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

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

Author affiliations:

[1] Laboratory of Microbiology, Ghent University, 9000 Ghent, Belgium

[2] Royal Botanical Gardens Kew, Richmond, London, TW9 3AE, United Kingdom

[3] Department of Prehistory and Institute of Environmental Science and Technology (ICTA), University of Barcelona, 08193 Bellaterra, Spain

[4] Meise Botanic Garden, 1860 Meise, Belgium

[5] Plant Conservation and Population Biology, KULeuven, 3000 Leuven, Belgium

[6] Department of Archaeology, University of York, Heslington, York, YO10 5DD, United Kingdom

[7] LIPM, Université de Toulouse, INRAE, CNRS, 31320 Castanet-Tolosan, France

*Correspondence and lead contact: aurelien.carlier@inrae.fr

Twitter handle: @AurelienCarlier

Summary

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

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

Results

Recovery of DNA from preserved Dioscorea sansibarensis *leaf glands*

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

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

Taxonomic composition of D. sansibarensis *leaf glands*

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

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

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

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

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

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

High specificity of leaf symbiosis without phylogenetic congruence

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

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

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

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

domain effector proteins (ODI_R0793, ODI_R0797 and ODI_R0809) were likely subjected to gene conversion or HGT. In addition, a pair of Rhs/VgrG proteins putative T6SS effector proteins was encoded in all genomes of clade II, but in none of the genomes of clade I. Apart from those, additional Rhs and/or VgrG proteins domains were also detected in 4 other MAGs (AMP9, BER1, BER2, and MK019).

Discussion

Herbarium samples provide a reliable source of symbiont DNA

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

Phylogeography of Dioscorea sansibarensis

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

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

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

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

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

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

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

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

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

Acknowledgments

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

Author Contributions

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

Declaration of Interests

The authors declare no conflict of interest.

FIGURE LEGENDS

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

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

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

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

SUPPLEMENTARY FIGURES AND TABLES

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

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

Figure S3: Prevalence of orthogroups with certain number of genomes

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

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

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

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

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

311

STAR Methods

KEY RESOURCE TABLE

RESOURCE AVAILABILITY

Lead contact

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

Materials Availability

This study did not generate new unique reagents.

Data and code availability

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

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Plants

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

METHOD DETAILS

Sampling and DNA-extraction

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

Total DNA-isolation and genomic library preparation of ten specimens (Herb1-Herb10), representing various geographic locations, different ages, and diverse gland sizes were performed in the palaeogenomics facility at the Department of Archaeology of the University of York (UK). Twenty-six specimens (MK001-MK026) were processed at the department of Microbiology of Ghent University in a room disinfected with bleach and under a disinfected PCR cabinet (AirClean 600 PCR Workstation, Starlab, Hamburg, Germany). All utensils were disinfected using bleach and/or followed a UV treatment prior to their usage. When possible, sterile disposable items were used. Extraction blanks were included to monitor possible DNA-contamination. Total genomic DNA from leaf nodules was extracted using the protocol described in Gilbert *et al.*[34], which was found to perform well on botanical specimens[35,36]. The leaf glands were cut into small pieces using a sterile scalpel and placed into sterile 2ml Eppendorf Lo-bind microfuge tubes. The samples were incubated on a shaker overnight at 55°C in 1200 µl of extraction buffer (10mM Tris-HCl pH 8, 10mM NaCl, 2% SDS, 5mM $CaCl_2$, 2.5mM EDTA pH 8, 0.5mg/ml Proteinase K, and 40mM DTT). Supernatants were extracted twice with an equal volume of 25:24:1 phenol/chloroform/isoamylalcohol. The resulting DNA was diluted in 13x binding buffer (5M guanidine hydrochloride, 40% isopropanol, 0.05% Tween-20, and 90mM Sodium Acetate pH 5.2)[37] and purified using a MinElute PCR purification kit (Qiagen, Hilden, Germany) following the manufacturer's recommendation. Extractions blanks yielded no detectable amounts of DNA (Table S1)

Library preparation and sequencing

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

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

Read processing and mapping

Sequencing adapters were removed using Cutadapt v2.10[40], and low-quality bases were removed using Trimmomatic v0.39[41]. Trimmed reads were mapped to the *Dioscorea sansibarensis* chloroplast sequence (NCBI accession GCA_900631875.1) and a draft version of the nuclear genome (unpublished) from a plant obtained from the Botanical Garden of Ghent University, to the associated *Orrella dioscoreae* LMG 29303^T genome (NCBI accession GCA_900089455.2), and to the human reference genome GRCh38 (NCBI accession GCF_000001405.39) using BWA aln v0.7.17, with seeding disabled[42]. Coverage for every genome was calculated using BEDTools v2.27.1 *genomecov* command[43]. Taxonomic composition of samples was determined using Metaphlan 3[7], Kraken v1.1.1[44] (using a custom database of bacterial and chloroplast sequences[8]), and blastn[45] searches against the NCBI non-redundant nucleotide database (accessed 03/2020), summarized using BASTA and Krona[46,47].

DNA damage analysis

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

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

Genome assembly

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

Phylogenetic analyses

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

Phylogenetic trees based on plastid data were constructed using PhyML v3.3.3[55] using the F81 model (selected using the CLC Main Workbench (Qiagen) model testing tool), 100 bootstrap replicates and the plastid sequence of *Dioscorea elephantipes* (NCBI accession NC_009601) as outgroup. For the *O. dioscoreae* phylogeny, Gubbins[56] was used to remove regions with elevated rates of base substitutions from the alignment, to mitigate the effect of recombination on whole genome or SNP-based phylogenies. A maximum-likelihood phylogeny was created using RAxML v8.2.12[57] (rapid bootstrapping and best-scoring ML mode, using 100 bootstrap replicates and the GTRGAMMA substitution model). The genome sequence of *Achromobacter xylosoxidans* ADAF13 (NCBI accession GCA_001566985) was used as outgroup to root the phylogeny. A haplotype network of the plastid sequences was created with the TCS algorithm, implemented in POPART[33,58], using the SNP-based alignment used for the phylogeny as input. Samples with more than 25% of gaps in their sequence were removed from downstream analyses. To further control for the effect of aDNA damage (mainly transitions) on the low-coverage plastome analysis, the phylogenetic analysis and haplotype network were recalculated using only transversions (89 transversions out of 168 SNPs). In an effort to assess the phylogeny of *D. sansibarensis* based on its nuclear genome, the sequence of the high-copy plant marker ITS (internal transcribed spacer) was extracted by searching metagenome sequences with blastn[45]. As a query, a known 18S-5S-28S region of *D. sansibarensis* (NCBI accession DQ267929.1) was used. The matching sequences were extracted from the metagenome contigs, including 250 extra flanking base pairs. Using this method, 20 herbarium ITS sequences could be extracted. Sequences were aligned using Muscle[59] and the alignment was trimmed using trimAL[60] to remove columns with >80% gaps. The resulting alignment resulted in no variable sites between samples

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

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

Comparative genomics of O. dioscoreae *genomes*

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

Python3[67] scripts used for summarizing DNA damage data, automating and filtering genome assemblies, and constructing the core-genome phylogeny can be found on Github: https://github.com/DanneelsBram/DioscoreaOrrellaHerbarium

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis on differences in DNA decay parameters were performed in Rstudio[68] using R v4.0.2[69] and the *dplyr* and *ggpubr* packages for visualisation of the results.

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

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

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

References

1. Orr, Y.M. (1923). The leaf glands of Dioscorea macroura. Notes from R. Bot. Gard. Edinburgh, 57–72.

2. Burkhil, H.M. (1985). The useful plants of West Tropical Africa. Vol. 1. Families A-D. (Royal Botanic Gardens, Kew).

3. De Meyer, F., Danneels, B., Acar, T., Rasolomampianina, R., Rajaonah, M.T., Jeannoda, V., and Carlier, A. (2019). Adaptations and evolution of a heritable leaf nodule symbiosis between Dioscorea sansibarensis and Orrella dioscoreae. ISME J. *13*, 1831–1844.

4. Mira, A., Ochman, H., and Moran, N.A. (2001). Deletional bias and the evolution of bacterial genomes. Trends Genet. *17*, 589–596.

5. Kuo, C.H., Moran, N.A., and Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. Genome Res. *19*, 1450–1454.

6. Manzano-Marín, A., Coeur d'acier, A., Clamens, A.-L., Orvain, C., Cruaud, C., Barbe, V., and Jousselin, E. (2018). A Freeloader? The Highly Eroded Yet Large Genome of the Serratia symbiotica Symbiont of Cinara strobi. Genome Biol. Evol. *10*, 2178–2189.

7. Beghini, F., McIver, L.J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Thomas, A.M., Manghi, P., Valles-Colomer, M., et al. (2020). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. bioRxiv, https://doi.org/10.1101/2020.11.19.388223.

8. Carlier, A., Cnockaert, M., Fehr, L., Vandamme, P., and Eberl, L. (2017). Draft genome and description of Orrella dioscoreae gen. nov. sp. nov., a new species of Alcaligenaceae isolated from leaf acumens of Dioscorea sansibarensis. Syst. Appl. Microbiol. *40*, 11–21.

9. Weiß, C.L., Schuenemann, V.J., Devos, J., Shirsekar, G., Reiter, E., Gould, B.A., Stinchcombe, J.R., Krause, J., and Burbano, H.A. (2016). Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. R. Soc. Open Sci. *3*, 160239.

10. Yoshida, K., Schuenemann, V.J., Cano, L.M., Pais, M., Mishra, B., Sharma, R., Lanz, C., Martin, F.N., Kamoun, S., Krause, J., et al. (2013). The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. Elife *2*, e00731.

11. Wales, N., Akman, M., Watson, R.H.B., Sánchez Barreiro, F., Smith, B.D., Gremillion, K.J., Gilbert, M.T.P., and Blackman, B.K. (2019). Ancient DNA reveals the timing and persistence of organellar genetic bottlenecks over 3,000 years of sunflower domestication and improvement. Evol. Appl. *12*, 38–53.

12. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055.

13. Lòpez-Fernàndez, S., Sonego, P., Moretto, M., Pancher, M., Engelen, K., Pertot, I., and Campisano, A. (2015). Whole-genome comparative analysis of virulence genes unveils similarities and differences between endophytes and other symbiotic bacteria. Front. Microbiol. *6*, 419.

14. Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U. S. A. *106*, 19126–19131.

15. Viruel, J., Conejero, M., Hidalgo, O., Pokorny, L., Powell, R.F., Forest, F., Kantar, M.B., Soto 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 Palaearctic 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. Csuos, 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.

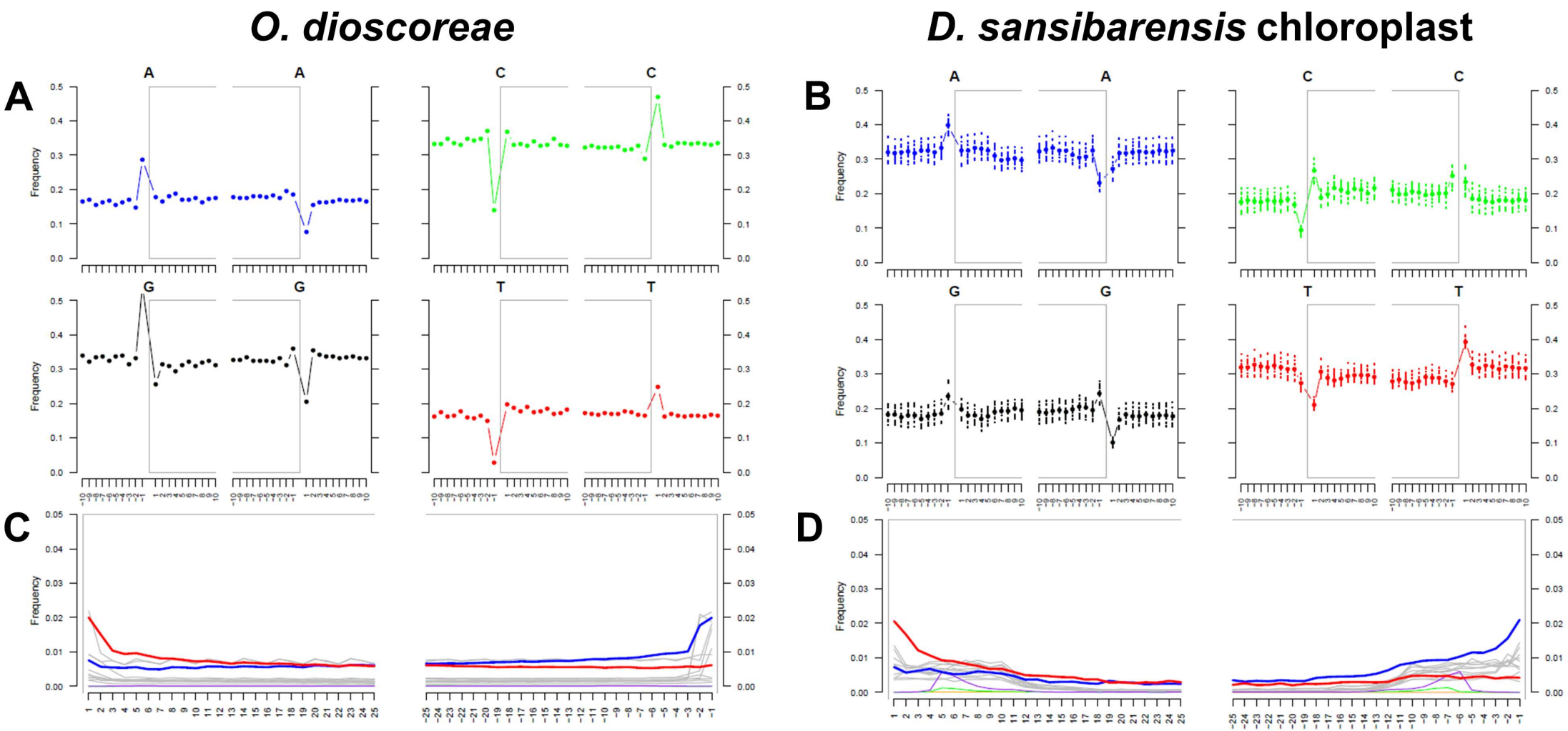
671 66. Sukumaran, J., and Holder, M.T. (2010). DendroPy: a Python library for phylogenetic
672 computing. Bioinformatics *26*, 1569–1571.

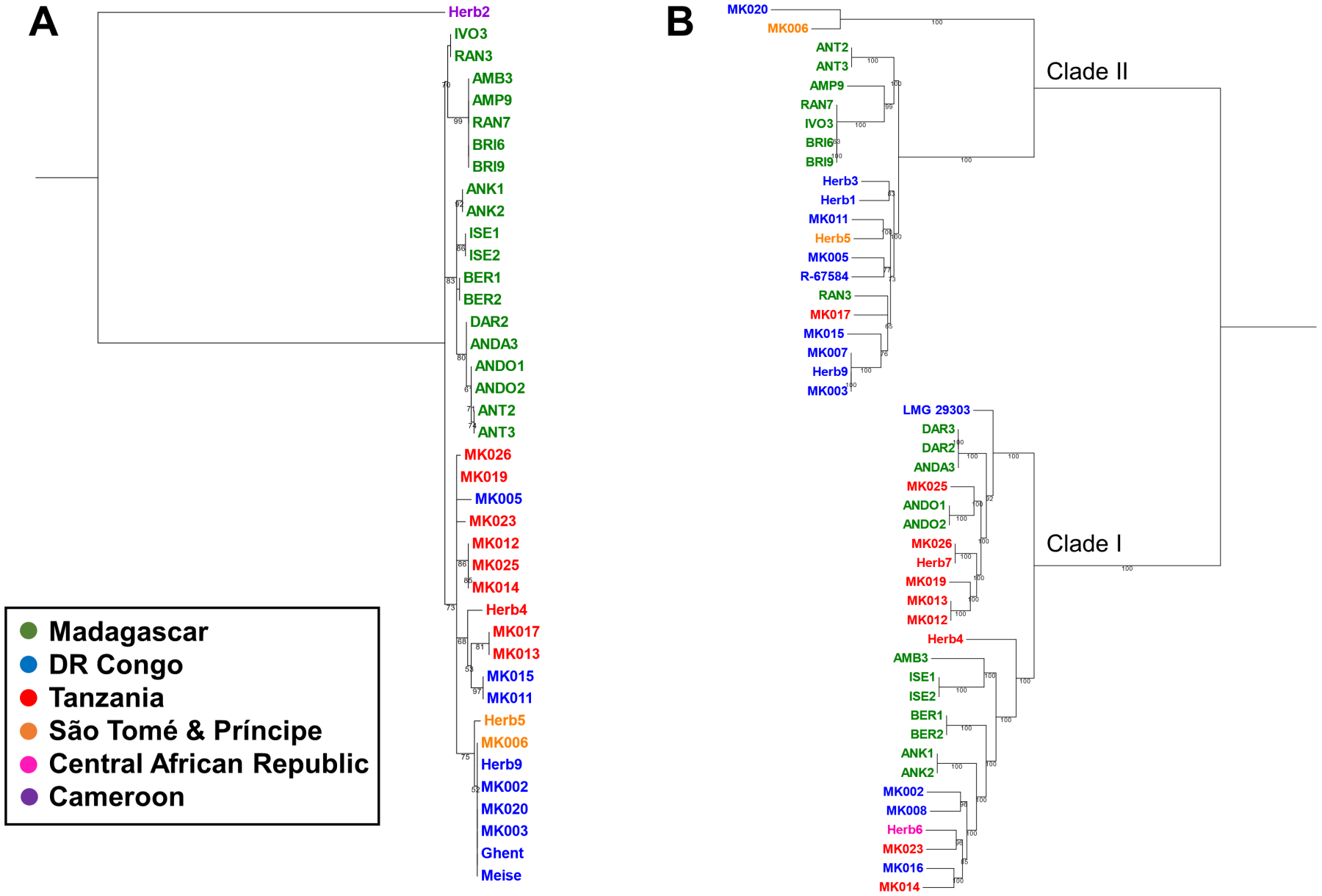
673 67. Van Rossum, G., and Drake, F.L. (2009). Python 3 Reference Manual.

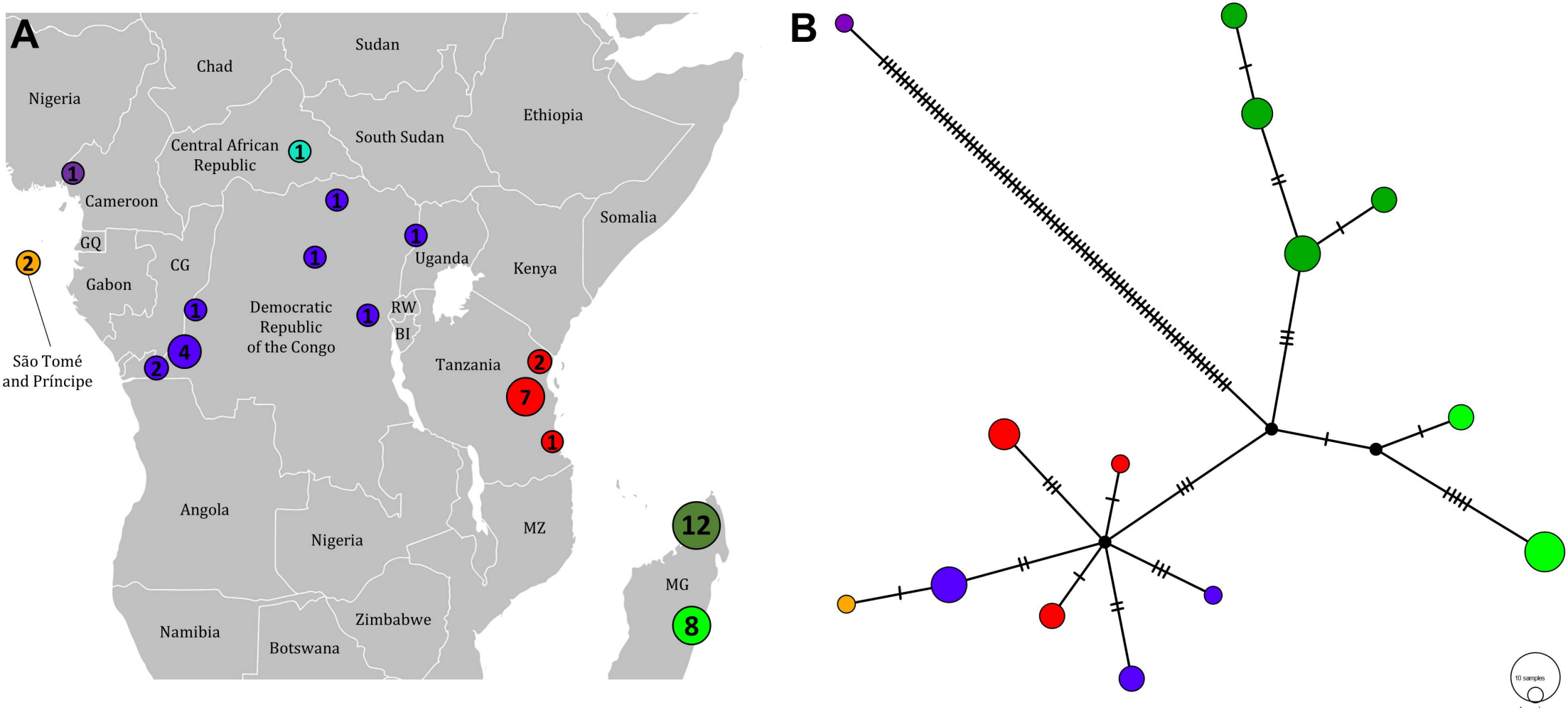
674 68. RStudio Team (2020). RStudio: Integrated Development for R. RStudio.

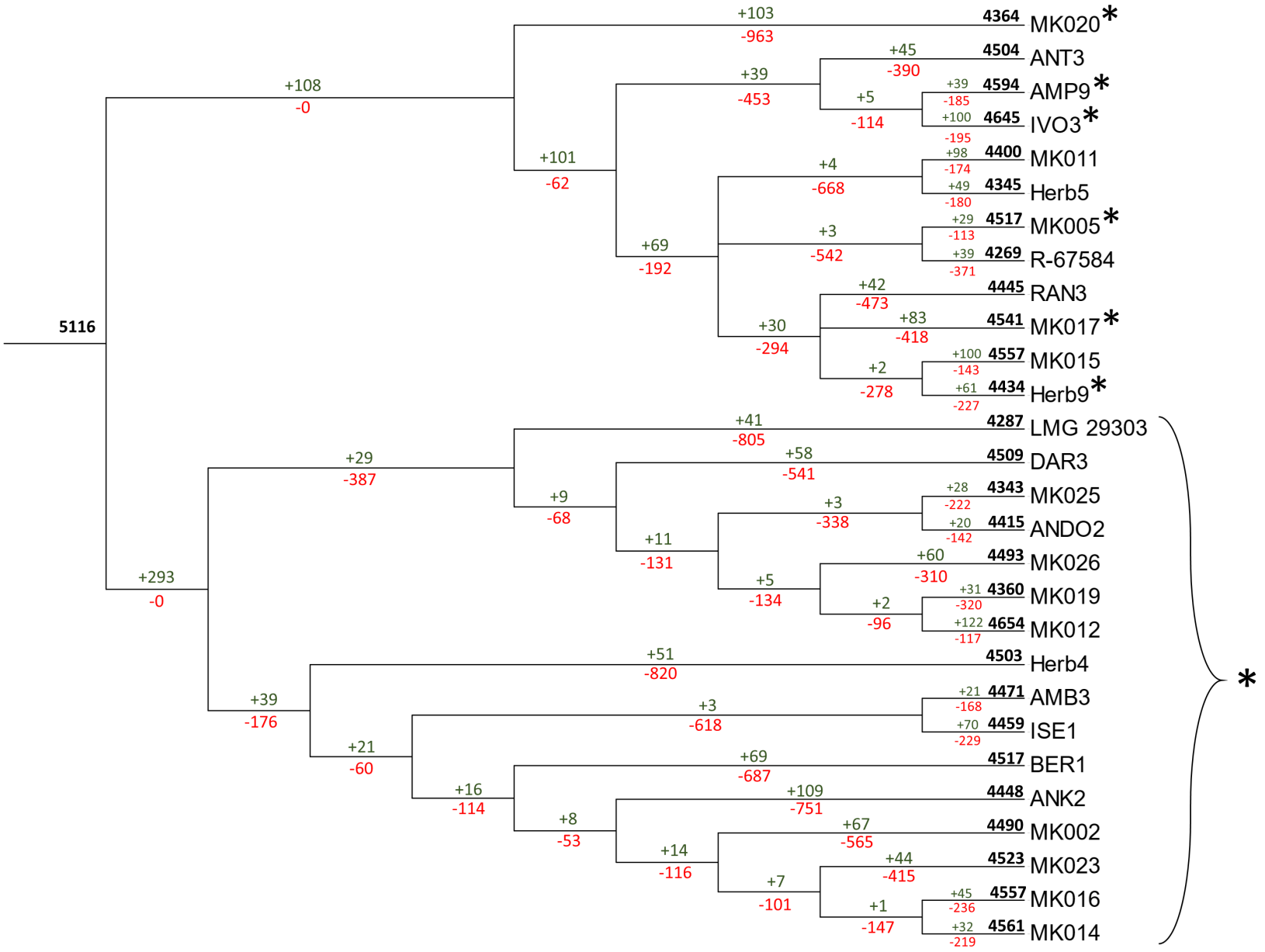
675 69. R Core Team (2020). R: A language and environment for statistical computing.

676


O. dioscoreae
D. sansibarensis chloroplast
A
B
C
D
A
A
C
C
G
G
T
T
Frequency


A
Herb2
IVO3
RAN3
AMB3
AMP9
RAN7
BRI6
BRI9
ANK1
ANK2
ISE1
ISE2
BER1
BER2
DAR2
ANDA3
ANDO1
ANDO2
ANT2
ANT3
MK026
MK019
MK005
MK023
MK012
MK025
MK014
Herb4
MK017
MK013
MK015
MK011
Herb5
MK006
Herb9
MK002
MK020
MK003
Ghent
Meise
Madagascar
DR Congo
Tanzania
São Tomé & Príncipe
Central African Republic
Cameroon
B
MK020
MK006
ANT2
ANT3
AMP9
RAN7
IVO3
BRI6
BRI9
Herb3
Herb1
MK011
Herb5
MK005
R-67584
RAN3
MK017
MK015
MK007
Herb9
MK003
Clade II
LMG 29303
DAR3
DAR2
ANDA3
MK025
ANDO1
ANDO2
MK026
Herb7
MK019
MK013
MK012
Herb4
AMB3
ISE1
ISE2
BER1
BER2
ANK1
ANK2
MK002
MK008
Herb6
MK023
MK016
MK014
Clade I


A
Sudan
Chad
Nigeria
Ethiopia
South Sudan
Central African Republic
Cameroon
Somalia
GQ
Uganda
CG
Kenya
Gabon
Democratic Republic of the Congo
RW
BI
São Tomé and Príncipe
Tanzania
Angola
MZ
Nigeria
MG
Zimbabwe
Namibia
Botswana
B
10 samples


5116
+108
-0
+103
-963
4364
MK020*
+45
-390
4504
ANT3
+39
-453
+39
4594
AMP9*
+5
-114
-185
+100
4645
IVO3*
-195
+101
-62
+98
4400
MK011
+4
-668
-174
+49
4345
Herb5
-180
+69
-192
+29
4517
MK005*
+3
-542
-113
+39
4269
R-67584
-371
+42
4445
RAN3
-473
+83
4541
MK017*
+30
-294
-418
+100
4557
MK015
-143
+2
-278
+61
4434
Herb9*
-227
+293
-0
+41
4287
LMG 29303
-805
+29
-387
+58
4509
DAR3
-541
+9
-68
+28
4343
MK025
-222
+3
-338
+20
4415
ANDO2
-142
+11
-131
+60
4493
MK026
-310
+5
-134
+31
4360
MK019
-320
+2
-96
+122
4654
MK012
-117
+51
4503
Herb4
-820
+39
-176
+21
4471
AMB3
-168
+3
-618
+70
4459
ISE1
-229
+21
-60
+69
4517
BER1
-687
+16
-114
+109
4448
ANK2
-751
+8
-53
+67
4490
MK002
-565
+14
-116
+44
4523
MK023
-415
+7
-101
+45
4557
MK016
-236
+1
-147
+32
4561
MK014
-219
*