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 Errbiï1, Jens Keilwagen2, Katharina J. Hoff3,4, Raphael Steffen1, Janine Altmüller5, Jan Oettler6, Lukas Schrader1

1Institute for Evolution and Biodiversity, University of Münster, 48149 Münster, Germany
2Julius Kühn Institute (JKI) – Federal Research Centre for Cultivated Plants, Institute for Biosafety in Plant Biotechnology, Erwin-Baur-Str. 27, 06484 Quedlinburg, Germany
3Institute of Mathematics and Computer Science, University of Greifswald, Walther-Rathenau-Str. 47, 17489 Greifswald, Germany
4Center for Functional Genomics of Microbes, University of Greifswald, Felix-Haussdorff-Str. 8, 17489 Greifswald, Germany
5Cologne Center for Genomics, Institute of Human Genetics, University of Cologne, 50931 Cologne, Germany
6Lehrstuhl für Zoologie/Evolutionsbiologie, University Regensburg, 93053 Regensburg, Germany

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

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

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

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample name</th>
<th>GPS</th>
<th>No. reads</th>
<th>Mapping percentage</th>
<th>Mean coverage</th>
<th>Mean mapping quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leiden, NL</td>
<td>leiden6</td>
<td>52°09'24.5&quot;N 4°28'59.2&quot;E</td>
<td>22.708.357</td>
<td>99,42%</td>
<td>16</td>
<td>45</td>
</tr>
<tr>
<td>Leiden, NL</td>
<td>leiden3</td>
<td>52°09'24.5&quot;N 4°28'59.2&quot;E</td>
<td>20.896.837</td>
<td>98,68%</td>
<td>15</td>
<td>45</td>
</tr>
<tr>
<td>Taipei, TW</td>
<td>taiwan1</td>
<td>-</td>
<td>20.694.396</td>
<td>91,99%</td>
<td>13</td>
<td>45</td>
</tr>
<tr>
<td>Taipei, TW</td>
<td>taiwan2</td>
<td>-</td>
<td>20.147.723</td>
<td>97,38%</td>
<td>14</td>
<td>45</td>
</tr>
<tr>
<td>Itabuna, BR</td>
<td>itabuna1</td>
<td>14°45'22.6&quot;S 39°13'50.6&quot;W</td>
<td>20.311.547</td>
<td>99,18%</td>
<td>14</td>
<td>47</td>
</tr>
<tr>
<td>Itabuna, BR</td>
<td>itabuna2</td>
<td>14°45'43.1&quot;S 39°13'57.1&quot;W</td>
<td>16.762.591</td>
<td>99,05%</td>
<td>12</td>
<td>47</td>
</tr>
<tr>
<td>Itabuna, BR</td>
<td>itabuna3</td>
<td>14°47'29.2&quot;S 39°11'14.1&quot;W</td>
<td>17.851.827</td>
<td>98,15%</td>
<td>12</td>
<td>47</td>
</tr>
<tr>
<td>Itabuna, BR</td>
<td>itabuna4</td>
<td>14°46'49.2&quot;S 39°12'54.6&quot;W</td>
<td>21.950.560</td>
<td>98,34%</td>
<td>15</td>
<td>47</td>
</tr>
<tr>
<td>Una, BR</td>
<td>una1</td>
<td>15°15'48.5&quot;S 39°05'00.1&quot;W</td>
<td>19.555.998</td>
<td>99,05%</td>
<td>14</td>
<td>47</td>
</tr>
<tr>
<td>Una, BR</td>
<td>una3</td>
<td>15°13'14.8&quot;S 39°02'08.6&quot;W</td>
<td>18.914.093</td>
<td>98,77%</td>
<td>13</td>
<td>47</td>
</tr>
<tr>
<td>Una, BR</td>
<td>una29</td>
<td>15°17'12.8&quot;S 39°03'49.1&quot;W</td>
<td>19.637.806</td>
<td>98,79%</td>
<td>14</td>
<td>47</td>
</tr>
<tr>
<td>Una, BR</td>
<td>una2</td>
<td>15°17'12.8&quot;S 39°03'49.1&quot;W</td>
<td>20.623.229</td>
<td>98,57%</td>
<td>14</td>
<td>47</td>
</tr>
<tr>
<td>Guaruja, BR</td>
<td>guaruja2</td>
<td>23°59'19.3&quot;S 46°14'31.3&quot;W</td>
<td>21.092.709</td>
<td>96,61%</td>
<td>15</td>
<td>46</td>
</tr>
<tr>
<td>Guaruja, BR</td>
<td>guaruja1</td>
<td>23°59'18.8&quot;S 46°14'31.8&quot;W</td>
<td>22.771.440</td>
<td>92,68%</td>
<td>15</td>
<td>46</td>
</tr>
<tr>
<td>Guaruja, BR</td>
<td>guarujaes</td>
<td>23°59'24.9&quot;S 46°14'33.9&quot;W</td>
<td>22.880.849</td>
<td>96,69%</td>
<td>16</td>
<td>47</td>
</tr>
<tr>
<td>Guaruja, BR</td>
<td>guaruja3</td>
<td>23°59'23.4&quot;S 46°14'27.5&quot;W</td>
<td>16.760.982</td>
<td>99,22%</td>
<td>12</td>
<td>45</td>
</tr>
<tr>
<td>Tenerife, ES</td>
<td>Pool_Tenerife</td>
<td>16 colonies; Figure 2</td>
<td>100.970.500</td>
<td>99,35%</td>
<td>75</td>
<td>45</td>
</tr>
<tr>
<td>Itabuna, ES</td>
<td>Pool_Itabuna</td>
<td>30 colonies; Figure 2</td>
<td>99.159.021</td>
<td>98,92%</td>
<td>73</td>
<td>47</td>
</tr>
</tbody>
</table>
**Table S2:** List of published ant genomes used to generate de novo repeat libraries with RepeatScout.

<table>
<thead>
<tr>
<th>Species</th>
<th>Common name</th>
<th>GenBank assembly accession</th>
<th>Assembly name</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Harpegnathos saltator</em></td>
<td>Jerdon’s jumping ant</td>
<td>GCA_003227715.1</td>
<td>Hsal_v8.5</td>
</tr>
<tr>
<td><em>Dinoponera quadriiceps</em></td>
<td></td>
<td>GCA_001313825.1</td>
<td>ASM131382v1</td>
</tr>
<tr>
<td><em>Ooceraea biroii</em></td>
<td>Clonal raider ant</td>
<td>GCA_003672135.1</td>
<td>Obir_v5.4</td>
</tr>
<tr>
<td><em>Pseudomyrmex gracilis</em></td>
<td>Graceful twig ant</td>
<td>GCA_002006095.1</td>
<td>ASM200609v1</td>
</tr>
<tr>
<td><em>Linepithema humile</em></td>
<td>Argentine ant</td>
<td>GCA_000217595.1</td>
<td>Lhum_UMD_V04</td>
</tr>
<tr>
<td><em>Formica exsecta</em></td>
<td>Wood ant</td>
<td>GCA_003651465.1</td>
<td>ASM365146v1</td>
</tr>
<tr>
<td><em>Camponotus floridanus</em></td>
<td>Florida carpenter ant</td>
<td>GCA_003227725.1</td>
<td>Cflo_v7.5</td>
</tr>
<tr>
<td><em>Pogonomyrmex barbatus</em></td>
<td>Red harvester ant</td>
<td>GCA_000187915.1</td>
<td>Pbar_UMD_V03</td>
</tr>
<tr>
<td><em>Solenopsis invicta</em></td>
<td>Red fire ant</td>
<td>GCA_000188075.2</td>
<td>Si_gnH</td>
</tr>
<tr>
<td><em>Monomorium pharaonis</em></td>
<td>Pharaoh ant</td>
<td>GCA_003260585.1</td>
<td>UPENN_Mphar_2.0</td>
</tr>
<tr>
<td><em>Temnothorax curvispinosus</em></td>
<td></td>
<td>GCA_003070985.1</td>
<td>ASM307098v1</td>
</tr>
<tr>
<td><em>Vollenhovia emeryi</em></td>
<td></td>
<td>GCA_000949405.1</td>
<td>V.emery_V1.0</td>
</tr>
<tr>
<td><em>Wasmannia auropunctata</em></td>
<td>Little fire ant</td>
<td>GCA_000956235.1</td>
<td>wasmannia.A_1.0</td>
</tr>
<tr>
<td><em>Cyphomyrmex costatus</em></td>
<td></td>
<td>GCA_001594065.1</td>
<td>Ccosl1.0</td>
</tr>
<tr>
<td><em>Trachymyrmex zeteki</em></td>
<td>Fungus-growing ant</td>
<td>GCA_001594055.1</td>
<td>Tzet1.0</td>
</tr>
<tr>
<td><em>Trachymyrmex cornetzi</em></td>
<td></td>
<td>GCA_001594075.1</td>
<td>Tcor1.0</td>
</tr>
<tr>
<td><em>Trachymyrmex septentrionalis</em></td>
<td></td>
<td>GCA_001594115.1</td>
<td>Tsep1.0</td>
</tr>
<tr>
<td><em>Acromyrmex echinatior</em></td>
<td></td>
<td>GCA_00204515.1</td>
<td>Aech_3.9</td>
</tr>
<tr>
<td><em>Atta cephalotes</em></td>
<td>Leafcutter ants</td>
<td>GCA_000143395.2</td>
<td>Attacep1.0</td>
</tr>
<tr>
<td><em>Atta colombica</em></td>
<td></td>
<td>GCA_001594045.1</td>
<td>Acol1.0</td>
</tr>
</tbody>
</table>
Table S3: TE-related functional annotation terms used for identifying TE-encoded proteins.

<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>G3DSA:3.30.70.270</td>
<td>Reverse transcriptase/Diguanilate cyclase domain</td>
</tr>
<tr>
<td>SSF53098</td>
<td>Ribonuclease H-like superfamily</td>
</tr>
<tr>
<td>PF00078</td>
<td>Reverse transcriptase (RNA-dependent DNA polymerase)</td>
</tr>
<tr>
<td>G3DSA:3.30.420.10</td>
<td>Ribonuclease H-like superfamily/Ribonuclease H</td>
</tr>
<tr>
<td>PS50878</td>
<td>Reverse transcriptase (RT) catalytic domain profile.</td>
</tr>
<tr>
<td>PS50994</td>
<td>Integrase catalytic domain profile.</td>
</tr>
<tr>
<td>G3DSA:3.10.10.10</td>
<td>HIV Type 1 Reverse Transcriptase, subunit A, domain 1</td>
</tr>
<tr>
<td>PF00665</td>
<td>Integrase core domain</td>
</tr>
<tr>
<td>G3DSA:3.10.20.370</td>
<td>retrotransposable element/transposon Tf2-type</td>
</tr>
<tr>
<td>PF17921</td>
<td>Integrase zinc binding domain</td>
</tr>
<tr>
<td>cd01647</td>
<td>RT_LTR</td>
</tr>
<tr>
<td>SSF57756</td>
<td>Retrovirus zinc finger-like domains superfamily</td>
</tr>
<tr>
<td>cd09274</td>
<td>RNase_HI_RT_Ty3</td>
</tr>
<tr>
<td>G3DSA:1.10.340.70</td>
<td>Ribonuclease H superfamily</td>
</tr>
<tr>
<td>PF17919</td>
<td>RNase H-like domain found in reverse transcriptase</td>
</tr>
<tr>
<td>cd00303</td>
<td>retropepsin_like</td>
</tr>
<tr>
<td>G3DSA:3.30.420.470</td>
<td>transposase type 1</td>
</tr>
<tr>
<td>cd01650</td>
<td>RT_nLTR_like</td>
</tr>
<tr>
<td>PF14529</td>
<td>Endonuclease-reverse transcriptase</td>
</tr>
<tr>
<td>PF05380</td>
<td>Pao retrotransposon peptidase</td>
</tr>
<tr>
<td>PF17917</td>
<td>RNase H-like domain found in reverse transcriptase</td>
</tr>
<tr>
<td>PF12259</td>
<td>Baculovirus F protein</td>
</tr>
<tr>
<td>PS50175</td>
<td>Aspartyl protease, retroviral-type family profile.</td>
</tr>
<tr>
<td>PF00077</td>
<td>Retroviral aspartyl protease</td>
</tr>
<tr>
<td>cd09077</td>
<td>R1-I-EN</td>
</tr>
<tr>
<td>PF07727</td>
<td>Reverse transcriptase (RNA-dependent DNA polymerase)</td>
</tr>
<tr>
<td>PF14223</td>
<td>gag-polypeptide of LTR copia-type</td>
</tr>
<tr>
<td>PF03732</td>
<td>Retrotransposon gag protein</td>
</tr>
<tr>
<td>PF04665</td>
<td>Poxvirus A32 protein</td>
</tr>
<tr>
<td>cd09272</td>
<td>RNase_HI_RT_Ty1</td>
</tr>
<tr>
<td>PF13359</td>
<td>DDE superfamily endonuclease</td>
</tr>
<tr>
<td>PF01498</td>
<td>Transposase</td>
</tr>
<tr>
<td>PF13976</td>
<td>GAG-pre-integrase domain</td>
</tr>
<tr>
<td>PF13975</td>
<td>gag-polyprotein putative aspartyl protease</td>
</tr>
<tr>
<td>Accession</td>
<td>Description</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
</tr>
<tr>
<td>PF14214</td>
<td>Helitron helicase-like domain at N-terminus</td>
</tr>
<tr>
<td>PF05699</td>
<td>hAT family C-terminal dimerisation region</td>
</tr>
<tr>
<td>cd09275</td>
<td>RNase_HI_RT_DIRS1</td>
</tr>
<tr>
<td>PF12017</td>
<td>Transposase protein</td>
</tr>
<tr>
<td>PF14787</td>
<td>GAG-polyprotein viral zinc-finger</td>
</tr>
<tr>
<td>cd09276</td>
<td>Rnase_HI_RT_non_LTR</td>
</tr>
<tr>
<td>PF01359</td>
<td>Transposase (partial DDE domain)</td>
</tr>
<tr>
<td>IPR000477</td>
<td>Reverse transcriptase domain</td>
</tr>
<tr>
<td>IPR012337</td>
<td>Ribonuclease H-like superfamily</td>
</tr>
<tr>
<td>IPR036397</td>
<td>Ribonuclease H superfamily</td>
</tr>
<tr>
<td>IPR041577</td>
<td>Reverse transcriptase/retrotransposon-derived protein, RNase H-like domain</td>
</tr>
<tr>
<td>IPR008042</td>
<td>Retrotransposon, Pao</td>
</tr>
<tr>
<td>IPR041373</td>
<td>Reverse transcriptase, RNase H-like domain</td>
</tr>
<tr>
<td>IPR022048</td>
<td>Envelope fusion protein-like</td>
</tr>
<tr>
<td>IPR001995</td>
<td>Peptidase A2A, retrovirus, catalytic</td>
</tr>
<tr>
<td>IPR013103</td>
<td>Reverse transcriptase, RNA-dependent DNA polymerase</td>
</tr>
<tr>
<td>IPR005162</td>
<td>Retrotransposon gag domain</td>
</tr>
<tr>
<td>IPR038717</td>
<td>Tc1-like transposase, DDE domain</td>
</tr>
<tr>
<td>IPR027806</td>
<td>Harbinger transposase-derived nuclease domain</td>
</tr>
<tr>
<td>IPR002492</td>
<td>Transposase, Tc1-like</td>
</tr>
<tr>
<td>IPR002156</td>
<td>Ribonuclease H domain</td>
</tr>
<tr>
<td>IPR025724</td>
<td>GAG-pre-integrase domain</td>
</tr>
<tr>
<td>IPR025476</td>
<td>Helitron helicase-like domain</td>
</tr>
<tr>
<td>IPR008906</td>
<td>HAT, C-terminal dimerisation domain</td>
</tr>
<tr>
<td>IPR021896</td>
<td>Transposase protein</td>
</tr>
<tr>
<td>IPR004211</td>
<td>Recombination endonuclease VII</td>
</tr>
<tr>
<td>IPR010998</td>
<td>Integrase/recombinase, N-terminal</td>
</tr>
<tr>
<td>IPR041426</td>
<td>Mos1 transposase, HTH domain</td>
</tr>
<tr>
<td>IPR034132</td>
<td>Retropepsin Saci-like domain</td>
</tr>
<tr>
<td>IPR026103</td>
<td>Harbinger transposase-derived nuclease, animal</td>
</tr>
<tr>
<td>IPR024445</td>
<td>ISXO2-like transposase domain</td>
</tr>
<tr>
<td>IPR029526</td>
<td>PiggyBac transposable element-derived protein</td>
</tr>
<tr>
<td>IPR018289</td>
<td>MULE transposase domain</td>
</tr>
<tr>
<td>IPR029472</td>
<td>Retrotransposon Copia-like, N-terminal</td>
</tr>
<tr>
<td>IPR025898</td>
<td>Tc3 transposase, DNA binding domain</td>
</tr>
</tbody>
</table>
Table S4: QUAST based statistics calculated for the Cobs2.1 genome assembly.

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Cobs.alpha.2.1</th>
</tr>
</thead>
<tbody>
<tr>
<td># contigs</td>
<td>127</td>
</tr>
<tr>
<td># contigs (&gt;= 1000 bp)</td>
<td>127</td>
</tr>
<tr>
<td># contigs (&gt;= 5000 bp)</td>
<td>125</td>
</tr>
<tr>
<td># contigs (&gt;= 10000 bp)</td>
<td>121</td>
</tr>
<tr>
<td># contigs (&gt;= 25000 bp)</td>
<td>105</td>
</tr>
<tr>
<td># contigs (&gt;= 50000 bp)</td>
<td>91</td>
</tr>
<tr>
<td>Total length (&gt;= 0 bp)</td>
<td>193051228</td>
</tr>
<tr>
<td>Total length (&gt;= 1000 bp)</td>
<td>193051228</td>
</tr>
<tr>
<td>Total length (&gt;= 5000 bp)</td>
<td>193047645</td>
</tr>
<tr>
<td>Total length (&gt;= 10000 bp)</td>
<td>193025568</td>
</tr>
<tr>
<td>Total length (&gt;= 25000 bp)</td>
<td>192755434</td>
</tr>
<tr>
<td>Total length (&gt;= 50000 bp)</td>
<td>192237042</td>
</tr>
<tr>
<td>Largest contig</td>
<td>13148674</td>
</tr>
<tr>
<td>Total length</td>
<td>193051228</td>
</tr>
<tr>
<td>GC (%)</td>
<td>41.02</td>
</tr>
<tr>
<td>N50</td>
<td>6290588</td>
</tr>
<tr>
<td>N75</td>
<td>4487289</td>
</tr>
<tr>
<td>L50</td>
<td>11</td>
</tr>
<tr>
<td>L75</td>
<td>21</td>
</tr>
<tr>
<td># N's per 100 kbp</td>
<td>94.76</td>
</tr>
</tbody>
</table>
Table S5: Location and length of each of the 34 TE islands identified in the Cobs2.1 genome assembly.

<table>
<thead>
<tr>
<th>Scaffold</th>
<th>Start</th>
<th>End</th>
<th>Length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6250000</td>
<td>8300000</td>
<td>2050000</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1500000</td>
<td>1500000</td>
</tr>
<tr>
<td>2</td>
<td>1200000</td>
<td>12360777</td>
<td>360777</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>500000</td>
<td>500000</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1000000</td>
<td>1000000</td>
</tr>
<tr>
<td>4</td>
<td>8400000</td>
<td>9000000</td>
<td>600000</td>
</tr>
<tr>
<td>5</td>
<td>7500000</td>
<td>8562622</td>
<td>1062622</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>900000</td>
<td>900000</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>800000</td>
<td>800000</td>
</tr>
<tr>
<td>8</td>
<td>6700000</td>
<td>7505125</td>
<td>805125</td>
</tr>
<tr>
<td>9</td>
<td>5400000</td>
<td>6290588</td>
<td>890588</td>
</tr>
<tr>
<td>10</td>
<td>5700000</td>
<td>6032196</td>
<td>332196</td>
</tr>
<tr>
<td>12</td>
<td>5200000</td>
<td>5755492</td>
<td>555492</td>
</tr>
<tr>
<td>13</td>
<td>7000000</td>
<td>7284943</td>
<td>284943</td>
</tr>
<tr>
<td>14</td>
<td>4400000</td>
<td>5477151</td>
<td>1077151</td>
</tr>
<tr>
<td>15</td>
<td>3650000</td>
<td>5070358</td>
<td>1420358</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>1000000</td>
<td>1000000</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>1200000</td>
<td>1200000</td>
</tr>
<tr>
<td>17</td>
<td>4400000</td>
<td>4616616</td>
<td>216616</td>
</tr>
<tr>
<td>18</td>
<td>4370000</td>
<td>4606763</td>
<td>236763</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
<td>1000000</td>
<td>1000000</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>1100000</td>
<td>1100000</td>
</tr>
<tr>
<td>20</td>
<td>4400000</td>
<td>4487289</td>
<td>87289</td>
</tr>
<tr>
<td>21</td>
<td>3700000</td>
<td>4384715</td>
<td>684715</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>300000</td>
<td>300000</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
<td>500000</td>
<td>500000</td>
</tr>
<tr>
<td>22</td>
<td>4200000</td>
<td>4241709</td>
<td>41709</td>
</tr>
<tr>
<td>23</td>
<td>3700000</td>
<td>4200000</td>
<td>500000</td>
</tr>
<tr>
<td>24</td>
<td>3200000</td>
<td>3831917</td>
<td>631917</td>
</tr>
<tr>
<td>25</td>
<td>5800000</td>
<td>6706168</td>
<td>906168</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
<td>500000</td>
<td>500000</td>
</tr>
<tr>
<td>26</td>
<td>3200000</td>
<td>3629807</td>
<td>429807</td>
</tr>
<tr>
<td>27</td>
<td>2600000</td>
<td>3200000</td>
<td>600000</td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>500000</td>
<td>500000</td>
</tr>
</tbody>
</table>
Table S6: Results of Gene Ontology Enrichment analyses with topGO of genes contained in TE islands in the Cobs2.1 genome assembly.

<table>
<thead>
<tr>
<th>GO.ID</th>
<th>Term</th>
<th>Annotated</th>
<th>Significant</th>
<th>Expected</th>
<th>parentChild</th>
<th>Ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0004984</td>
<td>olfactory receptor activity</td>
<td>232</td>
<td>30</td>
<td>16,26</td>
<td>0,00012</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0050660</td>
<td>flavin adenine dinucleotide binding</td>
<td>55</td>
<td>10</td>
<td>3,85</td>
<td>0,00056</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0005549</td>
<td>odorant binding</td>
<td>245</td>
<td>30</td>
<td>17,17</td>
<td>0,00261</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0003676</td>
<td>nucleic acid binding</td>
<td>982</td>
<td>99</td>
<td>68,82</td>
<td>0,00270</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0031177</td>
<td>phosphopantetheine binding</td>
<td>7</td>
<td>3</td>
<td>0,49</td>
<td>0,00647</td>
<td>MF</td>
</tr>
<tr>
<td>GO:1901363</td>
<td>heterocyclic compound binding</td>
<td>1852</td>
<td>155</td>
<td>129,79</td>
<td>0,00669</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0008173</td>
<td>RNA methyltransferase activity</td>
<td>19</td>
<td>5</td>
<td>1,33</td>
<td>0,00689</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0097159</td>
<td>organic cyclic compound binding</td>
<td>1855</td>
<td>155</td>
<td>130,00</td>
<td>0,00721</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0033218</td>
<td>amide binding</td>
<td>18</td>
<td>5</td>
<td>1,26</td>
<td>0,00753</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0004312</td>
<td>fatty acid synthase activity</td>
<td>7</td>
<td>3</td>
<td>0,49</td>
<td>0,01416</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0048037</td>
<td>cofactor binding</td>
<td>267</td>
<td>29</td>
<td>18,71</td>
<td>0,01622</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0043169</td>
<td>cation binding</td>
<td>710</td>
<td>64</td>
<td>49,76</td>
<td>0,01830</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0016491</td>
<td>oxidoreductase activity</td>
<td>386</td>
<td>36</td>
<td>27,05</td>
<td>0,01884</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0072341</td>
<td>modified amino acid binding</td>
<td>10</td>
<td>3</td>
<td>0,70</td>
<td>0,03081</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0008442</td>
<td>3-hydroxyisobutyrate dehydrogenase activity</td>
<td>1</td>
<td>1</td>
<td>0,07</td>
<td>0,03226</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0016614</td>
<td>oxidoreductase activity, acting on CH-OH...</td>
<td>51</td>
<td>9</td>
<td>3,57</td>
<td>0,03327</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0016830</td>
<td>carbon-carbon lyase activity</td>
<td>22</td>
<td>3</td>
<td>1,54</td>
<td>0,03526</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0000062</td>
<td>fatty-acyl-CoA binding</td>
<td>6</td>
<td>2</td>
<td>0,42</td>
<td>0,04163</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0004516</td>
<td>nicotinate phosphoribosyltransferase act...</td>
<td>1</td>
<td>1</td>
<td>0,07</td>
<td>0,04762</td>
<td>MF</td>
</tr>
<tr>
<td>GO:0019725</td>
<td>cellular homeostasis</td>
<td>26</td>
<td>5</td>
<td>1,57</td>
<td>0,00710</td>
<td>BP</td>
</tr>
<tr>
<td>GO:0030008</td>
<td>system process</td>
<td>279</td>
<td>38</td>
<td>16,87</td>
<td>0,00940</td>
<td>BP</td>
</tr>
<tr>
<td>GO:0045454</td>
<td>cell redox homeostasis</td>
<td>15</td>
<td>3</td>
<td>0,91</td>
<td>0,01680</td>
<td>BP</td>
</tr>
<tr>
<td>GO:0000393</td>
<td>spliceosomal conformational changes to g...</td>
<td>1</td>
<td>1</td>
<td>0,06</td>
<td>0,02940</td>
<td>BP</td>
</tr>
<tr>
<td>GO:007062</td>
<td>sister chromatid cohesion</td>
<td>3</td>
<td>1</td>
<td>0,18</td>
<td>0,03900</td>
<td>BP</td>
</tr>
<tr>
<td>GO:0006396</td>
<td>RNA processing</td>
<td>171</td>
<td>12</td>
<td>10,34</td>
<td>0,04370</td>
<td>BP</td>
</tr>
<tr>
<td>GO:0042592</td>
<td>homeostatic process</td>
<td>37</td>
<td>5</td>
<td>2,24</td>
<td>0,04510</td>
<td>BP</td>
</tr>
<tr>
<td>GO:0016042</td>
<td>lipid catabolic process</td>
<td>10</td>
<td>2</td>
<td>0,60</td>
<td>0,04690</td>
<td>BP</td>
</tr>
<tr>
<td>GO:0010256</td>
<td>endomembrane system organization</td>
<td>10</td>
<td>2</td>
<td>0,60</td>
<td>0,04700</td>
<td>BP</td>
</tr>
<tr>
<td>GO:0044423</td>
<td>virion part</td>
<td>3</td>
<td>3</td>
<td>0,17</td>
<td>0,00018</td>
<td>CC</td>
</tr>
<tr>
<td>GO:0019012</td>
<td>virion</td>
<td>3</td>
<td>3</td>
<td>0,17</td>
<td>0,00018</td>
<td>CC</td>
</tr>
<tr>
<td>GO:0016020</td>
<td>membrane</td>
<td>1254</td>
<td>82</td>
<td>71,18</td>
<td>0,00760</td>
<td>CC</td>
</tr>
<tr>
<td>GO:0072546</td>
<td>ER membrane protein complex</td>
<td>1</td>
<td>1</td>
<td>0,06</td>
<td>0,04167</td>
<td>CC</td>
</tr>
</tbody>
</table>
Table S7: Most abundant repeat families identified in populations of *C. obscurior* from Tenerife and Itabuna, using dnaPipeTE with 0.1x coverage per samples.

<table>
<thead>
<tr>
<th>Family</th>
<th>Tenerife (%)</th>
<th>Itabuna (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple_repeat</td>
<td>8.76</td>
<td>7.33</td>
</tr>
<tr>
<td>Unclassified</td>
<td>8.1</td>
<td>10.99</td>
</tr>
<tr>
<td>Gypsy</td>
<td>2.01</td>
<td>1.39</td>
</tr>
<tr>
<td>R1</td>
<td>0.79</td>
<td>0.83</td>
</tr>
<tr>
<td>R2</td>
<td>0.03</td>
<td>0.12</td>
</tr>
<tr>
<td>Helitron</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>TcMar-Tc</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td>Low_complexity</td>
<td>0.22</td>
<td>0.04</td>
</tr>
<tr>
<td>Copia</td>
<td>0.04</td>
<td>0.1</td>
</tr>
<tr>
<td>Total</td>
<td>20.12</td>
<td>21</td>
</tr>
</tbody>
</table>
**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.

<table>
<thead>
<tr>
<th></th>
<th>Low frequency</th>
<th></th>
<th>Common</th>
<th></th>
<th>Fixed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TE islands</td>
<td>LDRs</td>
<td>TE islands</td>
<td>LDRs</td>
<td>TE islands</td>
<td>LDRs</td>
</tr>
<tr>
<td>Itabuna</td>
<td>182</td>
<td>2587</td>
<td>1091</td>
<td>4236</td>
<td>1218</td>
<td>66</td>
</tr>
<tr>
<td>Tenerife</td>
<td>188</td>
<td>2350</td>
<td>999</td>
<td>2506</td>
<td>914</td>
<td>31</td>
</tr>
</tbody>
</table>
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.
**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.
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.
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.
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.
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.
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 $D \approx 0$) or under balancing selection (Tajima’s $D > 0$) is much higher in Itabuna than in Tenerife (Fisher’s exact test, $p < 2.2e^{-16}$).
**Figure S9:** Genome-wide distribution of genetic diversity ($\pi$) 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.
**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.
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.
**Figure S12**: Genetic diversity and differentiation in introgressed regions compared to the remainder of LDRs. (A) Genetic differentiation (B) nucleotide diversity ($\pi$) and (C) Tajima’s $D$ in two populations of *C. obscurior*, from Tenerife and Itabuna. *p<0.05, ***p < 0.0001.
Figure S13: Genetic diversity and differentiation within LDRs and TE islands in each population after excluding introgressed regions. (A) Genetic differentiation, (B) nucleotide diversity (\(\pi\)) and (D) Tajima’s \(D\) in two populations of \textit{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 \(\pi\) \((\chi^2 = 599.79, df = 3, p < 0.0001)\) and Tajima’s \(D\) \((\chi^2 = 450.72, df = 3, p < 0.0001)\), followed by pairwise Wilcoxon rank sum \textit{post hoc} tests. Different letters represent significant differences \((p < 0.001)\).
**Figure S14**: Association plots showing the signed contribution to Pearson’s $\chi^2$ (left panels) and the Pearson’s $\chi^2$ standardized residuals (right panels) calculated for low frequency ($f < 0.25$), common ($0.25 \leq f \leq 0.95$) 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.
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.