

This is a repository copy of *Patterns of transmission and horizontal gene transfer in the Dioscorea sansibarensis leaf symbiosis revealed by whole-genome sequencing*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/173164/>

Version: Accepted Version

---

**Article:**

Danneels, Bram, Viruel, Juan, Mcgrath, Krista et al. (4 more authors) (2021) Patterns of transmission and horizontal gene transfer in the *Dioscorea sansibarensis* leaf symbiosis revealed by whole-genome sequencing. *Current Biology*. pp. 2666-2673. ISSN 0960-9822

<https://doi.org/10.1016/j.cub.2021.03.049>

---

**Reuse**

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

1 **TITLE: Patterns of transmission and horizontal gene transfer in the Zanzibar yam leaf**  
2 **symbiosis revealed by whole genome sequencing**

3 **Authors: Bram Danneels<sup>1</sup>, Juan Viruel<sup>2</sup>, Krista Mcgrath<sup>3</sup>, Steven B. Janssens<sup>4,5</sup>, Nathan**  
4 **Wales<sup>6</sup>, Paul Wilkin<sup>2</sup> & Aurélien Carlier<sup>1,7,\*</sup>**

5

6 **Author affiliations:**

7 <sup>1</sup> Laboratory of Microbiology, Ghent University, 9000 Ghent, Belgium

8 <sup>2</sup> Royal Botanical Gardens Kew, Richmond, London, TW9 3AE, United Kingdom

9 <sup>3</sup> Department of Prehistory and Institute of Environmental Science and Technology (ICTA),  
10 University of Barcelona, 08193 Bellaterra, Spain

11 <sup>4</sup> Meise Botanic Garden, 1860 Meise, Belgium

12 <sup>5</sup> Plant Conservation and Population Biology, KULeuven, 3000 Leuven, Belgium

13 <sup>6</sup> Department of Archaeology, University of York, Heslington, York, YO10 5DD, United  
14 Kingdom

15 <sup>7</sup> LIPM, Université de Toulouse, INRAE, CNRS, 31320 Castanet-Tolosan, France

16 \*Correspondence and lead contact: [aurelien.carlier@inrae.fr](mailto:aurelien.carlier@inrae.fr)

17 Twitter handle: @AurelienCarlier

## 18 **Summary**

19 Leaves of the wild yam species *Dioscorea sansibarensis* display prominent acumens or “drip-  
20 tips” filled with extracellular bacteria of the species *Orrella dioscoreae*<sup>1</sup>. This species of yam  
21 is native to Madagascar and tropical Africa, and reproduces mainly asexually through aerial  
22 bulbils and underground tubers, which also contain a small population of *O. dioscoreae*<sup>2,3</sup>.  
23 Despite apparent vertical transmission, the genome of *O. dioscoreae* does not show any of  
24 the hallmarks of genome erosion often found in hereditary symbionts (e.g. small genome  
25 size and accumulation of pseudogenes<sup>4–6</sup>). We investigated here the range and distribution  
26 of leaf symbiosis between *D. sansibarensis* and *O. dioscoreae* using preserved leaf samples  
27 from herbarium collections that were originally collected from various locations in Africa.  
28 We recovered DNA from the extracellular symbiont in all samples, showing that the  
29 symbiosis is widespread throughout continental Africa and Madagascar. Despite the  
30 degraded nature of this DNA, we constructed 17 *de novo* symbiont genomes from short DNA  
31 fragments, without having to rely on reference sequences. Phylogenetic and genomic  
32 analyses revealed that horizontal transmission of symbionts and horizontal gene transfer  
33 shape the evolution of the symbiont. These mechanisms could help explain why the  
34 symbiont genomes do not display clear signs of reductive genome evolution despite an  
35 obligate host-associated lifestyle. Furthermore, phylogenetic analysis of *D. sansibarensis*  
36 plastid genomes revealed a strong geographical clustering of samples and provided evidence  
37 that the symbiosis originated at least 13 Mya, earlier than previously estimated.

38 **Keywords:** Symbiosis, herbarium, evolution, yam, plant-microbe interactions,  
39 phylogeography, bacterial genomics

## 40 **Results**

### 41 *Recovery of DNA from preserved Dioscorea sansibarensis leaf glands*

42 Leaf gland weights and yields from DNA extraction greatly varied, with an average of 1.15 µg  
43 of DNA recovered, and ranging from 1 ng up to 5.5 µg (Table S1). We did not detect any  
44 significant correlation between the size of the glands and DNA yield, even after leaving out 10  
45 specimens for which we had processed only a fragment of the gland (Spearman correlation *p*-  
46 value > 0.1). In addition, specimen age did not correlate with the amount of DNA extracted

47 (Spearman correlation  $p$ -value > 0.1), or with the number of reads from shotgun sequencing  
48 (Table S1).

#### 49 *Taxonomic composition of D. sansibarensis leaf glands*

50 On average, 60.5% of sequencing reads per sample mapped to the *O. dioscoreae* LMG 29303<sup>T</sup>  
51 reference genome, 7.92% mapped to a draft version of the *D. sansibarensis* genome, 0.64%  
52 mapped to the *D. sansibarensis* chloroplast, and 5.24% mapped to the human genome (Table  
53 S1). We could detect *O. dioscoreae*-specific markers in all samples using Metaphlan3<sup>7</sup>, except  
54 in the low-output sample MK024. In all but two samples (MK010 and MK018), the only  
55 bacterial gene marker sequences corresponded to *O. dioscoreae* (Table S2). MK010 and  
56 MK018, however, contained a large proportion of other bacterial species, mostly human  
57 commensals. This correlated with high levels of human DNA contamination (Table S1). Reads  
58 mapping to the human genome were longer on average, and did not show elevated levels of  
59 C-to-T conversions, as is typical for aDNA. This contamination most likely occurred during  
60 collection of the samples from the herbarium or processing of the DNA samples and we did  
61 not analyse these samples further. Interestingly, only 10% of the reads from sample Herb2, a  
62 herbarium specimen with characteristics that did not fully match the taxonomic morphotype  
63 (Table S3), showed significant homology to sequences in the database. Nearly 40% of the  
64 classified reads did not map to taxa related to either *D. sansibarensis* or *O. dioscoreae*.

#### 65 *DNA damage patterns vary between chloroplast and symbiont DNA but are consistent with* 66 *long-term preservation in historical specimens*

67 Assessment of DNA damage patterns in historical specimens is critical for validating their  
68 authenticity. Leaf glands of *D. sansibarensis* are populated by clonal bacteria<sup>8</sup> as well as plant  
69 cells and plastids. We observed an average read length of 52.5 bp in our historical specimens,  
70 a degree of fragmentation that is similar to previously reported herbarium DNA<sup>9–11</sup>, although  
71 reads mapping to the chloroplast tended to be larger than bacterial reads (55.10 bp vs 37.10  
72 bp, Wilcoxon paired rank sum test  $p$ -value < 0.001; Figure S1A). Read length was not  
73 significantly correlated to the age of the specimens in the chloroplast or the symbiont  
74 (Pearson correlation  $p$ -values > 0.1). Consistent with patterns typical of aDNA, the first base  
75 of sequencing reads is enriched in C-to-T mismatches in both the chloroplast and symbiont  
76 genomes (Figure 1). The absolute proportion of C-to-T mismatches showed significant

77 correlation with the age of the specimens (Pearson correlation  $p$ -values  $< 0.01$ ). Similarly,  
78 purines were enriched before strand breaks in both *O. dioscoreae* and plastid DNA, a common  
79 feature of ancient DNA, although in different relative proportions (Figure 1, Figure S1C-D). The  
80 proportion of purines before strand breaks was larger in the *O. dioscoreae* genome compared  
81 to the *D. sansibarensis* plastome. (66% vs. 17% increase, Wilcoxon signed-rank test  $p$ -value  $<$   
82 0.001) (Figure S1D).

### 83 *Herbarium specimens provide insight into the dispersal of D. sansibarensis over continental* 84 *Africa*

85 Most plastid sequences were nearly identical (168 SNPs over an alignment of 121 366 bp),  
86 resulting in a phylogenetic topology with very short branches (Figure 2A). Concordant with the  
87 haplotype network (Figure 3), specimens from Madagascar form distinct clades, with  
88 continental specimens originating from a single ancestor. In contrast, the plastid sequence of  
89 the Herb2 sample is very divergent from the rest and constitutes a basal branch in the  
90 phylogenetic tree. Samples collected in Madagascar all clustered together, and according to  
91 the sampling region, which is in accordance with what we reported previously<sup>3</sup>. Because post-  
92 mortem base transitions can affect the phylogenetic signal and introduce artefacts, we also  
93 analyzed a dataset keeping only transversions. Although less well resolved, transversion-only  
94 phylogenies and haplotype networks are consistent with the results of analyses using the full  
95 dataset (Figure S2B). Phylogenetic dating revealed that our *D. sansibarensis* specimens  
96 diverged about 13.54 million years ago (95% confidence interval: 4.93 Mya – 25.19 Mya). This  
97 high age estimate is mostly due to the very divergent nature of the Herb2 specimen. The  
98 remaining specimens share their most recent common ancestor at 3.31 million year ago (95%  
99 confidence interval: 0.63 Mya – 7.71 Mya).

### 100 *High specificity of leaf symbiosis without phylogenetic congruence*

101 The genomes of the symbionts are more diverse, with two main phylogenetic clades (clade I  
102 and clade II, Figure 2B). Samples do not cluster according to location of specimen collection,  
103 and the phylogenies of host and symbiont are not congruent (Figure S2A). To gain insight into  
104 the population structure of *O. dioscoreae*, we assembled *de novo* *O. dioscoreae* genomes from  
105 17 out of 36 herbarium specimens. Analysis using CheckM<sup>12</sup> indicated low levels of  
106 contamination ( $< 1\%$ ) in the metagenome-assembled genomes (MAGs). Most *O. dioscoreae*

107 MAGs could be reconstructed in less than 100 contigs, and were of similar sizes to the  
108 reference genome (4.7 to 5.2 Mbp) (Table S4). Whole genome alignment using Mauve<sup>13</sup>  
109 showed high synteny, without large-scale rearrangements. Average nucleotide identity (ANI)  
110 values confirmed that all symbiont MAGs belonged to the same species, with a minimum of  
111 96.02% ANI, well above the commonly accepted 95% threshold for species delineation<sup>14</sup>.  
112 Interestingly, two MAGs from specimens collected 35 years apart in different phytoregions of  
113 the DR Congo were almost identical (Herb9 and MK003, 2 SNPs out of 4 846 400 bp). Cross-  
114 sample contamination is unlikely since these samples were processed in different facilities and  
115 sequenced at a different sequencing centre. In contrast, some glands from plants collected at  
116 the same site in Madagascar contained bacteria belonging to distinct phylogenetic clusters,  
117 highlighting the distributed biogeography of *O. dioscoreae*<sup>3</sup>.

#### 118 *Comparative genomics of wild-collected and herbarium-assembled O. dioscoreae genomes*

119 The total amount of predicted genes is approximately the same in all MAGs (4300-4700, Table  
120 S4), with a core genome taking up an average of 77% of the gene inventory (3541 genes). The  
121 pan genome of *O. dioscoreae* is large given the narrow range of ANI values, consisting of 7406  
122 genes over 28 genomes. The accessory genome mostly consists of genes that are unique to  
123 one, or very few samples (30% of orthogroups only consist of three or less members, Figure  
124 S3). There is a general trend towards gene loss, with most lineages having lost on average  
125 1024 genes, while only gaining an average of 380 genes, for an average net gene loss of 644  
126 genes per lineage (Figure 4). Most frequently occurring patterns of gene loss involved long  
127 branches (e.g. in MK020), or genes that are specific to a certain (sub)clade in the phylogeny.  
128 Most genes are lost as single genes or in small clusters and correspond to hypothetical  
129 proteins, indicating that gene loss is unlikely to be adaptive. An exception is a large gene  
130 cluster that is lost in some lineages: a cluster of 34 genes related to Type III secretion. This  
131 cluster is present all MAGs of clade I, but has been lost multiple times in lineages of clade II  
132 (Figure 4). In contrast, functions highly expressed in the *D. sansibarensis* leaf gland and linked  
133 to specialized metabolism and type VI secretion are conserved in all *O. dioscoreae* MAGs<sup>3</sup>.  
134 Despite this high degree of conservation, the phylogenetic trees of several genes from one of  
135 the two Type VI secretion systems of *O. dioscoreae* are incongruent with the species tree, as  
136 reflected by significantly higher weighted Robinson-Fould distance than average (0.35 in T6SS-  
137 2 vs 0.20 in T6SS-1, Wilcoxon rank sum test  $p$ -value < 0.01). Among these, three putative VgrG-

138 domain effector proteins (ODI\_R0793, ODI\_R0797 and ODI\_R0809) were likely subjected to  
139 gene conversion or HGT. In addition, a pair of Rhs/VgrG proteins putative T6SS effector  
140 proteins was encoded in all genomes of clade II, but in none of the genomes of clade I. Apart  
141 from those, additional Rhs and/or VgrG proteins domains were also detected in 4 other MAGs  
142 (AMP9, BER1, BER2, and MK019).

## 143 **Discussion**

### 144 *Herbarium samples provide a reliable source of symbiont DNA*

145 Herbarium specimens are seeing applications going beyond taxonomic and systematic studies  
146 and are an increasingly useful resource for studies of plant biology and evolution<sup>15</sup>. Here, we  
147 leveraged herbarium specimens to gain insights into the genome evolution and transmission  
148 mode of the symbiosis between *D. sansibarensis* and *O. dioscoreae*. Preserved samples of leaf  
149 acumens mostly yielded DNA of sufficient quality for high-throughput shotgun sequencing  
150 analysis, which demonstrated the ubiquity of the association with *O. dioscoreae* in a broad  
151 cross-section of *D. sansibarensis*' range. Moreover, we did not find consistent or unambiguous  
152 evidence for the presence of other microorganisms in the leaf gland. Only one sample (Herb2)  
153 contained a large proportion of reads assigned to bacterial taxa other than *O. dioscoreae*.  
154 Specimen Herb2, collected in Cameroon, formed a divergent basal branch in the chloroplast  
155 phylogeny (Figure 2A). Interestingly this herbarium specimen did not fully fit the taxonomic  
156 morphotype of the species, and was tentatively identified as "*Dioscorea cf. sansibarensis*".  
157 These observations indicate that this specimen represents an early-diverging lineage of the  
158 species, a sub-species, or even an entirely new species. Further investigation and sampling will  
159 be necessary to confirm the exact taxonomic placement of this specimen, and link it to the  
160 evolution of *D. sansibarensis*. Nevertheless, the presence of the symbiont *O. dioscoreae* in the  
161 Herb2 specimen suggests that the symbiosis might not be confined to the *D. sansibarensis*  
162 species and is possibly established much earlier than expected. The fact that the bacterial  
163 communities in this sample were more complex could indicate that it represents an older  
164 branch of *D. sansibarensis* where strict specificity or vertical transmission has not yet evolved.

### 165 *Phylogeography of Dioscorea sansibarensis*

166 Most plastid sequences across *Dioscorea sansibarensis* representative of the distribution  
167 range were highly similar, which resulted in a phylogenetic topology containing many

168 unresolved branches. There is however a strong biogeographic separation of samples, with  
169 specimens from the same region clustering together. Continental African specimens form a  
170 clade separated from specimens from Madagascar, which is in concordance with the earlier  
171 hypothesis that *D. sansibarensis* originated in Madagascar and was dispersed to Africa<sup>16</sup>.  
172 *Dioscorea sansibarensis* appears to rely largely, or in places exclusively, on vegetative  
173 reproduction for propagation and dispersal. Despite extensive field research collecting  
174 *Dioscorea* in Africa and Madagascar, one author (PW) has never seen mature seeds or juvenile  
175 plants not arising from bulbils (axillary perennating organs) *in situ*, even in areas where it is  
176 abundant and flowers extensively such as the far North of Madagascar. Wilkin<sup>17</sup> reported that  
177 no seed bearing plants had been seen among all the herbarium specimens collected in  
178 southern Africa, although they were occasionally encountered elsewhere in Africa. This  
179 suggests that *O. dioscoreae* would be most likely to move between plants via bulbil-mediated  
180 vertical transmission<sup>3</sup>. It also suggests that patterns of genetic variation within *D.*  
181 *sansibarensis* would reflect its mode of reproduction, with low levels of within-population  
182 genetic divergence in local clones that are occasionally further dispersed. This is congruent  
183 with the plastid tree topology, and haplotype network (Figure 2B, Figure 3) with an eastern, a  
184 western and a mixed East-West Africa clade. Furthermore, there is some variation in bulbil  
185 traits, which tend to be black or purple and smooth in Africa and brown or green and warty in  
186 Madagascar.

#### 187 *Insights into the evolutionary history of O. dioscoreae from de novo assembly of genomes*

188 Despite direct evidence of vertical transmission of *O. dioscoreae* through vegetative  
189 propagules<sup>3</sup>, the phylogenetic trees of *D. sansibarensis* and *O. dioscoreae* are highly  
190 incongruent (Figure S2A). This indicates a mixed mode of transmission, where symbionts can  
191 occasionally transfer from one plant lineage to another. Horizontal transmission, for example  
192 by insect vectors, could result in phylogenetic incongruence between host and symbionts.  
193 Acquisition from an environmental reservoir seems unlikely since we could not reliably detect  
194 the symbiont anywhere outside of the plant<sup>3</sup>, but cannot be fully ruled out.

195 While bacterial genomes have been assembled from archaeological remains and herbaria  
196 accessions before<sup>18,19</sup>, these methods almost exclusively relied on either mapping reads to a  
197 reference, or on prior taxonomic classification of reads. The binary nature of the symbiosis,  
198 and genomic differences between *D. sansibarensis* and *O. dioscoreae* allowed us to construct



199 qualitative *de novo* genome sequences from herbarium material suitable for detailed  
200 comparative genomics analyses. Functions thought to play a role in the symbiotic lifestyle of  
201 *O. dioscoreae*, such as specialised metabolism and T6SS, are conserved in all samples<sup>3</sup>.  
202 Interestingly, we found differences in the complement of putative T6SS effector genes, in  
203 addition to evidence of intra-clade HGT or gene conversion. Effector repertoires define the  
204 target specificity of the T6SS, and these could play an important role in the ecology of *O.*  
205 *dioscoreae*<sup>20</sup>. For example, we found a combination of Rhs and VgrG domain-containing genes  
206 that is conserved in all genomes of clade II, but is not present in genomes of clade I. As T6SS  
207 plays important roles in microbe-microbe interactions<sup>20,21</sup>, this could indicate that effector  
208 inventories partially diverged in response to different threats from competitors or  
209 alternatively may play some role in signalling and adaptation to a new host<sup>22</sup>.

210 The *O. dioscoreae* core genome accounts for 78% of the gene complement in *O. dioscoreae*,  
211 while the pangenome is much larger, being approximately twice the size of the core genome.  
212 The membership distribution of genes of the pangenome is bimodal, with a strong bias  
213 towards genes only found in very few genomes. This either indicates that new genes can still  
214 be acquired, or more likely that genes affected by genetic drift are quickly purged<sup>23</sup>. In general,  
215 *O. dioscoreae* genomes show an overall trend toward gene loss, following a general trend in  
216 prokaryotes<sup>4,24,25</sup>. Net gene loss can lead to genome erosion, a feature commonly found in  
217 host-restricted bacteria, including leaf symbionts<sup>5,26–30</sup>. For example, a cluster of 34 genes  
218 containing most T3SS genes is conserved in clade I, but has been lost multiple times in  
219 genomes of clade II. Genes of the T3SS of *O. dioscoreae* LMG29303<sup>T</sup> were not upregulated *in*  
220 *planta*<sup>3</sup>, suggesting that loss of T3SS genes is likely due to genetic drift rather than adaptive  
221 selection<sup>31</sup>. Despite this apparent gene volatility, *O. dioscoreae* do not display the hallmarks  
222 of genome reduction, such as accumulation of pseudogenes and insertion elements, or AT-  
223 bias<sup>4,5,30</sup>. Occasional horizontal transmission of the symbiotic bacteria may slow down or  
224 alleviate entirely reductive genome evolution, and may explain these patterns.

225 In conclusion, our data demonstrate that aDNA and metagenomics methods are a powerful  
226 combination to probe dynamic associations between plants and microorganisms from  
227 herbarium samples. The discovery that symbiont switching or horizontal transfer occurs  
228 frequently between *D. sansibarensis* and *O. dioscoreae* despite up to 13 Mya of co-evolution  
229 suggests a degree of plasticity not previously seen in vertically-transmitted leaf symbioses.

230 This illustrates the potential of leaf symbioses as model systems to understand the  
231 mechanisms of host-microbe specificity in the leaf.

### 232 **Acknowledgments**

233 We would like to thank Mathijs Deprez, who helped out with herbarium DNA-extractions as  
234 a part of his work in preparation of his master dissertation. This work was supported by the  
235 Flemish Fonds Wetenschappelijk Onderzoek under grant G017717N to AC. AC also  
236 acknowledges support from the French National Research Agency under grant agreement  
237 ANR-19-TERC-0004-01 and from the French Laboratory of Excellence project "TULIP" (ANR-  
238 10-LABX-41; ANR-11-IDEX-0002-02). The funders had no role in study design, data collection  
239 and analysis, decision to publish, or preparation of the manuscript. We thank the Danish  
240 National High-throughput Sequencing Centre and the Oxford Genomics Centre at the  
241 Wellcome Centre for Human Genetics for assistance in generating and initial processing of  
242 the sequencing data.

### 243 **Author Contributions**

244 Conceptualization, B.D. and A.C.; Methodology, B.D., K.M., N.W.; Investigation, B.D. and J.V.;  
245 Resources, S.J. and A.C.; Writing – Original draft, B.D., J.V., P.W. and A.C.; Writing – Review &  
246 Editing, B.D., J.V., N.W., S.J., P.W. and A.C.; Supervision, A.C.; Funding Acquisition, A.C.

### 247 **Declaration of Interests**

248 The authors declare no conflict of interest.

249

250

251 **FIGURE LEGENDS**

252 **Figure 1: DNA damage patterns in *O. dioscoreae* and *D. sansibarensis* chloroplast.** Output  
253 graphs from MapDamage 2.0<sup>32</sup> of sample MK023, showing different DNA damage patterns.  
254 **(A-B)** Frequency of bases around read ends (grey brackets) mapped to the *Orrella dioscoreae*  
255 (A) and *Dioscorea sansibarensis* chloroplast (B) reference genomes. Numbers on the x-axis  
256 represent the relative position from the read end. The dotted lines on the chloroplast plot  
257 show the higher variability due to lower sequencing coverage **(C-D)** Frequency of  
258 mismatches along mapped reads. Numbers on the x-axis represent the position along the  
259 mapped read, lines represent the observed frequency of certain mismatches. Red: C-to-T  
260 mismatch; Blue: G-to-A mismatch; Grey: Other mismatches.

261 **Figure 2: SNP-based phylogenies of *D. sansibarensis* chloroplast (A) and *O. dioscoreae* (B).**  
262 SNP-based phylogenies based on the alignment of 121 366 nucleotides of the plastome  
263 (containing 168 variant sites) and 188 138 non-recombinant variant sites for *O. dioscoreae*.  
264 Branch support values are given as % (bootstrap). Branches with support < 50% were  
265 collapsed. Font colours correspond to where the specimens were collected originally. Plants  
266 from the botanical gardens of Meise and Ghent were originally collected in DR Congo, and  
267 are annotated as such on the tree. The chloroplast SNP-based phylogeny contains fewer  
268 samples than the symbiont phylogeny as samples with fewer than ten thousand reads  
269 mapped to the plastome were excluded from the analysis.

270 **Figure 3: *D. sansibarensis* specimen sampling locations and haplotype network of plastid**  
271 **sequences. (A)** Sampling locations of *D. sansibarensis* specimens. Samples derived from the  
272 same country share the same colour. Numbers on sample locations depict the number of  
273 specimens sampled in that region. Abbreviations: BI – Burundi; CG – Republic of Congo; GQ –  
274 Equatorial Guinea; MG – Madagascar; MZ – Mozambique; RW – Rwanda; **(B)** Haplotype  
275 network based on the chloroplast alignment used for phylogenetic analysis, constructed  
276 using the TCS algorithm<sup>33</sup>. Samples with more than 25% gaps were excluded. Colours  
277 represent geographical origin of the samples (see panel A). Circle size in the haplotype  
278 network are scaled by the number of samples of that haplotype. Ticks on connecting lines  
279 represent point mutations between nodes.

280

281 **Figure 4: Gene gains and losses in the *Orrella dioscoreae* genome.** Reconstruction of gene  
282 gain and loss based on Dollo's parsimony principle. Numbers on branches represent gained  
283 (+) and lost (-) genes. Numbers in black, bold font above branches represent the estimated  
284 size of the ancestral gene pool (left), or represent the current number of genes in a certain  
285 genome (right).

286

## 287 **SUPPLEMENTARY FIGURES AND TABLES**

288

289 **Figure S1: Differences in DNA damage parameters between *D. sansibarensis* chloroplast  
290 and *O. dioscoreae***

291

292 **Figure S2: SNP-based phylogenies and haplotype network of *D. sansibarensis* and *O.*  
293 *dioscoreae***

294

295 **Figure S3: Prevalence of orthogroups with certain number of genomes**

296

297 **Figure S4: Manual binning of metagenome contigs derived from *O. dioscoreae***

298

299 **Table S1: DNA and sequencing yields, and mapping results of herbarium specimens to *D.*  
300 *sansibarensis* nuclear and plastid genome, *O. dioscoreae* genome, and the human  
301 reference genome, related to figure 3A and STAR Methods**

302

303 **Table S2: Presence of bacterial markers in the trimmed sequencing reads**

304

305 **Table S3: Herbarium specimen metadata, as recorded in the archives of the Meise Botanic  
306 Garden herbarium collection, related to STAR Methods**

307

308 **Table S4: Genome statistics of *O. dioscoreae* genomes**

309

310



312 **STAR Methods**

313 **KEY RESOURCE TABLE**

314 **RESOURCE AVAILABILITY**

315 *Lead contact*

316 Further information and requests for resources and reagents should be directed to and will  
317 be fulfilled by the Lead Contact, Aurélien Carlier ([aurelien.carlier@inrae.fr](mailto:aurelien.carlier@inrae.fr)).

318 *Materials Availability*

319 This study did not generate new unique reagents.

320 *Data and code availability*

321 Sequencing reads generated from herbarium specimens of *Dioscorea sansibarensis* are  
322 deposited under SRA project PRJNA646369. Generated metagenome-associated genomes of  
323 *Orrella dioscoreae* are deposited in the Zenodo repository, accession 3946545. The draft  
324 genome of *D. sansibarensis* used in this paper is not yet published, but can be requested by  
325 contacting the lead author. Scripts used for analysis in this paper are available at  
326 <https://github.ugent.be/brdanee/DioscoreaHerbarium>.

327 **EXPERIMENTAL MODEL AND SUBJECT DETAILS**

328 ***Plants***

329 Leaf nodules of herbarium specimens of *Dioscorea sansibarensis* were provided by the Meise  
330 Botanic Garden. Glands from wild *D. sansibarensis* specimens from Madagascar were  
331 collected under research permit 158/16/MEEF/SG/DGF/DSAP/SCB.Re issued by the Ministry  
332 of Environment, Ecology and Forests of the Republic of Madagascar. Collection information  
333 for all samples is available in Table S3.

334 **METHOD DETAILS**

335 *Sampling and DNA-extraction*

336 Leaf glands of 36 herbarium specimens (Table S3) of the Meise Botanic Garden herbarium  
337 (Belgium) were dissected and tissues were stored at 4°C with silica until further processing.

338 Total DNA-isolation and genomic library preparation of ten specimens (Herb1-Herb10),  
339 representing various geographic locations, different ages, and diverse gland sizes were  
340 performed in the palaeogenomics facility at the Department of Archaeology of the University  
341 of York (UK). Twenty-six specimens (MK001-MK026) were processed at the department of  
342 Microbiology of Ghent University in a room disinfected with bleach and under a disinfected  
343 PCR cabinet (AirClean 600 PCR Workstation, Starlab, Hamburg, Germany). All utensils were  
344 disinfected using bleach and/or followed a UV treatment prior to their usage. When possible,  
345 sterile disposable items were used. Extraction blanks were included to monitor possible DNA-  
346 contamination. Total genomic DNA from leaf nodules was extracted using the protocol  
347 described in Gilbert *et al.*<sup>34</sup>, which was found to perform well on botanical specimens<sup>35,36</sup>. The  
348 leaf glands were cut into small pieces using a sterile scalpel and placed into sterile 2ml  
349 Eppendorf Lo-bind microfuge tubes. The samples were incubated on a shaker overnight at  
350 55°C in 1200 µl of extraction buffer (10mM Tris-HCl pH 8, 10mM NaCl, 2% SDS, 5mM CaCl<sub>2</sub>,  
351 2.5mM EDTA pH 8, 0.5mg/ml Proteinase K, and 40mM DTT). Supernatants were extracted  
352 twice with an equal volume of 25:24:1 phenol/chloroform/isoamylalcohol. The resulting DNA  
353 was diluted in 13x binding buffer (5M guanidine hydrochloride, 40% isopropanol, 0.05%  
354 Tween-20, and 90mM Sodium Acetate pH 5.2)<sup>37</sup> and purified using a MinElute PCR purification  
355 kit (Qiagen, Hilden, Germany) following the manufacturer's recommendation. Extractions  
356 blanks yielded no detectable amounts of DNA (Table S1)

#### 357 *Library preparation and sequencing*

358 Genomic libraries adapted for ancient DNA were constructed for all samples, the extraction  
359 blanks, and library blanks containing molecular biology-grade water, following the double-  
360 stranded protocol from Wales *et al.*<sup>38</sup>, and using the adapters described in Meyer & Kircher<sup>39</sup>.  
361 DNA fragment ends were repaired using the NEBNext End Repair module (New England  
362 BioLabs, Ipswich, MA, USA), and purified on MinElute (Qiagen, Hilden, Germany) columns,  
363 followed by adapter ligation using the NEBNext Quick Ligation module (New England BioLabs,  
364 Ipswich, MA, USA) and purification using QiaQuick (Qiagen, Hilden, Germany) columns. Gaps  
365 were filled using *Bst* DNA polymerase (New England Biolabs, Ipswich, MA, USA). PCR-amplified  
366 DNA libraries were quantified using either a Quantus (Promega, Madison, WI, USA) or Qubit  
367 (Invitrogen, Carlsbad, CA, USA) fluorometer with respective dsDNA kits. Libraries were pooled  
368 in equimolar concentrations and sequenced at the National High-throughput DNA Sequencing

369 Centre, Copenhagen, Denmark (Illumina HiSeq 4000, samples Herb1 to 10) or at the Wellcome  
370 Trust Human Center for Human Genetics, Oxford, UK (Illumina NovaSeq 6000, samples  
371 MK001-MK026), yielding single-ended 80 bp reads. Library blanks containing molecular grade  
372 water showed no library amplification and yielded only a few thousands of reads (Table S1).  
373 Samples with sequencing output below their respective extraction blanks (Herb10, MK001,  
374 and MK022) were not used in further analysis. Raw sequencing reads were deposited in the  
375 SRA archive under bioproject PRJNA646369.

#### 376 *Read processing and mapping*

377 Sequencing adapters were removed using Cutadapt v2.10<sup>40</sup>, and low-quality bases were  
378 removed using Trimmomatic v0.39<sup>41</sup>. Trimmed reads were mapped to the *Dioscorea*  
379 *sansibarensis* chloroplast sequence (NCBI accession GCA\_900631875.1) and a draft version of  
380 the nuclear genome (unpublished) from a plant obtained from the Botanical Garden of Ghent  
381 University, to the associated *Orrella dioscoreae* LMG 29303<sup>T</sup> genome (NCBI accession  
382 GCA\_900089455.2), and to the human reference genome GRCh38 (NCBI accession  
383 GCF\_000001405.39) using BWA aln v0.7.17, with seeding disabled<sup>42</sup>. Coverage for every  
384 genome was calculated using BEDTools v2.27.1 *genomecov* command<sup>43</sup>. Taxonomic  
385 composition of samples was determined using Metaphlan 3<sup>7</sup>, Kraken v1.1.1<sup>44</sup> (using a custom  
386 database of bacterial and chloroplast sequences<sup>8</sup>), and blastn<sup>45</sup> searches against the NCBI non-  
387 redundant nucleotide database (accessed 03/2020), summarized using BASTA and Krona<sup>46,47</sup>.

#### 388 *DNA damage analysis*

389 Mapping files created by BWA aln against the *O. dioscoreae* LMG 29303<sup>T</sup> genome, the *D.*  
390 plastome, and the human HRCh38 reference genome were dereplicated using Samtools  
391 MarkDup<sup>48</sup>, and used as input for MapDamage 2<sup>32</sup> to rescale quality scores of likely damaged  
392 bases, and estimate DNA damage patterns. To increase the reliability of the DNA damage  
393 analysis, only samples with average coverage above 5x for both plastid and symbiont genomes  
394 were considered for statistical analysis (17 samples in total). DNA damage was assessed by  
395 measuring: (i) the absolute number of C-to-T mismatches on the first base of the reads; (ii) the  
396 relative amount of C-to-T mismatches (calculated by subtracting the background amount of  
397 C-to-T mismatches, estimated as the average C-to-T mismatches on bases 10 to 20, from the  
398 C-to-T mismatches on the first base); (iii) the relative increase of purine bases before strand



399 breaks, calculated by dividing the proportions of purine bases in the reference genome at  
400 position -1 and at position -5 relative to the start of the mapped read.

#### 401 *Genome assembly*

402 *De novo* assembly of bacterial genomes was performed in 2 steps using SPAdes v3.14<sup>49</sup> as  
403 described previously<sup>3</sup>. First, a low-stringency assembly was done in unpaired mode using *k*-  
404 mer sizes of 21, 25, 33, 37, and 45. Bacterial contigs were visually binned according to base  
405 composition (% G+C) and average coverage (Figure S4), as *O. dioscoreae* has significantly  
406 higher G+C content than the plant, and is present in high numbers in leaf glands. Reads  
407 mapping to the selected contigs (mapped using SMALT<sup>50</sup> and extracted using Samtools<sup>48</sup>) were  
408 reassembled using SPAdes v3.14 in *careful* mode, using *k*-mer sizes of 21, 27, 33, and 41. The  
409 final assemblies were filtered to remove contamination, by removing contigs assigned as  
410 eukaryotic or with discordant taxonomic assignment by Kraken v1.1.1<sup>44</sup> and blastn searches  
411 to the NCBI nr database, analysed and visualised using BASTA<sup>46</sup> and Krona<sup>47</sup> as described  
412 previously<sup>3</sup>. Assembly quality, completeness and contamination were determined using  
413 Quast<sup>51</sup>, BUSCO v4.0.6<sup>52</sup> and CheckM v1.1.2<sup>12</sup> respectively. Variants between the aDNA reads  
414 and the assembled genomes were called using bcftools v1.11<sup>53</sup>, to assess overrepresentation  
415 of problematic transition mutations (C-to-T or G-to-A) in the genomes. The herbarium  
416 metagenome-assembled genomes are available on the Zenodo repository (DOI:  
417 10.5281/zenodo.3946545)

#### 418 *Phylogenetic analyses*

419 SNP-based phylogenies of both plastid and symbiont genome of the herbarium specimens,  
420 previously sequenced fresh leaf glands from Madagascar<sup>3</sup>, and a specimen collected from the  
421 living collection of the Meise Botanic Garden, Belgium (accession CD-0-BR-1960001),  
422 containing *O. dioscoreae* strain R-67584, were constructed using Realphy v.122<sup>54</sup>. Sequencing  
423 reads with quality scores rescaled to account for DNA damage (see above) were used as input  
424 for the Realphy pipeline, allowing at most 10% of disagreement on mapped bases, allowing  
425 gaps in at most 20% of the samples, and requiring a minimum coverage of 5 reads for bacterial  
426 alignments and 3 for the chloroplast. The aforementioned *Dioscorea sansibarensis* plastome  
427 and the *Orrella dioscoreae* LMG 29303<sup>T</sup> genome were used as references for mapping.  
428 Samples with fewer than 10 000 mapped reads were discarded from further analysis.

429 Phylogenetic trees based on plastid data were constructed using PhyML v3.3.3<sup>55</sup> using the F81  
430 model (selected using the CLC Main Workbench (Qiagen) model testing tool), 100 bootstrap  
431 replicates and the plastid sequence of *Dioscorea elephantipes* (NCBI accession NC\_009601) as  
432 outgroup. For the *O. dioscoreae* phylogeny, Gubbins<sup>56</sup> was used to remove regions with  
433 elevated rates of base substitutions from the alignment, to mitigate the effect of  
434 recombination on whole genome or SNP-based phylogenies. A maximum-likelihood  
435 phylogeny was created using RAxML v8.2.12<sup>57</sup> (rapid bootstrapping and best-scoring ML  
436 mode, using 100 bootstrap replicates and the GTRGAMMA substitution model). The genome  
437 sequence of *Achromobacter xylosoxidans* ADAF13 (NCBI accession GCA\_001566985) was used  
438 as outgroup to root the phylogeny. A haplotype network of the plastid sequences was created  
439 with the TCS algorithm, implemented in POPART<sup>33,58</sup>, using the SNP-based alignment used for  
440 the phylogeny as input. Samples with more than 25% of gaps in their sequence were removed  
441 from downstream analyses. To further control for the effect of aDNA damage (mainly  
442 transitions) on the low-coverage plastome analysis, the phylogenetic analysis and haplotype  
443 network were recalculated using only transversions (89 transversions out of 168 SNPs). In an  
444 effort to assess the phylogeny of *D. sansibarensis* based on its nuclear genome, the sequence  
445 of the high-copy plant marker ITS (internal transcribed spacer) was extracted by searching  
446 metagenome sequences with blastn<sup>45</sup>. As a query, a known 18S-5S-28S region of *D.*  
447 *sansibarensis* (NCBI accession DQ267929.1) was used. The matching sequences were  
448 extracted from the metagenome contigs, including 250 extra flanking base pairs. Using this  
449 method, 20 herbarium ITS sequences could be extracted. Sequences were aligned using  
450 Muscle<sup>59</sup> and the alignment was trimmed using trimAL<sup>60</sup> to remove columns with >80% gaps.  
451 The resulting alignment resulted in no variable sites between samples

452 Age estimation of the common ancestor of all investigated specimens was performed using  
453 BEAST v1.10.4<sup>61</sup> based on Viruel *et al.*<sup>16</sup>, and as described previously<sup>3</sup>. Gene alignments for  
454 three chloroplast genes (*matK*, *rbcl*, *atpB*) were constructed, as high-quality sequences for  
455 these genes are available for many *Dioscorea* species<sup>16</sup>. the *trnLF* (trnL intron–trnL exon–  
456 trnL/trnF spacer) region was not used, as this region could not be reliably extracted from the  
457 herbarium plastid sequences. Markers of *Dioscorea* species described in Viruel *et al.*<sup>16</sup>, three  
458 herbarium specimens with enough coverage and representing most variety in the SNP-based  
459 phylogeny (MK014, MK017, MK023), and the chloroplast sequences obtained from a

460 specimen kept in the botanical garden of Ghent University were used to construct the  
461 phylogeny. The same parameters (uncorrelated relaxed molecular clock and Yule speciation  
462 model) and calibration points as described in Viruel *et al.*<sup>16</sup> were used to run the dating  
463 analysis: two calibrations using a normal distribution were used for Dioscoreaceae (*Dioscorea*,  
464 *Tacca*, *Trichopus* and *Stenomeris*; mean = 108.0, stdev = 10.0) and for Burmanniaceae and  
465 Dioscoreaceae node (mean = 115.0, stdev = 4.0); and two calibrations using a lognormal  
466 distribution for *D. cochleari-apiculata*, *D. dregeana* and *D. dumetorum* clade (mean = 27.32,  
467 stdev = 1.0), and *Dioscorea* clade (mean = 48.2, stdev = 1.0).

#### 468 *Comparative genomics of O. dioscoreae genomes*

469 Average Nucleotide Identity (ANI) values between available *O. dioscoreae* genomes were  
470 calculated using PyANI v0.3<sup>62</sup>. Orthologs between herbarium genomes, genomes assembled  
471 from fresh glands<sup>3</sup>, and the R-67584 strain isolated from a *D. sansibarensis* specimen from the  
472 Botanic Garden of Meise, were predicted using Orthofinder v2.3.9<sup>63</sup>. Patterns of gene gain and  
473 loss were computed based on the gene presence/absence output of Orthofinder, using the  
474 Dollo analysis implement in Count<sup>64</sup>. The gains and losses were mapped on a pruned  
475 phylogeny created using the ete3 python package<sup>65</sup>, only including non-redundant genomes  
476 (<99% identical). Weighted Robinson-Fould distances were calculated using the DendroPy  
477 python package<sup>66</sup>, comparing the gene and species tree as reported by Orthofinder. To assess  
478 if assembly errors could affect the detection of gene losses, reads of herbarium specimens  
479 were mapped to the closest reliable fresh-specimen genome, and compared the proportions  
480 of unmapped sequence to the amount of observed gene losses. This showed no suspicious  
481 discrepancies between gaps in the genome and regions without mapped reads.

482 Python<sup>3</sup><sup>67</sup> scripts used for summarizing DNA damage data, automating and filtering genome  
483 assemblies, and constructing the core-genome phylogeny can be found on Github:  
484 <https://github.com/DanneelsBram/DioscoreaOrrellaHerbarium>

#### 485 **QUANTIFICATION AND STATISTICAL ANALYSIS**

486 Statistical analysis on differences in DNA decay parameters were performed in Rstudio<sup>68</sup>  
487 using R v4.0.2<sup>69</sup> and the *dplyr* and *ggpubr* packages for visualisation of the results.

488 Comparisons of average read length, absolute and relative enrichment of C-to-T  
489 substitutions on the first base of the reads, and proportion of purines before strand breaks

490 between *D. sansibarensis* plastome and *O. dioscoreae* genome were performed only on  
491 samples with high enough coverage (at least 5x on both genomes; n=17). Comparisons were  
492 performed using the Wilcoxon signed-rank test to compare paired data, and the significance  
493 was assessed at a significance level of 0.05.

494 Correlations between average read length, absolute and relative enrichment of C-to-T  
495 substitutions on the first base, and proportion of purines before strand breaks with sample  
496 age (for both *D. sansibarensis* plastome and *O. dioscoreae* genome) were performed on  
497 samples with high enough coverage (at least 5x on both genomes), and that had an  
498 annotated collection date on the herbarium sheet (n=12). Correlations were calculated using  
499 the Pearson method, at a significance level of 0.05.

500 **References**

- 501 1. Orr, Y.M. (1923). The leaf glands of *Dioscorea macroura*. Notes from R. Bot. Gard. Edinburgh,  
502 57–72.
- 503 2. Burkhill, H.M. (1985). The useful plants of West Tropical Africa. Vol. 1. Families A-D. (Royal  
504 Botanic Gardens, Kew).
- 505 3. De Meyer, F., Danneels, B., Acar, T., Rasolomampianina, R., Rajaonah, M.T., Jeannoda, V., and  
506 Carlier, A. (2019). Adaptations and evolution of a heritable leaf nodule symbiosis between  
507 *Dioscorea sansibarensis* and *Orrella dioscoreae*. ISME J. *13*, 1831–1844.
- 508 4. Mira, A., Ochman, H., and Moran, N.A. (2001). Deletional bias and the evolution of bacterial  
509 genomes. Trends Genet. *17*, 589–596.
- 510 5. Kuo, C.H., Moran, N.A., and Ochman, H. (2009). The consequences of genetic drift for  
511 bacterial genome complexity. Genome Res. *19*, 1450–1454.
- 512 6. Manzano-Marín, A., Coeur d’acier, A., Clamens, A.-L., Orvain, C., Cruaud, C., Barbe, V., and  
513 Jousset, E. (2018). A Freeloader? The Highly Eroded Yet Large Genome of the *Serratia*  
514 *symbiotica* Symbiont of *Cinara strobilifera*. Genome Biol. Evol. *10*, 2178–2189.
- 515 7. Beghini, F., McIver, L.J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A.,  
516 Thomas, A.M., Manghi, P., Valles-Colomer, M., et al. (2020). Integrating taxonomic, functional,  
517 and strain-level profiling of diverse microbial communities with bioBakery 3. bioRxiv,  
518 <https://doi.org/10.1101/2020.11.19.388223>.
- 519 8. Carlier, A., Cnockaert, M., Fehr, L., Vandamme, P., and Eberl, L. (2017). Draft genome and  
520 description of *Orrella dioscoreae* gen. nov. sp. nov., a new species of Alcaligenaceae isolated  
521 from leaf acumens of *Dioscorea sansibarensis*. Syst. Appl. Microbiol. *40*, 11–21.
- 522 9. Weiß, C.L., Schuenemann, V.J., Devos, J., Shirsekar, G., Reiter, E., Gould, B.A., Stinchcombe,  
523 J.R., Krause, J., and Burbano, H.A. (2016). Temporal patterns of damage and decay kinetics of  
524 DNA retrieved from plant herbarium specimens. R. Soc. Open Sci. *3*, 160239.
- 525 10. Yoshida, K., Schuenemann, V.J., Cano, L.M., Pais, M., Mishra, B., Sharma, R., Lanz, C., Martin,  
526 F.N., Kamoun, S., Krause, J., et al. (2013). The rise and fall of the *Phytophthora infestans*  
527 lineage that triggered the Irish potato famine. Elife *2*, e00731.
- 528 11. Wales, N., Akman, M., Watson, R.H.B., Sánchez Barreiro, F., Smith, B.D., Gremillion, K.J.,  
529 Gilbert, M.T.P., and Blackman, B.K. (2019). Ancient DNA reveals the timing and persistence of  
530 organellar genetic bottlenecks over 3,000 years of sunflower domestication and  
531 improvement. Evol. Appl. *12*, 38–53.
- 532 12. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM:  
533 assessing the quality of microbial genomes recovered from isolates, single cells, and  
534 metagenomes. Genome Res. *25*, 1043–1055.
- 535 13. López-Fernández, S., Sonogo, P., Moretto, M., Pancher, M., Engelen, K., Pertot, I., and  
536 Campisano, A. (2015). Whole-genome comparative analysis of virulence genes unveils  
537 similarities and differences between endophytes and other symbiotic bacteria. Front.  
538 Microbiol. *6*, 419.
- 539 14. Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the  
540 prokaryotic species definition. Proc. Natl. Acad. Sci. U. S. A. *106*, 19126–19131.
- 541 15. Viruel, J., Conejero, M., Hidalgo, O., Pokorny, L., Powell, R.F., Forest, F., Kantar, M.B., Soto  
542 Gomez, M., Graham, S.W., Gravendeel, B., et al. (2019). A Target Capture-Based Method to

- 543 Estimate Ploidy From Herbarium Specimens. *Front. Plant Sci.* *10*, 937.
- 544 16. Viruel, J., Segarra-Moragues, J.G., Raz, L., Forest, F., Wilkin, P., Sanmartín, I., and Catalán, P.  
545 (2016). Late Cretaceous-Early Eocene origin of yams ( *Dioscorea* , *Dioscoreaceae*) in the  
546 Laurasian Palaeartic and their subsequent Oligocene-Miocene diversification. *J. Biogeogr.* *43*,  
547 750–762.
- 548 17. Wilkin, P. (2001). *Dioscoreaceae* of South-Central Africa. *Kew Bull.* *56*, 361.
- 549 18. Schuenemann, V.J., Singh, P., Mendum, T.A., Krause-Kyora, B., Jäger, G., Bos, K.I., Herbig, A.,  
550 Economou, C., Benjak, A., Busso, P., et al. (2013). Genome-wide comparison of medieval and  
551 modern *Mycobacterium leprae*. *Science* *341*, 179–83.
- 552 19. Weiß, C.L., Gansauge, M.-T., Aximu-Petri, A., Meyer, M., and Burbano, H.A. (2020). Mining  
553 ancient microbiomes using selective enrichment of damaged DNA molecules. *BMC Genomics*  
554 *21*, 432.
- 555 20. Bernal, P., Llamas, M.A., and Filloux, A. (2018). Type VI secretion systems in plant-associated  
556 bacteria. *Environ. Microbiol.* *20*, 1–15.
- 557 21. Costa, T.R.D., Felisberto-Rodrigues, C., Meir, A., Prevost, M.S., Redzej, A., Trokter, M., and  
558 Waksman, G. (2015). Secretion systems in Gram-negative bacteria: structural and mechanistic  
559 insights. *Nat. Rev. Microbiol.* *13*, 343–359.
- 560 22. Mehrabi, R., Bahkali, A.H., Abd-Elsalam, K.A., Moslem, M., Ben M’barek, S., Gohari, A.M.,  
561 Jashni, M.K., Stergiopoulos, I., Kema, G.H.J., and de Wit, P.J.G.M. (2011). Horizontal gene and  
562 chromosome transfer in plant pathogenic fungi affecting host range. *FEMS Microbiol. Rev.* *35*,  
563 542–54.
- 564 23. Kuo, C.H., and Ochman, H. (2010). The extinction dynamics of bacterial pseudogenes. *PLoS*  
565 *Genet.* *6*, e1001050.
- 566 24. Bolotin, E., and Hershberg, R. (2016). Bacterial intra-species gene loss occurs in a largely  
567 clocklike manner mostly within a pool of less conserved and constrained genes. *Sci. Rep.* *6*,  
568 35168.
- 569 25. Danneels, B., Pinto-Carbó, M., and Carlier, A. (2018). Patterns of Nucleotide Deletion and  
570 Insertion Inferred from Bacterial Pseudogenes. *Genome Biol. Evol.* *10*, 1792–1802.
- 571 26. Lemaire, B., Vandamme, P., Merckx, V., Smets, E., and Dessein, S. (2011). Bacterial leaf  
572 symbiosis in angiosperms: host specificity without co-speciation. *PLoS One* *6*, e24430.
- 573 27. Van Oevelen, S., De Wachter, R., Vandamme, P., Robbrecht, E., and Prinsen, E. (2002).  
574 Identification of the bacterial endosymbionts in leaf galls of *Psychotria* (*Rubiaceae*,  
575 angiosperms) and proposal of “*Candidatus Burkholderia kirkii*” sp. nov. *Int. J. Syst. Evol.*  
576 *Microbiol.* *52*, 2023–2027.
- 577 28. Carlier, A., and Eberl, L. (2012). The eroded genome of a *Psychotria* leaf symbiont: Hypotheses  
578 about lifestyle and interactions with its plant host. *Environ. Microbiol.* *14*, 2757–2769.
- 579 29. Alonso, D.P., Mancini, M.V., Damiani, C., Cappelli, A., Ricci, I., Alvarez, M.V.N., Bandi, C.,  
580 Ribolla, P.E.M., and Favia, G. (2019). Genome Reduction in the Mosquito Symbiont *Asaia*.  
581 *Genome Biol. Evol.* *11*, 1–10.
- 582 30. Manzano-Marín, A., and Latorre, A. (2016). Snapshots of a shrinking partner: Genome  
583 reduction in *Serratia symbiotica*. *Sci. Rep.* *6*, 32590.
- 584 31. Kuo, C.-H., and Ochman, H. (2009). Deletional Bias across the Three Domains of Life. *Genome*

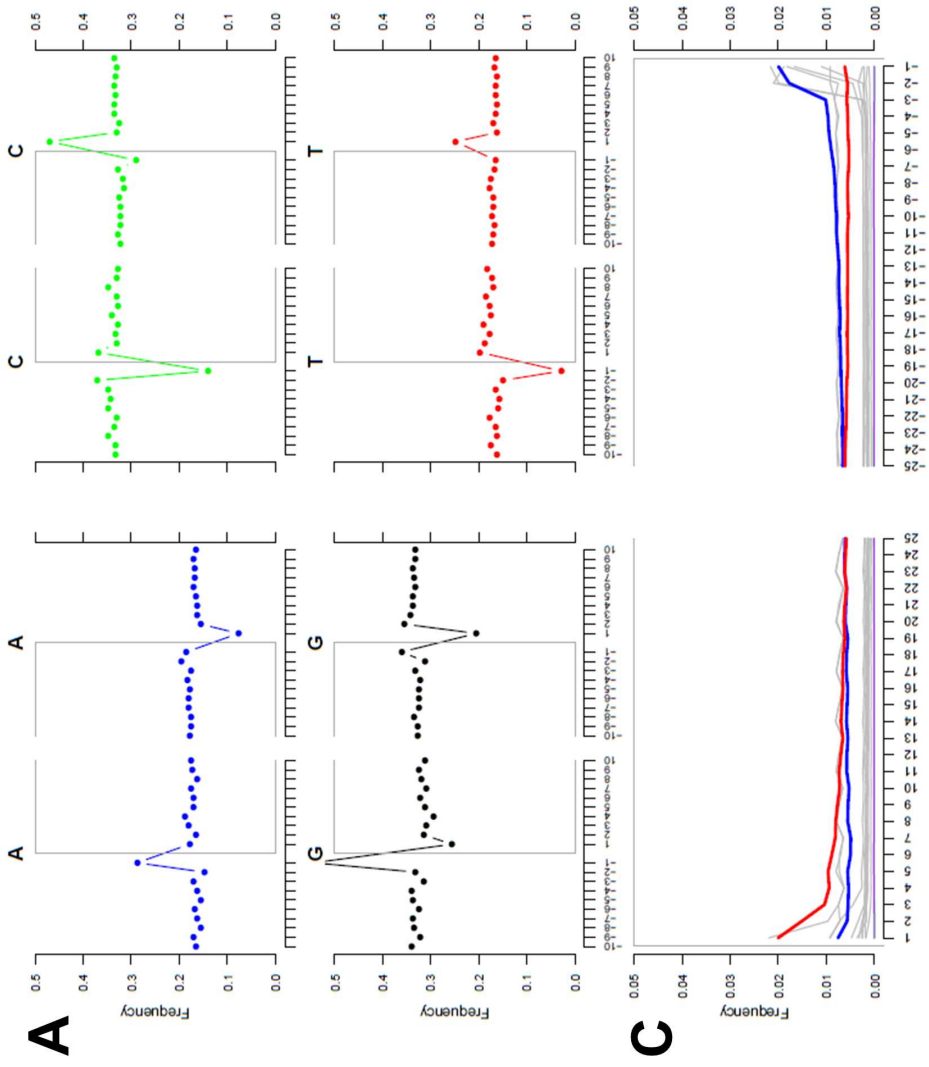
- 585 Biol. Evol. 1, 145–152.
- 586 32. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., and Orlando, L. (2013).  
587 mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters.  
588 Bioinformatics 29, 1682–1684.
- 589 33. Clement, M., Posada, D., and Crandall, K.A. (2000). TCS: A computer program to estimate gene  
590 genealogies. Mol. Ecol. 9, 1657–1659.
- 591 34. Gilbert, M.T.P., Wilson, A.S., Bunce, M., Hansen, A.J., Willerslev, E., Shapiro, B., Higham, T.F.,  
592 Richards, M.P., O’Connell, T.C., Tobin, D.J., et al. (2004). Ancient mitochondrial DNA from hair.  
593 Curr. Biol. 14, R463–R464.
- 594 35. Wales, N., Andersen, K., Cappellini, E., Ávila-Arcos, M.C., and Gilbert, M.T.P. (2014).  
595 Optimization of DNA recovery and amplification from non-carbonized archaeobotanical  
596 remains. PLoS One 9, e86827.
- 597 36. Cappellini, E., Gilbert, M.T.P., Geuna, F., Fiorentino, G., Hall, A., Thomas-Oates, J., Ashton,  
598 P.D., Ashford, D.A., Arthur, P., Campos, P.F., et al. (2010). A multidisciplinary study of  
599 archaeological grape seeds. Naturwissenschaften 97, 205–217.
- 600 37. Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C.,  
601 Garcia, N., Paabo, S., Arsuaga, J.-L., et al. (2013). Complete mitochondrial genome sequence  
602 of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc. Natl.  
603 Acad. Sci. 110, 15758–15763.
- 604 38. Wales, N., Carøe, C., Sandoval-Velasco, M., Gamba, C., Barnett, R., Samaniego, J.A., Madrigal,  
605 J.R., Orlando, L., and Gilbert, M.T.P. (2015). New insights on single-stranded versus double-  
606 stranded DNA library preparation for ancient DNA. Biotechniques 59, 368–371.
- 607 39. Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly  
608 multiplexed target capture and sequencing. Cold Spring Harb. Protoc. 2010, pdb.prot5448.
- 609 40. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing  
610 reads. EMBnet.journal 17, 10.
- 611 41. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina  
612 sequence data. Bioinformatics 30, 2114–2120.
- 613 42. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler  
614 transform. Bioinformatics 25, 1754–1760.
- 615 43. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing  
616 genomic features. Bioinformatics 26, 841–842.
- 617 44. Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification  
618 using exact alignments. Genome Biol. 15, R46.
- 619 45. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L.  
620 (2009). BLAST+: architecture and applications. BMC Bioinformatics 10, 421.
- 621 46. Kahlke, T., and Ralph, P.J. (2019). BASTA – Taxonomic classification of sequences and  
622 sequence bins using last common ancestor estimations. Methods Ecol. Evol. 10, 100–103.
- 623 47. Ondov, B.D., Bergman, N.H., and Phillippy, A.M. (2011). Interactive metagenomic visualization  
624 in a Web browser. BMC Bioinformatics 12, 385.
- 625 48. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and  
626 Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25,

- 627 2078–2079.
- 628 49. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M.,  
629 Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: A new genome assembly  
630 algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* *19*, 455–477.
- 631 50. Ponsting, H., and Ning, Z. (2010). SMALT - A New Mapper for DNA Sequencing Reads.  
632 *F1000Posters* *1*, 1.
- 633 51. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUASt: quality assessment tool for  
634 genome assemblies. *Bioinformatics* *29*, 1072–1075.
- 635 52. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E. V, and Zdobnov, E.M. (2015).  
636 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.  
637 *Bioinformatics* *31*, 3210–2.
- 638 53. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping  
639 and population genetical parameter estimation from sequencing data. *Bioinformatics* *27*,  
640 2987–2993.
- 641 54. Bertels, F., Silander, O.K., Pachkov, M., Rainey, P.B., and van Nimwegen, E. (2014). Automated  
642 reconstruction of whole-genome phylogenies from short-sequence reads. *Mol. Biol. Evol.* *31*,  
643 1077–88.
- 644 55. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010).  
645 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the  
646 performance of PhyML 3.0. *Syst. Biol.* *59*, 307–21.
- 647 56. Croucher, N.J., Page, A.J., Connor, T.R., Delaney, A.J., Keane, J.A., Bentley, S.D., Parkhill, J., and  
648 Harris, S.R. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial  
649 whole genome sequences using Gubbins. *Nucleic Acids Res.* *43*, e15.
- 650 57. Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of  
651 large phylogenies. *Bioinformatics* *30*, 1312–1313.
- 652 58. Leigh, J.W., and Bryant, D. (2015). POPART: full-feature software for haplotype network  
653 construction. *Methods Ecol. Evol.* *6*, 1110–1116.
- 654 59. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high  
655 throughput. *Nucleic Acids Res.* *32*, 1792–7.
- 656 60. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: A tool for  
657 automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* *25*, 1972–  
658 1973.
- 659 61. Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., and Rambaut, A. (2018).  
660 Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* *4*,  
661 vey016.
- 662 62. Pritchard, L., Glover, R.H., Humphris, S., Elphinstone, J.G., and Toth, I.K. (2016). Genomics and  
663 taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal.*  
664 *Methods* *8*, 12–24.
- 665 63. Emms, D.M., and Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for  
666 comparative genomics. *Genome Biol.* *20*, 238.
- 667 64. Csuo, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and  
668 likelihood. *Bioinformatics* *26*, 1910–1912.

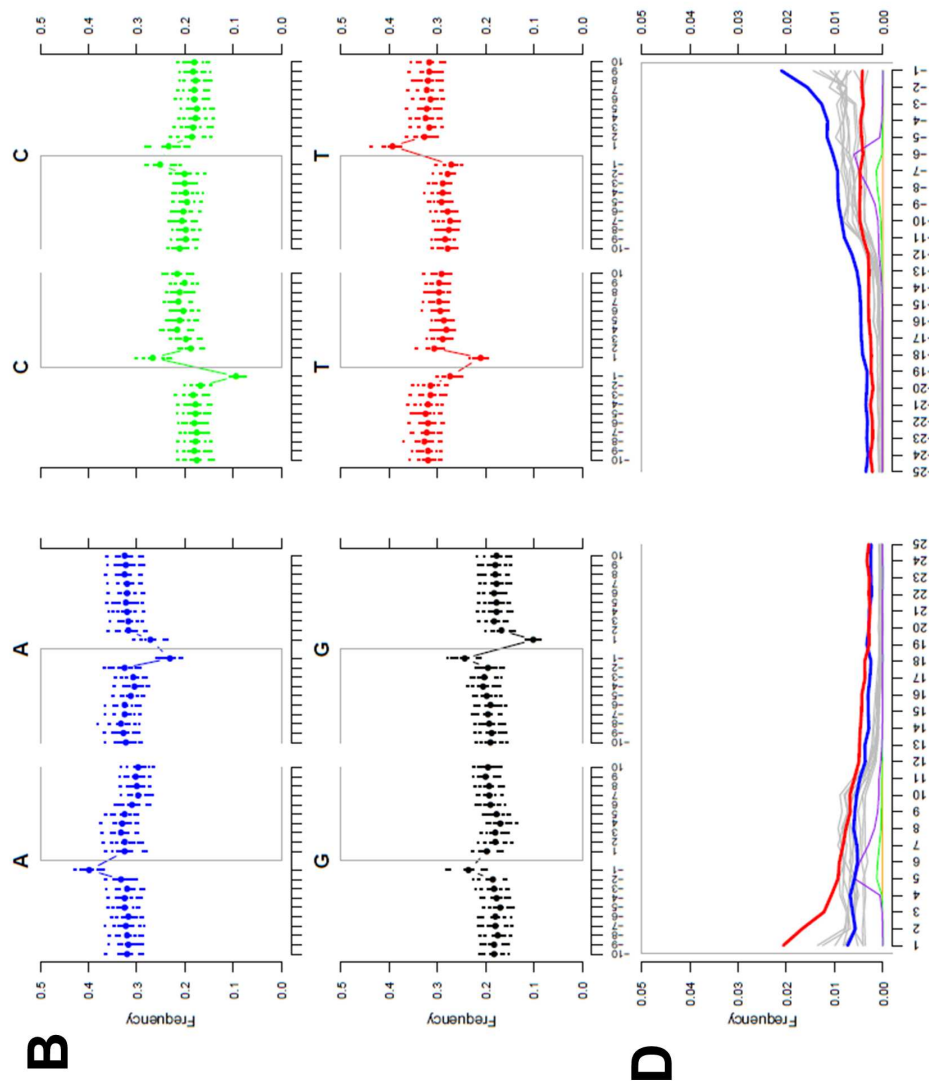


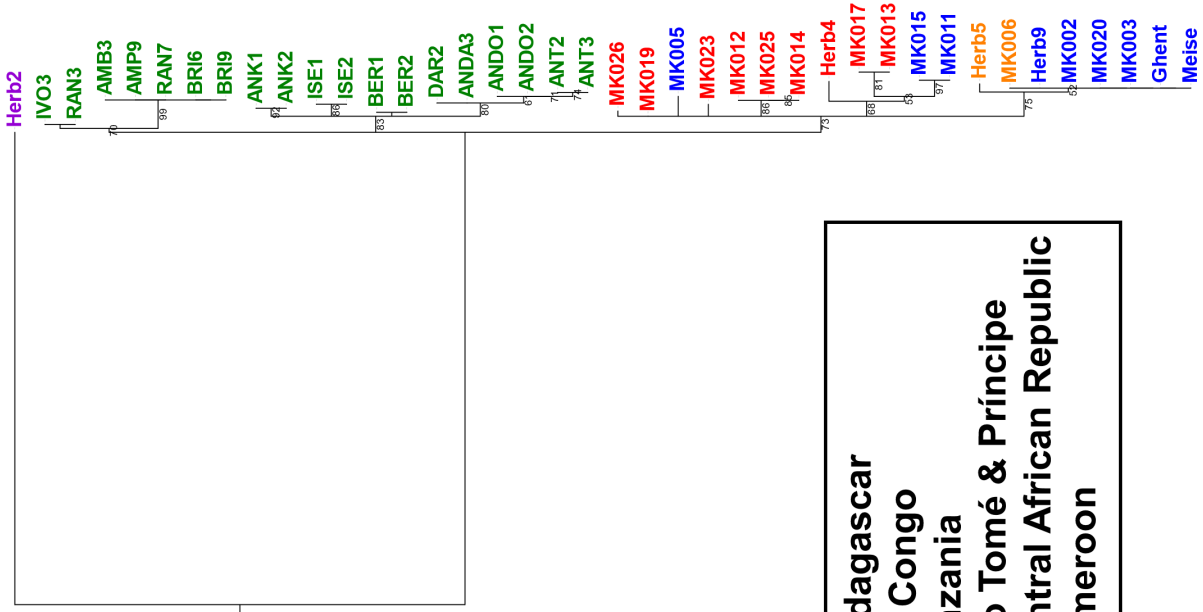
- 669 65. Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, Analysis, and  
670 Visualization of Phylogenomic Data. *Mol. Biol. Evol.* 33, 1635–1638.
- 671 66. Sukumaran, J., and Holder, M.T. (2010). DendroPy: a Python library for phylogenetic  
672 computing. *Bioinformatics* 26, 1569–1571.
- 673 67. Van Rossum, G., and Drake, F.L. (2009). Python 3 Reference Manual.
- 674 68. RStudio Team (2020). RStudio: Integrated Development for R. RStudio.
- 675 69. R Core Team (2020). R: A language and environment for statistical computing.  
676

# *O. dioscoreae*



# *D. sansibarensis* chloroplast



**A****B**