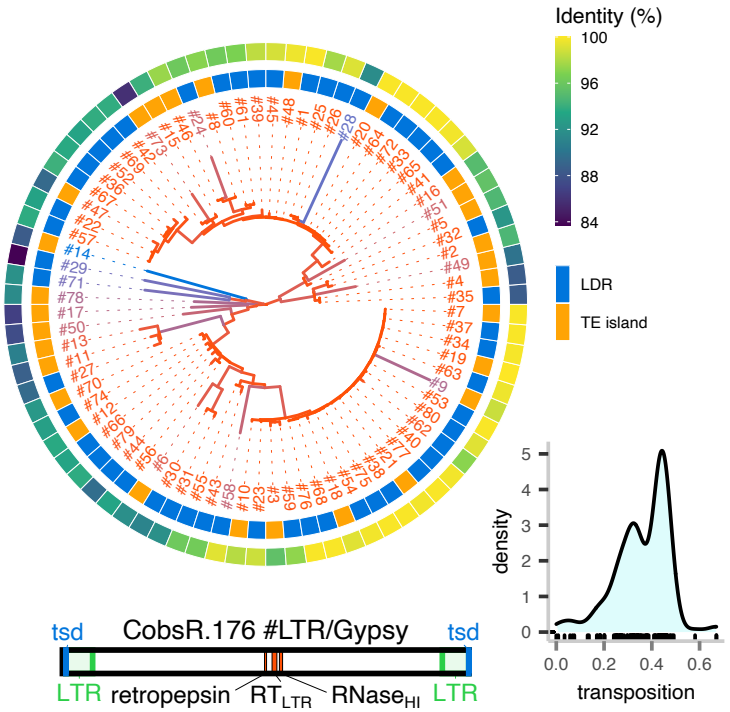

Figure S15: Phylogenetic analysis of the 79 copies of the LTR/Gypsy *CobsR.176* element in the *C. obscurior* genome. The inner circle shows whether a copy is in LDRs or TE islands. The outer circle shows divergence (as calculated from RepeatMasker). The bottom illustration is the structure of the Gypsy with tandem site duplication (TSD), long terminal repeat (LTR) and the conserved protein domains. The density plot at the bottom right shows the dynamic in the phylogeny, with two periods in time (the two peaks) where most duplications occurred.