Patterns of transmission and horizontal gene transfer in the Dioscorea sansibarensis leaf symbiosis revealed by whole-genome sequencing

Bram Danneels, Juan Viruel, Krista Mcgrath, Steven Janssens, Nathan Wales, Paul Wilkin, Aurélien Carlier

TITLE: Shedding light on the evolution of the Zanzibar yam leaf symbiosis using whole genome sequences from historical herbarium specimens

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, Castanet-Tolosan, France

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

Summary

Herbaria contain an invaluable record of plant specimens from across the world. They can be used to study the evolutionary history and geographic distribution of plants that have experienced recent demographic changes, and provide opportunities to study species that are otherwise difficult to collect. Several plant species also establish highly specific interactions with microorganisms, which can be preserved in the herbarium tissues. In this work, we investigated the leaf symbiosis between the yam *Dioscorea sansibarensis* and its bacterial symbiont *Orrella dioscoreae* using preserved leaf samples collected from different locations in Africa. We recovered DNA from the extracellular symbiont in all samples, showing that the symbionts are widespread in continental Africa. We also observed both similarities and differences in the dynamics of DNA decay in both plastid and symbiont DNA. Despite the degraded nature of this ancient DNA, we managed to construct 17 *de novo* symbiont genomes from short DNA fragments. Phylogenetic and genomic analyses revealed that horizontal transmission of symbionts and horizontal gene transfer shape the symbiont's evolution despite the captive nature of symbiont populations. These mechanisms could help explain why the symbiont genomes do not display clear signs of reductive genome evolution. Furthermore, phylogenetic analysis of the *Dioscorea sansibarensis* plastid genome revealed a strong geographical clustering of samples, providing further insight in *D. sansibarensis*' spread, and provide evidence of that the symbiosis was established earlier than previously estimated.

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

Introduction

Herbaria are a tremendous resource for biological research, containing plant specimens and associated data collected over at least the last 300 years all around the world [1]. With an estimated 350 million specimens, they enable studies in plant biogeography, history, and population dynamics [2–4]. Furthermore, with the development of novel high-throughput sequencing (HTS) methods and the continuous optimisation of laboratory techniques, especially ancient DNA (aDNA) methodologies, herbarium specimens can now also be used for large-scale genomic studies [5–7]. In its early days, aDNA research was mainly focused on investigating archaeological and museum specimens, with the first isolated aDNA being a 213 bp sequence from the extinct quagga [8]. Further development of PCR technologies, and later HTS, made it possible to sequence full genomes, such as that of the woolly mammoth [9], or certain plant species like maize, barley and grapevine [10–12]. Early reports on ancient DNA were however fraught with issues stemming from the contamination of samples with exogenous DNA [13–17]. In fact, these setbacks were a direct consequence of two major issues in the aDNA research field: the rapid degradation of DNA in non-living tissue, leading to low yields of highly-degraded DNA, and a high risk of contamination from ubiquitous, chemically intact "modern" DNA. Although contamination can never be completely ruled out, the risk can be mitigated by adherence to strict laboratory procedures [18–21]. The use of high-throughput sequencing technology further provides tools ideally suited for investigating aDNA: short read sequencing is ideal for aDNA inserts which are naturally very short, greater amounts of data allow better quantification and characterisation of contaminant DNA, and it allows for the study of aDNA damage patterns.

While the general dynamics of DNA decay are rather well understood at a molecular level [22], the degradation dynamics in various types of preserved tissue (e.g. bone, herbarium, museum specimens) remains difficult to model [23–25]. In general, most aDNA shows similar patterns of degradation. First, aDNA is highly fragmented, with DNA strands breaking most often near purine bases [26]. This bias is likely caused by preferential hydrolytic depurination over depyrimidination, followed by DNA breakage at abasic sites [22]. The level of fragmentation does not seem to correlate with the age of the specimens, with regional climate a better predictor of DNA fragmentation [24]. Second, aDNA tends to accumulate deaminated cytosine residues (creating uracil residues), preferably at the edge of single-stranded fragments or overhangs [26]. This can be detected after sequencing, as the polymerases will erroneously pair uracil with adenosine, resulting in mismatches. This type of DNA damage is correlated with specimen age [26–29], and the presence of uracil residues can be used to enrich samples for aDNA [30]. Furthermore, the presence of these degradation patterns can be used to authenticate aDNA [17,19,20,28,31].

In contrast to studies of animal and plant ancient or historical remains, there are relatively few examples of aDNA methods applied to microbial samples. Microbial aDNA studies have mostly focused on pathogens such as *Mycobacterium leprae* and *Salmonella enterica* in human remains [32–34], the human oral microbiome [35,36], *Phytophtora infestans* in plants

[37,38], or on the general microbiome of archaeological specimens [30,39,40]. Studying microbiomes of archaeological specimens is challenging, as it is often difficult to discern pre-mortem microbiota from post-mortem communities [6,40–42]. When investigating pathogenic microbes, this risk is lower, as often the specific pathogen is known, and unlikely to have been introduced post-mortem. However, the identification of symptoms on ancient specimens is necessary. This is a non-trivial task which requires expert knowledge, with the difficulty compounded by the fact that pathogens may not leave identifiable signs or symptoms on certain preserved tissue types (e.g. bone), or may even kill their host before symptoms develop. Symbiotic microorganisms, however, are often known beforehand, making detection easier. In many symbiotic systems, especially obligate symbioses, the presence of a symbiont is highly predictable based on taxonomic and collection information. Because mostly plant aerial organs are stored in herbaria, microbial DNA from symbioses affecting leaves or flowers is more likely to be preserved.

Stable associations between plants and microbes in aerial organs are rare, and leaf symbioses represent perhaps some of the most intimate examples. Leaf symbioses affect numerous species of the Rubiaceae and Primulaceae, and bacteria of the *Burkholderiaceae* family of β-Proteobacteria [43]. In these associations, bacteria reside extracellularly in specialised leaf structures called leaf glands or nodules. The symbionts are generally vertically transmitted, and can produce secondary metabolites such as the insecticidal kirkamide or the herbicidal streptol-glucoside [44,45]. Detection of *Burkholderia* symbionts in herbarium specimens by PCR has been reported previously, showing that bacterial DNA can be preserved in leaf nodules [46,47]. Recently, we described a symbiotic system that shares many of the features of leaf nodule symbioses, i.e. the association between the monocot *Dioscorea sansibarensis* and the bacterium *Orrella dioscoreae* [48]. Also here, the symbiont resides in specialised leaf structures, is vertically transmitted, and has a high investment in secondary metabolism [49]. Despite evidence of vertical transmission and in contrast to obligate symbionts of Rubiaceae and Primulaceae, the genome of *O. dioscoreae* does not display clear signs of genome erosion [48].

We took advantage here of herbarium specimens to study the *Dioscorea* – *Orrella* symbiosis and its evolution. We reconstructed *de novo* whole bacterial genomes from degraded bacterial aDNA present in the leaf glands, enabling detailed comparative genomics across a broad variety of preserved, historic and modern specimens across the natural range of *D. sansibarensis*. The high quality of our dataset and the low intraspecific diversity allowed us to document consistent, distinct damage patterns in the preserved DNA of hosts and symbionts which are independent of sample age or region of collection. Further, we show that the symbiosis with *O. dioscoreae* is ubiquitous and exclusive throughout the geographical range of *D. sansibarensis*, with co-phylogenetic patterns suggestive of a mixed vertical and horizontal mode of transmission.

Material & Methods

Sampling and DNA-extraction

Leaf glands of 36 herbarium specimens (Table 1, Figure 1) 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, Table 1), 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 PCR cabinet (AirClean 600 PCR Workstation, Starlab, Hamburg, Germany). All tools used were disinfected using bleach and/or followed a UV treatment prior to their usage. When possible, sterile disposable items were used. Total genomic DNA from leaf nodules was extracted using the protocol described in Gilbert *et al.* [50], which was found to perform well on botanical specimens [51,52]. 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) [53] and purified using a MinElute PCR purification kit (Qiagen) following the manufacturer's recommendation.

Library preparation and sequencing

Genomic libraries adapted for ancient DNA were constructed following the double-stranded protocol from Wales *et al*. [54], and using the adapters described in Meyer & Kircher [55]. 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). DNA libraries were quantified using either a Quantus (Promega, Madison, WI, USA) or Qubit (Invitrogen, Carlsbad, CA, USA) fluorometers with respective dsDNA kits, and amplified using PCR. Libraries were pooled in equimolar concentrations and sequenced at the National High-throughput DNA Sequencing Centre, Copenhagen, Denmark (samples Herb1 to 10, Table 1) or at the Wellcome Trust Human Center for Human Genetics, Oxford, UK (samples MK001-MK026, Table 1), using Illumina technology, single-end 80 bp reads. Raw sequencing reads were deposited in the SRA archive under bioproject PRJNA646369.

Read processing and mapping

Sequencing adapters were removed using Cutadapt v2.10 [56], and low-quality bases were removed using Trimmomatic v0.39 [57]. Smalt v0.7.6 [58] and BEDTools v2.27.1 [59] were used for mapping and coverage estimations, respectively. Reads were mapped to the *Dioscorea sansibarensis* chloroplast sequence from a plant from the Botanical Garden of Ghent University (NCBI accession GCA_900631875.1) and its associated *Orrella dioscoreae* LMG 29303^T genome (NCBI accession GCA_900089455.2). To estimate the diversity of microbes in the glands, and detect possible contamination, Metaphlan 3 [60] and Kraken v1.1.1 [61] (using a custom database of bacterial and chloroplast sequences [48]) were used. Samples with low amounts of *O. dioscoreae* content were further analysed using Blastn [62] against the NCBI nt database (accessed 03/2020).

Genome assembly

De novo assembly of bacterial genomes was performed in 2 steps using SPAdes v3.14 [63] as described previously [49]. 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 identified based on base composition (% G+C) and average coverage (Figure S1). Reads used in the assembly of these contigs were extracted and reassembled using SPAdes v3.4 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 [61] and BASTA [64]. Quast v5.0.2 [65] was used to determine quality of assembly, and BUSCO v4.0.6 [66] was used to asses genome completeness using the Burkholderiales conserved marker set. The herbarium metagenome-assembled genomes are submitted to Zenodo (DOI: 10.5281/zenodo.3946545).

DNA damage analysis

Trimmed reads of each sample were mapped to the *Orrella dioscoreae* LMG 29303^T reference genome and the aforementioned *Dioscorea sansibarensis* plastome using bwa-mem [67]. Duplicates were removed using Samtools MarkDup [68]. The resulting alignments were used as input for MapDamage 2 [31] to estimate DNA damage patterns. Only samples with average coverage above 5x for both plastid and symbiont genomes were considered for further analysis (17 samples). The assessed patterns were the amount of C-to-T mutations on the first base of the reads, and 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). To assess the influence of DNA-methylation on strand breakage, methylated sites in the *Orrella dioscoreae* LMG 29303^T reference genome were predicted using PacBio long reads obtained as part of the genome sequencing effort [49] and the RS_Modification_and_Motif_Analysis protocol of the SMRT Analysis software suite v. 2.3.0 (PacBio, Menlo Park, CA, USA). The amount of data included in the analysis amounted to a total of about 1.25 Gb, for an average coverage of the reference genome of 114x. This strategy allowed for the detection of m4C (a vast majority of the detected modifications) and m6A modified bases. To assess if methylation has an effect on strand breakage, the amount of 5’ read ends mapping in a neighbourhood of up to 5 bases up- and downstream of methylated sites in the *O. dioscoreae* reference genome was calculated using BedTools [59]. This was performed for four samples with high coverage (MK003, MK005, MK019, MK025).

Phylogeny & comparative genomics

SNP-based phylogenies of the herbarium specimens, previously sequenced fresh leaf glands collected in Madagascar [49], 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 (with settings -polyThreshold 0.9, -gapThreshold 0.2 -perBaseCov 5 (3 for the chloroplast)) [69] to create reference alignments of all samples including the *Dioscorea sansibarensis* chloroplast and the *Orrella dioscoreae* LMG 29303^T genome. For the chloroplast alignment, samples where less than 40% of reference bases could be covered with >5x coverage were discarded to increase the robustness of the phylogeny. Phylogenetic trees based on plastid data were constructed using PhyML v3.3.3 [70] using the F81 model, 100 bootstrap replicates and the plastid sequence of *Dioscorea elephantipes* (NCBI accession NC_009601) as outgroup. Symbiont trees were constructed using FastTree [71] with the GTR model and the genome sequence of *Achromobacter xylosoxidans* ADAF13 (NCBI accession GCA_001566985) as outgroup. A haplotype network of the plastid sequences was created with TCS [72] using the SNP-based alignment used for the phylogeny.

Average Nucleotide Identity (ANI) values between genomes were calculated using PyANI v0.3 [73]. Orthologs between herbarium genomes, genomes assembled from fresh glands [49], and the specimen from Meise, were predicted using Orthofinder v2.3.9 [74]. A core-genome phylogeny was constructed by aligning the protein sequences of the single-copy core genes using Muscle v3.8.31 [75], back-translating the alignments into nucleotide sequences using T-Coffee v12 and concatenating [76]. The concatenated alignment was then used to construct a maximum likelihood phylogeny using RAxML v8.2.12 [77] (rapid bootstrapping and best-scoring ML mode, using 100 bootstrap replicates and the GTRGAMMA substitution model). Patterns of gene gain and loss were computed based on the gene presence/absence output of Orthofinder, using the Dollo analysis implement in Count [78]. Only non-redundant genomes (genomes <99% identical) were used in this analysis. 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.

Age estimation of the common ancestor of all investigated specimens was performed using BEAST v1.10.4 [79] based on Viruel *et al.* [80], and as described before [49]. Gene alignments for three chloroplast genes (*matK*, *rbcL*, *atpB*) were constructed using the sequences of

Dioscorea species described in[80], 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. The same parameters and calibration points as described in [49] and [80] were used to run the dating analysis.

Python 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.ugent.be/brdannee/DioscoreaHerbarium

Results

Recovery of DNA from preserved Dioscorea sansibarensis *leaf glands*

Leaf glands from herbaria had an average weight of 6.9 mg and varied in size from 1.4 mg to 18.4 mg. (Table 2). Yields from DNA extraction also greatly varied, with an average of 1.15 µg DNA recovered (Table 2), and ranging from 1 ng up to 5.5 µg. 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).

Characteristics of the sequencing libraries are consistent with historic DNA specimens

The number of sequencing reads was also highly variable for each library within a sequencing run (Table 2). One library failed (MK021), some showed extremely low yields (MK001: 74 thousand reads; MK022: 13.5 thousand reads) while others had over 30 million reads (MK011: 36.8 million; MK003: 40 million). There was no correlation between the age of the specimens and the number of reads produced during sequencing, even when accounting for the different sequencing run (Spearman correlation *p*-value > 0.1). As expected for highly degraded DNA, 83.4 % of reads were shorter than the maximum read length of 80 bp. On average, adapter-trimmed reads were 53 bp long, with shorter reads mostly occurring in specimens with low sequencing output. In total, an average of 35% of raw bases originated from sequencing adapters.

Taxonomic composition of D. sansibarensis *leaf glands*

On average 66% of reads per sample mapped to the *O. dioscoreae* LMG 29303^T reference genome. The proportion of *O. dioscoreae* reads in samples MK001, MK010, MK018, and Herb2 was significantly lower, with an average of 50% of reads or less (Table 3). Because the percentage of mapped reads could be influenced by the degree of divergence to the reference sequence, we also classified sequencing reads using a more robust blastn search against the NCBI nucleotide database (Figure S2). With this approach, 26% of MK001 reads were classified as *O. dioscoreae*, while another 26% matched *Viridiplantae* sequences. The remaining reads were classified as other bacteria (of which 17% were γ-Proteobacteria and 9% were α-proteobacteria). Furthermore, 14% accounted for various eukaryotes, of which 6% matched with human DNA. Samples MK010 and MK018 in particular contained high proportions of human contamination (85% and 43% respectively). Finally, 31% of Herb2 reads matched the reference *O. dioscoreae* sequence, whereas 16% matched with *Viridiplantae* sequences. The remaining reads were dominated by α-proteobacterial and Actinobacterial sequences (23% and 12% respectively). Because shotgun read abundance may not accurately reflect cell numbers and spurious hits in the database may confuse the analysis, we used Metaphlan 3 to infer normalized abundance counts (Table S1). In all other samples, *O. dioscoreae* represented 100% of eubacterial DNA, except in MK010 and MK018. Sequences classified as *Cutibacterium acnes*, a human commensal, represented 41% and 17% of the bacterial relative abundances in samples MK010 and MK018, respectively, indicating possible post-mortem contamination. Due to these high levels of contamination, neither samples were used for further analysis. Similarly, we did not analyse further samples MK001 and MK024 due to a low overall sequencing yield.

DNA damage patterns vary between chloroplast and symbiont DNA

Assessment of DNA damage patterns in historical specimens is critical for validating their authenticity. Leaf glands of *D. sansibarensis* are populated by clonal bacteria [48] as well as plant cells and plastids. This within-sample homogeneity allowed us to test whether DNA degradation patterns or dynamics differ between microbial, plastid or nuclear DNA within the same historic specimen. We observed an average read length of 53 bp in our historical specimens, a degree of fragmentation that is similar to previously reported herbarium DNA [28,81,82]. We did not observe significant differences in read length between reads mapping to the chloroplast and reads mapping to the symbiont (Wilcoxon paired rank sum test *p*-value

> 0.1). 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 2). We observed a small, statistically significant increase in the average proportion of C-to-T mismatches between *O. dioscoreae* and chloroplast DNA (Figure S3a; paired t-test *p*-value < 0.05). This can be explained by the higher inter-sample diversity in *O. dioscoreae* sequences compared to plastids. This interpretation is confirmed by the fact that specimens phylogenetically further away from the reference tend to see an increase in background mismatches, which also translates to an increase in C-to-T conversions (Figure S5). The proportion of C-to-T mismatches showed significant correlation with the age of the specimens in both sources (Figure S4a; Pearson correlation *p*-values < 0.01).

Purines were enriched before strand breaks in *O. dioscoreae* and plastid DNA, a common feature of ancient DNA (Figure 2). Unexpectedly, the proportion of purines before strand breaks was larger in the *O. dioscoreae* genome compared to the *D. sansibarensis* plastome. (45% vs. 30% increase, Wilcoxon signed-rank test *p*-value < 0.005) (Figure S3b). More specifically, positions preceding strand breaks were more enriched in adenine in the symbiont than in plastid DNA (23% vs. 56% increase, Wilcoxon signed-rank test *p*-value < 0.001; Figure S3c-d), whereas the relative proportion of guanines remains unchanged (Wilcoxon signed-rank test *p*-value > 0.1). In addition, there was no correlation between specimen age and the relative increase in purines (Figure S4c-d; Pearson correlation *p*-values > 0.05). We wondered whether DNA base modifications could influence the relative proportions of purines before strand breaks. We examined whether strand breakage is influenced by DNA-methylation by examining the occurrence of strand breaks in four samples with high coverage around known methylation sites of the LMG29303^T reference genome. In all four samples, we observed a lower proportion of 5' ends mapping in the immediate vicinity of N^4-methyl cytosines (m4C) (Fig S5). However, the proportion of mapped 5' ends plateaus quickly further away from the m4C residue. The proportion of strand breakage at cytosine residues was less affected by the proximity of a m4C residue. We observed a higher proportion of purines before strand breaks, although strand breaks did not occur more often after adenosine compared to guanine residues. This indicates that base modifications may influence the rate of degradation in historical DNA, yet do not explain the different rates of strand breakage at purines when comparing plastid and bacterial genomes.

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

Most plastid sequences used in SNP-based alignment were nearly identical, resulting in a phylogenetic topology with very short branches (Figure 5). In contrast, the plastid sequence of the Herb2 sample is divergent from the rest and constitutes a basal branch in the phylogenetic tree, while the other samples appear to cluster together based on geographic origin (Figure 5). Samples collected from fresh glands in Madagascar all clustered together, and according to the sampling region, which is in concordance with what we previously described [49]. The herbarium specimens collected in continental Africa form a different clade in the phylogeny. Three main clusters can be distinguished: a group mainly comprising specimens from Tanzania, a group with specimens from DR Congo and São Tomé, and a group with specimens from both Tanzania and DR Congo. The genomes of the symbionts are more diverse, with 2 main clusters (Figure 5). In contrast to the chloroplast phylogeny, samples do not cluster together according to specimen location. For example, the closest relatives of the RAN3 sample from Madagascar are all herbarium specimens collected in DR Congo. The core-genome phylogeny of the symbiont is mostly congruent with the SNP based phylogeny, with the subdivision of the two big groups, and fresh-collected samples mixed with herbarium samples (Figure 6). Phylogenetic dating of the most recent common ancestor of the herbarium specimens and the Madagascar specimens revealed that the *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). This is slightly older than previously estimated based on the fresh specimens from Madagascar alone [49], estimated between 20 000 and 3.19 million years ago.

Nearly complete bacterial assemblies can be retrieved from herbarium specimens

We were able to produce *de novo O. dioscoreae* genome assemblies from 17 out of 36 herbarium specimens. In general, at least 30x coverage of the *Orrella dioscoreae* reference genome was required for constructing a *de novo* genome with > 98% BUSCO completeness from the short reads. Most genomes could be reconstructed in less than 100 contigs, and showed very similar sizes to the reference genome (4.7 to 5.2 Mbp) (Table 4). Whole genome alignment using Mauve showed high synteny, without large rearrangements. Average nucleotide identity (ANI) values confirmed that all symbiont genomes belonged to the same species, with a minimum of 96.02% ANI, well above the commonly accepted 95% threshold for species delineation [83]. Interestingly, two genomes from specimens collected 35 years apart in different phytoregions of the DRC were almost identical (Herb9 and MK003, 2 SNPs over the whole genome). 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 highly distributed biogeography of *O. dioscoreae* [49].

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

The total amount of predicted genes is approximately the same in all genomes (4300-4700, Table 4), 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 (herbarium & fresh). 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 S7). On average, each genome has 50 genes for which no orthologs were found in other genomes, while accessory genes shared between more than 5 genomes are rare. Using an analysis of gene gain and loss over time with the Dollo parsimony principle, we estimated an ancestral genome of 5116 genes (Figure 7). Thus, there is a general trend towards gene loss, with most lineages having lost on average 998 genes, while only gaining an average of 385 genes for a net gene loss of 614 genes per lineage. Gene loss seemed to occur mostly at random across lineages. Most frequently occurring patterns of gene loss involve long branches (e.g. in MK020), or genes that are specific to a certain (sub)group in the phylogeny. Most lost genes are hypothetical genes, and are lost as single genes or in small clusters, indicating that gene loss is unlikely to be adaptive. An exception is a large gene cluster that is lost in some lineages, is a cluster of 34 Type III secretion system genes. This cluster is present all genomes of the LMG 29303^T reference genome subgroup, but is lost multiple times in lineages of the other subgroup (Figure S8). In contrast, functions highly expressed in the *D. sansibarensis* leaf nodule and linked to specialized metabolism and type VI secretion are conserved in all *O. dioscoreae* genomes [49]. We also wondered if frequent host-switching, as evidenced by the incongruence between host and symbiont phylogenetic trees, would also be reflected in HGT of symbiotic factors. Interestingly, the phylogenetic trees of 10 genes that comprise one of the two Type VI secretion systems of O. dioscoreae are incongruent with the species tree. These 10 genes include the two putative VgrG-domain effector proteins of the cluster. In addition, a pair of Rhs/VgrG proteins putative T6SS effector proteins was encoded in all genomes of one of the phylogenetic clusters but not the other. Apart from those, additional Rhs and/or VgrG proteins domains were also detected in 4 other genomes (AMP9, BER1, BER2, and MK019).

Discussion

Herbarium specimens are an increasingly useful resource for studies of plant biology and evolution, including molecular techniques [84]. Here, we leverage herbarium specimens to gain novel insights into the genome evolution and transmission mode of the symbiosis between *Dioscorea sansibarensis* and its obligate symbiont *Orrella dioscoreae*. We could detect *O. dioscoreae* DNA in all successfully sequenced libraries, highlighting the ubiquity of the association in a broad cross-section of *D. sansibarensis'* range. Several factors influenced the amount of recovered symbiont reads. Sample complexity, such as high amounts of plant DNA or contaminants, resulted in smaller amounts of recovered microbial DNA. DNA quality also played a role: in highly degraded samples, reliably mapping reads to the reference genomes is more difficult due to their short length. Similarly, accurate taxonomic classification of reads is less reliable with shorter reads. Highly degraded samples MK001 and MK0024 did not yield usable gene marker sequences, and unbiased analysis of reads using yielded taxonomically ambiguous results. Interestingly, in sample MK001, 12% of the reads showed a best hit against the genome of *Pantoea stewartii*, a known plant pathogen which affects aerial tissues, including leaves [85]. Our sample predates the first detection of this species in Africa by more than 60 years (Benin and Togo [86,87]). However, many other *Pantoea* species are known to infect a range of plant species across the world [88]. Thus, the high level of *Pantoea* in this sample could indicate an infected state of the gland, demonstrating the potential of shotgun metagenomics of herbarium specimens to investigate plant diseases. In the Herb2 specimen, almost 40% of reads showed a best hit either against α-proteobacteria or Actinobacteria, without reliable taxonomic assignment at the species level. We cannot say if these sequences represent post-mortem contamination, spurious hits, or actual bacteria present in the leaf gland, but *O. dioscoreae* was still largely dominant. However, some of these species are also known to be common lab contaminants [89]. Interestingly, the two samples which showed excessive human contamination (MK010 and MK018) were specimens from the same collection (Ghesquière J. 2709). It is thus possible that specimens from this collection were contaminated during collection or preservation.

We compared the presence and severity of some commonly investigated DNA-damage patterns between plastid and bacterial DNA. Conversions of cytosine to uracil/thymidine on fragment ends is a clear signature of ancient DNA [26]. We observed elevated levels of these C-to-T conversions in herbarium specimens, validating the DNA as ancient. We also found that these conversions were slightly but significantly more present in *O. dioscoreae* than in the chloroplast of *D. sansibarensis*. This may however be the result of higher substitution rates in the symbiont genomes compared to the plastomes, resulting in elevated numbers of mismatches between the samples and reference genomes. This is an important factor to consider when performing aDNA studies, especially when working with non-model systems and/or species without reliable genome references. Similar to other studies, we found that the proportion of C-to-T conversions is correlated with specimen age [26,28,29]. The proportion of C-to-T conversions in our samples is similar to that of plant specimens of similar age [28,82], but also of the oomycete *Phytophtora infestans* collected from potato leaves [37,38], as well as preserved molluscs and primate specimens [25,90].

Both plastid and symbiont DNA sequences showed fragmentation similar to what has been described in most aDNA studies [25,26,28,82]. However, purines were significantly enriched before strands breaks in the symbiont DNA compared to the plastid DNA. This enrichment was solely due to an elevated relative abundance of adenosines before strand breaks, while the proportion of guanines remained unchanged. This purine bias has so far been scarcely discussed in the literature [25,26]. Sawyer and colleagues observed a shift in bias from adenosines to purines over time and hypothesized that enzymatic processes could be responsible, such as nuclease activity, and act differently on adenosines than on guanines [25]. However, they did not control for specimens of different origins, species, and conservation methods. In their review on ancient DNA damage, Dabney and colleagues [26] argued that differences in resonance structure between A and G could be responsible for this bias. Higher relative increase of guanines is indeed found in many specimens from the Pleistocene era [53,91,92], but not all [93]. All these data were gathered from mammalian specimens, mostly targeting mitochondrial DNA, which is generally AT-rich. Furthermore, purine enrichment varies greatly between specimens. In DNA from a Pleistocene horse bone preserved in permafrost, guanines were enriched over two-fold before strand breaks, while adenosine only increased 1.35 fold [91]. Specimens derived from Neanderthal, mammoth, and cave bear bones showed lower rates of purine enrichments (both adenosine and guanine between 1.1 and 1.3 fold increase [27]). While these differences could be attributed to different storage conditions, possibly affecting enzymatic processes, evidence from our specimens show that factors other than preservation method clearly play a role. As both symbiont and host DNA are preserved for the same time, in the same conditions, storage environment alone cannot explain the differences. Micro-environment (e.g. conditions in the gland, or within bacterial/plastid cells) or inherent differences in DNA content and/or structure (%G+C, methylation status, presence of histones etc…) could perhaps account for the difference in the chemistry of strand breaks in plant and bacteria in herbarium specimens.

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/country clustering together. Continental African specimens form a nested clade within specimens from Madagascar, which is in concordance with the earlier hypothesis that *D. sansibarensis* originated in Madagascar and was dispersed to Africa [80]. *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 [94] reported that no seed bearing plants had been seen among all the herbarium specimens collected in southern Africa, although they were occasionally encountered from elsewhere in Africa. This suggests that *O. dioscoreae* would be most likely to move between plants via bulbil-mediated vertical transmission [49]. 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 (Fig. 5, Fig. S9) 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 which is also congruent with the observed tree topology and haplotype network. However, specimen Herb2, collected in Cameroon, formed a divergent basal branch on the tree. Interestingly this herbarium specimen did not fully fit the taxonomic type of the species, and was identified as "*Dioscorea* cf. *sansibarensis*". These observations could 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 is interesting, as this indicates that the symbiosis might not be confined to the *D. sansibarensis* species and is possibly established much earlier than expected. Indeed, when adding this sample to our dating analysis, the estimated time of divergence of the symbiont-containing specimens is ca. 13 Mya, much older than when excluding the sample (ca. 3.3 Mya). This last estimate is in line with what we previously estimated based on fresh specimens from Madagascar alone [49]. Alternatively, Herb2 may represent a lineage in which the symbiont has been acquired independently, as it has been observed that other *Dioscorea* species can engage in leaf symbiotic interactions [95]. However, the observation that in the SNP-based phylogeny the Herb2 symbiont clusters with other symbionts from *D. sansibarensis* may suggest that instead horizontal transfer of the symbiont could have taken place.

Despite direct evidence of vertical transmission under laboratory conditions [49], the phylogenetic trees of *D. sansibarensis* and *O. dioscoreae* are incongruent. This confirms that a mixed transmission mode is likely a feature of this symbiosis [49]. Horizontal transmission, for example by insect vectors, could create the observed patterns. Acquisition from an environmental reservoir (e.g. the soil) seems unlikely, as we could not reliably detect the symbiont anywhere outside of the plant [49], but cannot be fully ruled out at this moment.

As the symbionts do not seem to co-evolve with their hosts, environmental selective pressures could play a role in the evolution of the symbiont. Functions implied to play a role in the symbiosis such as secondary metabolism and type VI secretion are conserved in all samples, further reinforcing their importance [49]. While the putative effectors within the two T6SS gene clusters are conserved, putative orphan effectors found elsewhere in the genome are not. While some spurious Rhs or VgrG domain hits could be found in some single genomes, we did find a combination of a Rhs and VgrG domain that is conserved in all genomes from one phylogenetic clade, but not in the other. As T6SSs play important roles in microbe-microbe interactions [96,97], this could indicate that T6SS effector inventories partially diverged in response to different threats from competitors. The fact that some genes of the second T6SS cluster in *O. dioscoreae* diverge greatly from the core genome phylogeny could indicate diversifying evolution, or more likely active horizontal gene transfer. This T6SS cluster may play a role in adaptation to the local environment and diverse threats, or could perhaps play a role in signalling and adaptation to a new host [98]. Another example of possible ongoing adaptation in the symbiont is the type III secretion system (T3SS). This cluster of 34 genes is conserved in one of the clades, but has been lost multiple times in the other clades (Figure S8). T3SSs are often used by pathogenic bacteria to inject effectors into eukaryotic hosts [97], but can also play a role in symbiosis [99]. However, genes of the T3SS of *O. dioscoreae* LMG29303^T were not upregulated *in planta* [49], suggesting that loss of T3SS genes is due to genetic drift rather than adaptive selection [100].

In general, *O. dioscoreae* genomes show an overall trend toward gene loss. While on average the core genome accounts for 78% of the gene complement in *O. dioscoreae*, the pangenome is large, being approximately twice the size of the core genome. The membership distribution of genes of the pan-genome is bimodal, with a strong bias towards genes only found in very few genomes. This could indicate that new genes can still be acquired, or more likely that genes affected by genetic drift are quickly purged [101]. Genome erosion is a feature commonly found in restricted symbionts, including leaf symbionts [46,102–106]. However, the genomes of *O. dioscoreae* do not display the hallmarks signs of genome reduction, such as accumulation of pseudogenes and IS elements [49]. Together, this indicates that *O. dioscoreae* may be undergoing the very early steps of genome streamlining. How the symbiont manages to escape the evolutionary rabbit hole of genome reduction remains unknown. The deleterious effects of excessive genetic drift could possibly be counteracted by avoiding stringent transmission bottlenecks, or through horizontal transfer and/or mixing of symbionts, at the cost of the eventual loss of symbiont effectiveness due to a proliferation of cheaters [104,107,108].

In conclusion, our data demonstrate that aDNA and metagenomics methods are a powerful combination to probe dynamic associations between plants and microorganisms from preserved 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 thought 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.

Table 1: Herbarium specimens metadata, as registered in the archives of the Meise Botanic Garden Herbarium.

Sample name	Herbarium Barcode	Collector	Collection Number	Collection Date	Country of collection	Collection details
Herb1	BR0000024463324	L. Pauwels	6041	05 DEC 1978	DR Congo	Zongo. Terr. Kasangulu
Herb2	BR0000024463003	R. Letouzey	13552	09 MAY 1975	Cameroon	Route Mamfe-Calabar; entre lac Fjagham et rivière Akegam (40km W. Mamfe) Mamfe
Herb3	BR0000024463461	Fred L. Hendrickx	5488	15 JUL 1948	DR Congo	Maniema ou culture à Mulungu
Herb4	BR0000024463171	R.B. Faden, S.M. Phillips, A.M. Muasya & E. Macha	96/12	01 JUN 1996	Tanzania	T3, Lushoto District. Western Usambara Mts., Mombo-Lushoto road, 3 km
Herb5	BR0000024462990	Joaquim Viegas do Graça do Espirito Santo	79	07 JAN 1949	São Tomé & Príncipe	Regio S. Tomé. S. Vicente
Herb6	BR0000024463034	Georges le Testu	4268	30 OCT 1922	Central African Republic	Yalinga. Dans la Haute-Kotto
Herb7	BR0000024463126	H.G. Faulkner	K.580	24 MAY 1950	Tanzania	Pangani District, Busheii Estate
Herb8	BR0000024463218	H.J. Schlieben	2063	09 APR 1932	Tanzania	Mahenge: Umgebung der Mahenge
Herb9	BR0000024463287	Flamigni	520	01 JUL 1913	DR Congo	Kito. Kitobola
Herb10	BR0000024463416	Em. & M. Laurent	-	11 DEC 1903	DR Congo	Près de Yumbi
MK001	BR0000024463249	H. Humbert	25450	15 FEB 1951	Madagascar	Bassin margen du Sambirano au S. de Marvato
MK002	BR0000024463409	S.C.	S.N.	-	DR Congo	-
MK003	BR0000024463423	J. Leonard	1914	17 SEP 1948	DR Congo	Yangambi, Service médical
MK004	BR0000024463454	D. Van der Ben	420	20 MAY 1953	DR Congo	Lac Albert. Mahagi-Port. Rivière Ori à la sortie de la montagne
MK005	BR0000024463485	Van Meel	872	-	-	-
MK006	BR0000024462983	J. Espirito Santo	79	07 JAN 1949	São Tomé & Príncipe	S. Vicente
MK007	BR0000024463294	Flamigni	520	17 APR 1913	DR Congo	Kito. Kitobola
MK008	BR0000024463256	C. Evrard	6998	-	DR Congo	Campus de Kinshasa. Territoire: Kinshasa
MK009	BR0000024463270	C. Evrard	6998	03 OCT 1973	DR Congo	Campus de Kinshasa. Territoire: Kinshasa

MK010	BR0000024463379	Ghesquière J.	2709	13 JUN 1936	DR Congo	Busiza
MK011	BR0000024463362	Body	471	01 JUL 1906	-	-
MK012	BR0000024463157	J.S. Bond	68	15 SEP 1969	Tanzania	T3. Pangani District. Kidifu
MK013	BR0000024463072	Haerdi	500/0	-	Tanzania	Itundufula oberhalb Kiberege, Mbangliau b/Mahenge, Itula b/Ifakara, in Dickichten des Hagelwaldes
MK014	BR0000024463102	G. de Nevers & S. Charnley	3230	11 APR 1984	Tanzania	Mikumi National Park; hills east of road; Miombo Woodland
MK015	BR0000024463430	Ph. Gerard	1465	09 JUL 1954	DR Congo	Tukpwo. Galerie de la Magidi
MK016	BR0000024463263	H. Breyne	-	28 APR 1971	DR Congo	Sao. Territoire: Maluku
MK017	BR0000024463058	M. Batty	1045	11 APR 1970	Tanzania	Morogoro. Kidatu
MK018	BR0000024463355	Ghesquière J.	2709	-	DR Congo	Busiza
MK019	BR0000024463133	M. Batty	1044	11 APR 1970	Tanzania	Morogoro Dist., Kidatu
MK020	BR0000024463331	F. Demeuse	106bis	02 NOV 1888	DR Congo	-
MK021	BR0000024463201	H.J. Schlieben	6133	17 MAR 1935	Tanzania	Lindi: 60 km W Lindi
MK022	BR0000024462976	M. Mathieu	-	-	Senegal	Marovoay
MK023	BR0000024463164	H.J. Schlieben	2095	18 APR 1932	Tanzania	Mahenge: Umgebung der Mahenge
MK024	BR0000024463188	E.M. Bruce	1040	15 APR 1935	Tanzania	Ulugurus, Kimbosa
MK025	BR0000024463195	R.E.S. Tanner	3515	27 MAY 1957	Tanzania	Pangani Dist., Tassini. Tanga Prov., Pangani Dist., Madanga, Tassini
MK026	BR0000024463119	H.G. Faulkner	580	-	-	-

577

Table 2: DNA-extraction and sequencing yields

Sample name	Nodule mass (mg)	DNA yield (ng)	Specimen age (years)	Country of collection	Sequencing Center	Sequencing yields (million reads)	Mean trimmed read length (bp)	Retained bases after trimming (%)	Mean read length (bp)
Herb1	2.2	219.8	41	DR Congo	Copenhagen, DK	1.17	49.31	61.51	55.77
Herb2	3.0	7.1	44	Cameroon	Copenhagen, DK	2.36	55.66	69.47	59.98
Herb3	3.7	1722.0	71	DR Congo	Copenhagen, DK	0.85	51.23	63.92	58.09
Herb4	5.2	3640.0	23	Tanzania	Copenhagen, DK	4.43	64.85	80.90	64.56
Herb5	6.5	2576.0	70	São Tomé & Príncipe	Copenhagen, DK	4.09	71.29	88.96	72.73
Herb6	7.8	1078.0	97	Central African Republic	Copenhagen, DK	0.70	51.72	64.56	60.90
Herb7	10.0	959.0	69	Tanzania	Copenhagen, DK	1.45	59.44	74.18	63.52
Herb8	10.4	3395.0	87	Tanzania	Copenhagen, DK	0.64	51.27	63.97	58.59
Herb9	14.4	303.1	106	DR Congo	Copenhagen, DK	4.16	71.10	88.73	71.95
Herb10	18.4	1470.0	116	DR Congo	Copenhagen, DK	0.43	39.69	49.54	47.27
MK001	1.4	163.8	68	Madagascar	Oxford, UK	0.07	37.60	46.12	40.83
MK002	13.8	35.7		DR Congo	Oxford, UK	19.62	50.07	62.21	54.67
MK003	1.5	139.7	71	DR Congo	Oxford, UK	40.10	48.19	59.87	49.99
MK004	4.4	910.0	66	DR Congo	Oxford, UK	0.31	40.17	47.60	46.81
MK005	1.6	27.6			Oxford, UK	28.22	43.91	54.47	49.54
MK006	2.6	455.0	70	São Tomé & Príncipe	Oxford, UK	1.57	48.31	59.95	50.76
MK007	3.5	220.5	106	DR Congo	Oxford, UK	1.01	43.16	53.61	46.31
MK008	13.0	252.0		DR Congo	Oxford, UK	1.25	45.23	56.17	51.26
MK009	5.6	241.5	46	DR Congo	Oxford, UK	0.42	40.01	46.84	47.33
MK010	8.9	4.2	83	DR Congo	Oxford, UK	6.76	58.79	73.05	55.39

MK011	2.9	72.5	113		Oxford, UK	36.84	45.75	56.74	55.41
MK012	2.1	5250.0	50	Tanzania	Oxford, UK	20.86	62.42	77.64	63.35
MK013	1.6	945.0		Tanzania	Oxford, UK	23.64	58.28	72.10	59.57
MK014	6.3	1960.0	35	Tanzania	Oxford, UK	13.38	62.01	77.12	63.08
MK015	8.0	136.5	65	DR Congo	Oxford, UK	26.35	50.09	62.24	57.95
MK016	12.5	2870.0	106	DR Congo	Oxford, UK	19.09	55.20	68.60	61.12
MK017	4.1	1050.0	49	Tanzania	Oxford, UK	27.77	54.02	67.11	55.74
MK018	10.3	0.9		DR Congo	Oxford, UK	5.37	53.68	66.66	55.24
MK019	7.0	2345.0	49	Tanzania	Oxford, UK	23.94	54.41	67.56	57.34
MK020	13.3	118.3	131	DR Congo	Oxford, UK	28.56	57.34	71.24	58.94
MK022	3.5	287.0		Senegal	Oxford, UK	0.01	49.31	61.33	52.87
MK023	5.4	665.0	87	Tanzania	Oxford, UK	15.04	60.52	75.25	60.58
MK024	6.2	26.6	84	Tanzania	Oxford, UK	0.30	36.10	44.47	43.16
MK025	3.3	1540.0	62	Tanzania	Oxford, UK	22.60	51.71	64.27	53.37
MK026	18.2	5530.0			Oxford, UK	22.94	66.19	82.28	66.69

Table 3: Results from read mapping to the *D. sansibarensis* chloroplast sequence and *O. dioscoreae* reference genome.

	O. dioscoreae				***D. sansibarensis* chloroplast**			
Sample name	**Mapped Reads (million reads)**	**Mean read length (bp)**	**Proportion of total (%)**	**Mean genomic coverage (X)**	**Mapped Reads (thousand reads)**	**Mean read length (bp)**	**Proportion of total (%)**	**Mean genomic coverage (X)**
Herb1	0.90	50.30	77.16	7.58	1.48	55.77	0.13	0.49
Herb2	0.44	56.54	18.65	2.70	33.97	59.98	1.44	13.39
Herb3	0.67	52.28	78.37	5.98	1.28	58.09	0.15	0.45
Herb4	3.43	65.49	77.68	39.91	21.64	64.56	0.49	8.75
Herb5	2.59	71.93	63.34	32.63	40.69	72.73	1.00	18.73
Herb6	0.60	52.34	86.30	5.44	0.24	60.90	0.03	0.07
Herb7	0.79	59.96	54.53	7.92	1.36	63.52	0.09	0.51
Herb8	0.50	52.06	78.89	4.42	0.23	58.59	0.04	0.06
Herb9	3.34	72.03	80.29	42.25	16.15	71.95	0.39	6.98
Herb10	0.33	39.23	77.59	2.13	2.46	47.27	0.57	0.72
MK001	0.01	43.72	8.55	0.04	1.90	40.83	2.61	0.59
MK002	16.44	50.65	84.30	167.63	12.04	54.67	0.06	4.02
MK003	27.06	49.18	67.90	264.31	350.68	49.99	0.88	130.24
MK004	0.19	43.04	62.52	1.59	0.75	46.81	0.25	0.26
MK005	18.54	44.24	66.18	163.15	530.65	49.54	1.89	198.95
MK006	0.77	49.71	48.97	7.24	13.11	50.76	0.84	4.79
MK007	0.75	43.99	74.34	6.57	4.25	46.31	0.42	1.45
MK008	1.09	45.77	87.59	10.11	2.13	51.26	0.17	0.75
MK009	0.28	42.99	70.00	2.40	1.58	47.33	0.40	0.55
MK010	0.31	52.61	4.59	1.95	37.46	55.39	0.56	7.99
MK011	20.81	43.61	56.93	179.45	277.79	55.41	0.76	103.99

MK012	16.65	63.04	80.21	211.93	57.02	63.35	0.27	23.71
MK013	15.32	59.75	65.49	184.44	130.53	59.57	0.56	52.70
MK014	9.69	62.75	72.74	121.75	74.16	63.08	0.56	32.39
MK015	20.12	51.06	76.82	202.60	33.94	57.95	0.13	12.51
MK016	16.58	55.81	87.38	186.05	15.12	61.12	0.08	5.14
MK017	19.30	54.99	69.93	209.64	179.76	55.74	0.65	72.66
MK018	1.88	46.01	35.30	15.58	28.81	55.24	0.54	8.67
MK019	21.12	54.93	88.84	236.03	26.01	57.34	0.11	9.24
MK020	22.96	58.53	80.88	264.85	80.84	58.94	0.28	33.92
MK022	0.01	49.77	77.64	0.10	0.06	52.87	0.47	0.02
MK023	11.99	61.15	80.10	146.98	46.47	60.58	0.31	18.60
MK024	0.21	37.22	71.11	1.56	0.30	43.16	0.10	0.09
MK025	19.03	52.16	84.70	202.22	67.51	53.37	0.30	24.93
MK026	11.01	67.44	48.23	136.54	18.05	66.69	0.08	7.68

Table 4: Genome statistics of reference genome (LMG 29303^T), symbionts from fresh collected specimens from Madagascar[49], and symbionts from herbarium specimens.

Assembly	Source	Nr. of contigs	Assembly length (bp)	Largest contig (bp)	N50 (bp)	GC (%)	Nr. of genes
LMG 29303	Typestrain	1	4848101	4848101	4848101	67.43	4361
R-67584	Botanical Garden Meise	22	4750654	1321707	810799	67.46	4636
AMB3	Sampling Madagascar	49	5042305	758077	168034	67.37	4536
AMP9	Sampling Madagascar	33	5239569	684227	355776	67.17	4704
ANDA3	Sampling Madagascar	31	5095598	575995	280825	67.15	4609
ANDO1	Sampling Madagascar	28	4985216	545984	283023	67.21	4485
ANDO2	Sampling Madagascar	19	4985273	800603	455071	67.21	4488
ANK1	Sampling Madagascar	13	4916307	953346	707415	67.62	4474
ANK2	Sampling Madagascar	8	4923321	1670418	828734	67.6	4475
ANT2	Sampling Madagascar	44	5127454	1151522	458767	67.21	4631
ANT3	Sampling Madagascar	34	5151797	1151522	458767	67.15	4643
BER1	Sampling Madagascar	38	5156740	637626	459432	67.15	4598
BER2	Sampling Madagascar	43	5152454	674381	245878	67.16	4600
BRI6	Sampling Madagascar	78	5304677	343171	130052	67.18	4812
BRI9	Sampling Madagascar	75	5304830	469735	148617	67.18	4812
DAR2	Sampling Madagascar	34	5069858	873190	222944	67.14	4588
DAR3	Sampling Madagascar	30	5070116	875833	280825	67.14	4590
ISE1	Sampling Madagascar	31	5004706	683316	275927	67.3	4496
ISE2	Sampling Madagascar	42	5005310	820128	220477	67.3	4493
IVO3	Sampling Madagascar	74	5304581	469351	164009	67.18	4811
RAN3	Sampling Madagascar	52	4950690	550950	167320	67.39	4479
RAN7	Sampling Madagascar	85	5303384	437753	127654	67.19	4802
Herb4	Herbarium	83	4990371	433249	118222	67.25	4608
Herb5	Herbarium	84	4777030	322322	127023	67.51	4385
Herb9	Herbarium	66	4855410	663656	144429	67.43	4460
MK002	Herbarium	79	5010997	346863	141145	67.23	4464
MK003	Herbarium	87	4846400	327900	134388	67.46	4572
MK005	Herbarium	109	4973616	372040	105137	67.35	4564
MK011	Herbarium	94	4965225	680317	105998	67.13	4762
MK012	Herbarium	75	5043392	717518	165392	67.01	4638
MK014	Herbarium	61	5023193	420498	157880	67.3	4627
MK015	Herbarium	88	5022797	568010	106353	67.33	4619
MK016	Herbarium	52	4972869	395858	172976	67.47	4623
MK017	Herbarium	86	5061349	392001	165944	67.3	4416
MK019	Herbarium	61	4845358	395326	129604	67.31	4446
MK020	Herbarium	68	4818048	401954	134771	67.41	4573
MK023	Herbarium	63	4970086	572808	177270	67.39	4386
MK025	Herbarium	57	4758493	376489	160045	67.46	4582
MK026	Herbarium	59	4963154	715678	242657	67.19	4312

References

1. Besnard, G., Gaudeul, M., Lavergne, S., Muller, S., Rouhan, G., Sukhorukov, A.P., Vanderpoorten, A., and Jabbour, F. (2018). Herbarium-based science in the twenty-first century. Bot. Lett. *165*, 323–327.

2. Bieker, V.C., and Martin, M.D. (2018). Implications and future prospects for evolutionary analyses of DNA in historical herbarium collections. Bot. Lett. *165*, 409–418.

3. Olofsson, J.K., Cantera, I., Van de Paer, C., Hong-Wa, C., Zedane, L., Dunning, L.T., Alberti, A., Christin, P.A., and Besnard, G. (2019). Phylogenomics using low-depth whole genome sequencing: A case study with the olive tribe. Mol. Ecol. Resour. *19*, 877–892.

4. Konrade, L., Shaw, J., and Beck, J. (2019). A rangewide herbarium-derived dataset indicates high levels of gene flow in black cherry (Prunus sclerotina). Ecol. Evol. *9*, 975–985.

5. Inglis, P.W., Pappas, M. de C.R., Resende, L. V., and Grattapaglia, D. (2018). Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. PLoS One *13*, e0206085.

6. Kistler, L., Bieker, V.C., Martin, M.D., Pedersen, M.W., Madrigal, J.R., and Wales, N. (2020). Ancient Plant Genomics in Archaeology, Herbaria, and the Environment. Annu. Rev. Plant Biol. *71*, annurev-arplant-081519-035837.

7. Zeng, C.X., Hollingsworth, P.M., Yang, J., He, Z.S., Zhang, Z.R., Li, D.Z., and Yang, J.B. (2018). Genome skimming herbarium specimens for DNA barcoding and phylogenomics. Plant Methods *14*, 43.

8. Higuchi, R., Bowman, B., Freiberger, M., Ryder, O.A., and Wilson, A.C. (1984). DNA sequences from the quagga, an extinct member of the horse family. Nature *312*, 282–284.

9. Miller, W., Drautz, D.I., Ratan, A., Pusey, B., Qi, J., Lesk, A.M., Tomsho, L.P., Packard, M.D., Zhao, F., Sher, A., *et al.* (2008). Sequencing the nuclear genome of the extinct woolly mammoth. Nature *456*, 387–390.

10. Mascher, M., Schuenemann, V.J., Davidovich, U., Marom, N., Himmelbach, A., Hübner, S., Korol, A., David, M., Reiter, E., Riehl, S., *et al.* (2016). Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley. Nat. Genet. *48*, 1089–1093.

11. Wales, N., Ramos Madrigal, J., Cappellini, E., Carmona Baez, A., Samaniego Castruita, J.A., Romero-Navarro, J.A., Carøe, C., Ávila-Arcos, M.C., Peñaloza, F., Moreno-Mayar, J.V., *et al.* (2016). The limits and potential of paleogenomic techniques for reconstructing grapevine domestication. J. Archaeol. Sci. *72*, 57–70.

12. Vallebueno-Estrada, M., Rodríguez-Arévalo, I., Rougon-Cardoso, A., González, J.M., Cook, A.G., Montiel, R., and Vielle-Calzada, J.P. (2016). The earliest maize from san marcos tehuacán is a partial domesticate with genomic evidence of inbreeding. Proc. Natl. Acad. Sci. U. S. A. *113*, 14151–14156.

13. Woodward, S.R., Weyand, N.J., and Bunnell, M. (1994). DNA sequence from cretaceous period bone fragments. Science (80-.). *266*, 1229–1232.

14. Golenberg, E.M., Giannasi, D.E., Clegg, M.T., Smiley, C.J., Durbin, M., Henderson, D., and Zurawski, G. (1990). Chloroplast DNA sequence from a Miocene Magnolia species. Nature *344*, 656–658.

15. Stankiewicz, B.A., Poinar, H.N., Briggs, D.E.G., Evershed, R.P., and Poinar, J. (1998). Chemical

preservation of plants and insects in natural resins. Proc. R. Soc. B Biol. Sci. *265*, 641–647.

16. Shapiro, B., and Hofreiter, M. (2012). Ancient DNA:methods and protocols B. Shapiro and M. Hofreiter, eds. (Totowa, NJ: Humana Press).

17. Weiß, C.L., Dannemann, M., Prüfer, K., and Burbano, H.A. (2015). Contesting the presence of wheat in the british isles 8,000 years ago by assessing ancient DNA authenticity from low-coverage data. Elife *4*.

18. Paabo, S. (1989). Ancient DNA: Extraction, characterization, molecular cloning, and enzymatic amplification. Proc. Natl. Acad. Sci. U. S. A. *86*, 1939–1943.

19. Cooper, A. (2000). Ancient DNA: Do It Right or Not at All. Science (80-.). *289*, 1139b – 1139.

20. Gilbert, M.T.P., Bandelt, H.J., Hofreiter, M., and Barnes, I. (2005). Assessing ancient DNA studies. Trends Ecol. Evol. *20*, 541–544.

21. Fulton, T.L., and Shapiro, B. (2019). Setting up an ancient DNA laboratory. In Methods in Molecular Biology (Humana Press, New York, NY), pp. 1–13.

22. Lindahl, T. (1993). Instability and decay of the primary structure of DNA. Nature *362*, 709–715.

23. Allentoft, M.E., Collins, M., Harker, D., Haile, J., Oskam, C.L., Hale, M.L., Campos, P.F., Samaniego, J.A., Gilbert, T.P.M., Willerslev, E., *et al.* (2012). The half-life of DNA in bone: Measuring decay kinetics in 158 dated fossils. Proc. R. Soc. B Biol. Sci. *279*, 4724–4733.

24. Kistler, L., Ware, R., Smith, O., Collins, M., and Allaby, R.G. (2017). A new model for ancient DNA decay based on paleogenomic meta-analysis. Nucleic Acids Res. *45*, 6310–6320.

25. Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. PLoS One *7*, e34131.

26. Dabney, J., Meyer, M., and Pääbo, S. (2013). Ancient DNA damage. Cold Spring Harb. Perspect. Biol. *5*.

27. Briggs, A.W., Stenzel, U., Johnson, P.L.F., Green, R.E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M., *et al.* (2007). Patterns of damage in genomic DNA sequences from a Neandertal. Proc. Natl. Acad. Sci. U. S. A. *104*, 14616–14621.

28. 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.

29. Staats, M., Cuenca, A., Richardson, J.E., Ginkel, R.V. van, Petersen, G., Seberg, O., and Bakker, F.T. (2011). DNA damage in plant herbarium tissue. PLoS One *6*, e28448.

30. Weiß, C.L., Gansauge, M.-T., Aximu-Petri, A., Meyer, M., and Burbano, H.A. (2019). Mining ancient microbiomes using selective enrichment of damaged DNA molecules. bioRxiv, 397927.

31. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., and Orlando, L. (2013). MapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. In Bioinformatics (Narnia), pp. 1682–1684.

32. Schuenemann, V.J., Singh, P., Mendum, T.A., Krause-Kyora, B., Jager, G., Bos, K.I., Herbig, A., Economou, C., Benjak, A., Busso, P., *et al.* (2013). Genome-Wide Comparison of Medieval and Modern Mycobacterium leprae. Science (80-.). *341*, 179–183.

33. Bos, K.I., Harkins, K.M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S.A., Bryant, J.M., Harris, S.R., Schuenemann, V.J., *et al.* (2014). Pre-Columbian mycobacterial genomes

reveal seals as a source of New World human tuberculosis. Nature *514*, 494–497.

34. Zhou, Z., Lundstrøm, I., Tran-Dien, A., Duchêne, S., Alikhan, N.F., Sergeant, M.J., Langridge, G., Fotakis, A.K., Nair, S., Stenøien, H.K., *et al.* (2018). Pan-genome Analysis of Ancient and Modern Salmonella enterica Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia. Curr. Biol. *28*, 2420-2428.e10.

35. Weyrich, L.S., Duchene, S., Soubrier, J., Arriola, L., Llamas, B., Breen, J., Morris, A.G., Alt, K.W., Caramelli, D., Dresely, V., *et al.* (2017). Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. Nature *544*, 357–361.

36. Jensen, T.Z.T., Niemann, J., Iversen, K.H., Fotakis, A.K., Gopalakrishnan, S., Vågene, Å.J., Pedersen, M.W., Sinding, M.H.S., Ellegaard, M.R., Allentoft, M.E., *et al.* (2019). A 5700 year-old human genome and oral microbiome from chewed birch pitch. Nat. Commun. *10*, 5520.

37. Martin, M.D., Cappellini, E., Samaniego, J.A., Zepeda, M.L., Campos, P.F., Seguin-Orlando, A., Wales, N., Orlando, L., Ho, S.Y.W., Dietrich, F.S., *et al.* (2013). Reconstructing genome evolution in historic samples of the Irish potato famine pathogen. Nat. Commun. *4*, 2172.

38. 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 *2013*.

39. Warinner, C., Herbig, A., Mann, A., Fellows Yates, J.A., Weiß, C.L., Burbano, H.A., Orlando, L., and Krause, J. (2017). A Robust Framework for Microbial Archaeology. Annu. Rev. Genomics Hum. Genet. *18*, 321–356.

40. Philips, A., Stolarek, I., Kuczkowska, B., Juras, A., Handschuh, L., Piontek, J., Kozlowski, P., and Figlerowicz, M. (2017). Comprehensive analysis of microorganisms accompanying human archaeological remains. Gigascience *6*.

41. Pedersen, M.W., Overballe-Petersen, S., Ermini, L., Der Sarkissian, C., Haile, J., Hellstrom, M., Spens, J., Thomsen, P.F., Bohmann, K., Cappellini, E., *et al.* (2015). Ancient and modern environmental DNA. Philos. Trans. R. Soc. B Biol. Sci. *370*, 20130383.

42. Bieker, V.C., Sánchez Barreiro, F., Rasmussen, J.A., Brunier, M., Wales, N., and Martin, M.D. (2020). Metagenomic analysis of historical herbarium specimens reveals a postmortem microbial community. In Molecular Ecology Resources (John Wiley & Sons, Ltd), pp. 1755–0998.13174.

43. Pinto-Carbó, M., Gademann, K., Eberl, L., and Carlier, A. (2018). Leaf nodule symbiosis: function and transmission of obligate bacterial endophytes. Curr. Opin. Plant Biol. *44*, 23–31.

44. Sieber, S., Carlier, A., Neuburger, M., Grabenweger, G., Eberl, L., and Gademann, K. (2015). Isolation and Total Synthesis of Kirkamide, an Aminocyclitol from an Obligate Leaf Nodule Symbiont. Angew. Chemie - Int. Ed. *54*, 7968–7970.

45. Pinto-Carbó, M., Sieber, S., Dessein, S., Wicker, T., Verstraete, B., Gademann, K., Eberl, L., and Carlier, A. (2016). Evidence of horizontal gene transfer between obligate leaf nodule symbionts. ISME J. *10*, 2092–2105.

46. Van Oevelen, S., De Wachter, R., Vandamme, P., Robbrecht, E., and Prinsen, E. (2002). Identification of the bacterial endosymbionts in leaf galls of Psychotria (Rubiaceae, angiosperms) and proposal of “Candidatus Burkholderia kirkii” sp. nov. Int. J. Syst. Evol. Microbiol. *52*, 2023–2027.

47. Verstraete, B., Janssens, S., Lemaire, B., Smets, E., and Dessein, S. (2013). Phylogenetic

lineages in Vanguerieae (Rubiaceae) associated with Burkholderia bacteria in sub-Saharan Africa. Am. J. Bot. *100*, 2380–2387.

48. 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.

49. 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., 1.

50. Gilbert, M.T.P., Wilson, A.S., Bunce, M., Hansen, A.J., Willerslev, E., Shapiro, B., Higham, T.F.G., Richards, M.P., O'Connell, T.C., Tobin, D.J., *et al.* (2004). Ancient mitochondrial DNA from hair. Curr. Biol. *14*, R463-4.

51. Wales, N., Andersen, K., Cappellini, E., Ávila-Arcos, M.C., and Gilbert, M.T.P. (2014). Optimization of DNA recovery and amplification from non-carbonized archaeobotanical remains. PLoS One *9*, e86827.

52. Cappellini, E., Gilbert, M.T.P., Geuna, F., Fiorentino, G., Hall, A., Thomas-Oates, J., Ashton, P.D., Ashford, D.A., Arthur, P., Campos, P.F., *et al.* (2010). A multidisciplinary study of archaeological grape seeds. Naturwissenschaften *97*, 205–217.

53. Dabney, J., Knapp, M., Glocke, I., Gansauge, M.T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.L., *et al.* (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc. Natl. Acad. Sci. U. S. A. *110*, 15758–15763.

54. Wales, N., Carøe, C., Sandoval-Velasco, M., Gamba, C., Barnett, R., Samaniego, J.A., Madrigal, J.R., Orlando, L., and Gilbert, M.T.P. (2015). New insights on single-stranded versus double-stranded DNA library preparation for ancient DNA. Biotechniques *59*, 368–371.

55. Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. *5*, pdb.prot5448-pdb.prot5448.

56. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10.

57. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

58. Ponsting, H., and Ning, Z. (2010). SMALT - A New Mapper for DNA Sequencing Reads. In F1000Posters, p. 1.

59. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

60. Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods *12*, 902–903.

61. Wood, D.E., and Salzberg, S.L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol. *15*, R46.

62. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

63. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., *et al.* (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. *19*, 455–477.

64. Kahlke, T., and Ralph, P.J. (2019). BASTA – Taxonomic classification of sequences and sequence bins using last common ancestor estimations. Methods Ecol. Evol. *10*, 100–103.

65. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075.

66. Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: Assessing genome assembly and annotation completeness. In Methods in Molecular Biology (Humana, New York, NY), pp. 227–245.

67. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

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

69. Bertels, F., Silander, O.K., Pachkov, M., Rainey, P.B., and Van Nimwegen, E. (2014). Automated reconstruction of whole-genome phylogenies from short-sequence reads. Mol. Biol. Evol. *31*, 1077–1088.

70. Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

71. Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. *26*, 1641–1650.

72. Clement, M., Posada, D., and Crandall, K.A. (2000). TCS: a computer program to estimate gene genealogies. Mol. Ecol. *9*, 1657–1659.

73. Pritchard, L., Glover, R.H., Humphris, S., Elphinstone, J.G., and Toth, I.K. (2016). Genomics and taxonomy in diagnostics for food security: Soft-rotting enterobacterial plant pathogens. Anal. Methods *8*, 12–24.

74. Emms, D.M., and Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238.

75. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

76. Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. *302*, 205–217.

77. Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

78. Csurös, M. (2010). Count: Evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics *26*, 1910–1912.

79. Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. *4*.

80. Viruel, J., Segarra-Moragues, J.G., Raz, L., Forest, F., Wilkin, P., Sanmartín, I., and Catalán, P. (2016). Late Cretaceous-Early Eocene origin of yams (Dioscorea, Dioscoreaceae) in the

Laurasian Palaearctic and their subsequent Oligocene-Miocene diversification. J. Biogeogr. *43*, 750–762.

81. 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 *2013*.

82. 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.

83. 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.

84. 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 Estimate Ploidy From Herbarium Specimens. Front. Plant Sci. *10*, 937.

85. Mergaert, J., Verdonck, L., and Kersters, K. (1993). Transfer of Erwinia ananas (synonym, Erwinia uredovora) and Erwinia stewartii to the genus Pantoea emend. as Pantoea ananas (Serrano 1928) comb. nov. and Pantoea stewartii (Smith 1898) comb. nov., respectively, and description of Pantoea stewartii subsp. . Int. J. Syst. Bacteriol. *43*, 162–173.

86. Kini, K., Agnimonhan, R., Afolabi, O., Milan, B., Soglonou, B., Koebnik, R., and Gbogbo, V. (2017). First report of a new bacterial leaf blight of rice caused by Pantoea ananatis and Pantoea stewartii in Benin. Plant Dis. *101*, 242.

87. Kini, K., Agnimonhan, R., Afolabi, O., Soglonou, B., Silué, D., and Koebnik, R. (2017). First report of a new bacterial leaf blight of rice caused by Pantoea ananatis and Pantoea stewartii in Togo. Plant Dis. *101*, 241.

88. Walterson, A.M., and Stavrinides, J. (2015). Pantoea: Insights into a highly versatile and diverse genus within the Enterobacteriaceae. FEMS Microbiol. Rev. *39*, 968–984.

89. Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., and Walker, A.W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. *12*, 87.

90. Der Sarkissian, C., Pichereau, V., Dupont, C., Ilsøe, P.C., Perrigault, M., Butler, P., Chauvaud, L., Eiríksson, J., Scourse, J., Paillard, C., *et al.* (2017). Ancient DNA analysis identifies marine mollusc shells as new metagenomic archives of the past. Mol. Ecol. Resour. *17*, 835–853.

91. Orlando, L., Ginolhac, A., Raghavan, M., Vilstrup, J., Rasmussen, M., Magnussen, K., Steinmann, K.E., Kapranov, P., Thompson, J.F., Zazula, G., *et al.* (2011). True single-molecule DNA sequencing of a pleistocene horse bone. Genome Res. *21*, 1705–1719.

92. Krause, J., Unger, T., Noçon, A., Malaspinas, A.S., Kolokotronis, S.O., Stiller, M., Soibelzon, L., Spriggs, H., Dear, P.H., Briggs, A.W., *et al.* (2008). Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. BMC Evol. Biol. *8*, 220.

93. Briggs, A.W., and Heyn, P. (2012). Preparation of next-generation sequencing libraries from damaged DNA. Methods Mol. Biol. *840*, 143–154.

94. Wilkin, P. (2001). Dioscoreaceae of South-Central Africa. Kew Bull. *56*, 361.

95. Herpell, J.B., Schindler, F., Bejtović, M., Fragner, L., Diallo, B., Bellaire, A., Kublik, S., Foesel,

B.U., Gschwendtner, S., Kerou, M., *et al.* (2020). The Potato Yam Phyllosphere Ectosymbiont Paraburkholderia sp. Msb3 Is a Potent Growth Promotor in Tomato. Front. Microbiol. *11*, 581.

96. Bernal, P., Llamas, M.A., and Filloux, A. (2018). Type VI secretion systems in plant-associated bacteria. Environ. Microbiol. *20*, 1–15.

97. Costa, T.R.D., Felisberto-Rodrigues, C., Meir, A., Prevost, M.S., Redzej, A., Trokter, M., and Waksman, G. (2015). Secretion systems in Gram-negative bacteria: structural and mechanistic insights. Nat. Rev. Microbiol. *13*, 343–359.

98. Mehrabi, R., Bahkali, A.H., Abd-Elsalam, K.A., Moslem, M., Ben M'Barek, S., Gohari, A.M., Jashni, M.K., Stergiopoulos, I., Kema, G.H.J., and De Wit, P.J.G.M. (2011). Horizontal gene and chromosome transfer in plant pathogenic fungi affecting host range. FEMS Microbiol. Rev. *35*, 542–554.

99. Reinhold-Hurek, B., and Hurek, T. (2011). Living inside plants: Bacterial endophytes. Curr. Opin. Plant Biol. *14*, 435–443.

100. Kuo, C.-H., and Ochman, H. (2009). Deletional Bias across the Three Domains of Life. Genome Biol. Evol. *1*, 145–152.

101. Kuo, C.H., and Ochman, H. (2010). The extinction dynamics of bacterial pseudogenes. PLoS Genet. *6*, e1001050.

102. Lemaire, B., Vandamme, P., Merckx, V., Smets, E., and Dessein, S. (2011). Bacterial leaf symbiosis in angiosperms: Host specificity without Co-Speciation. PLoS One *6*, e24430.

103. Carlier, A., and Eberl, L. (2012). The eroded genome of a Psychotria leaf symbiont: Hypotheses about lifestyle and interactions with its plant host. Environ. Microbiol. *14*, 2757–2769.

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

105. Alonso, D.P., Mancini, M.V., Damiani, C., Cappelli, A., Ricci, I., Vinicius Niz Alvarez, M., Bandi, C., Ribolla, P., and Favia, G. (2018). Genome reduction in the mosquito symbiont Asaia. Genome Biol. Evol. *10*, 2716–2733.

106. Manzano-Marín, A., and Latorre, A. (2016). Snapshots of a shrinking partner: Genome reduction in Serratia symbiotica. Sci. Rep. *6*, 32590.

107. Sung, W., Ackerman, M.S., Miller, S.F., Doak, T.G., and Lynch, M. (2012). Drift-barrier hypothesis and mutation-rate evolution. Proc. Natl. Acad. Sci. *109*, 18488–18492.

108. Bobay, L.-M., and Ochman, H. (2018). Factors driving effective population size and pan-genome evolution in bacteria. BMC Evol. Biol. *18*, 153.

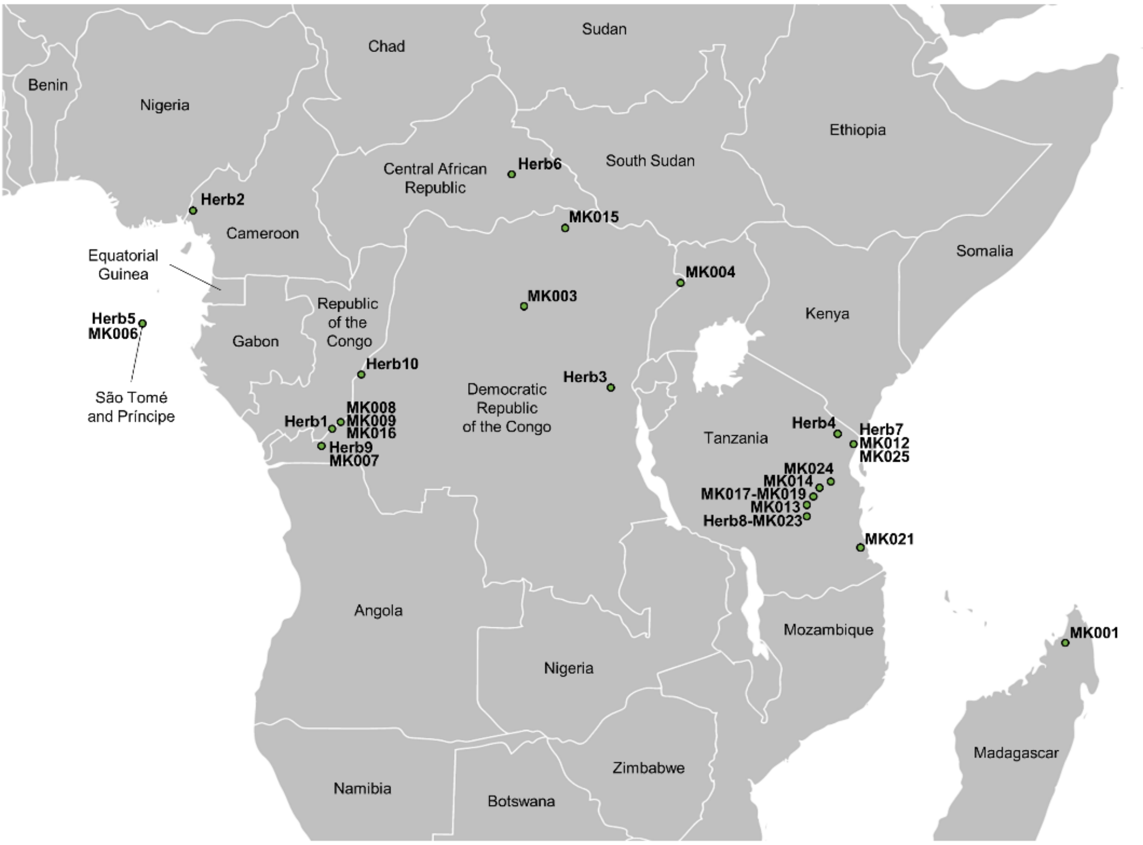

FIGURE 1: Sample location of herbarium specimens. Approximate sample location based on information available from the herbarium sheets. No accurate location could be determined for specimens MK002, MK005, MK010, MK011, MK018, MK020, MK022, and MK026.

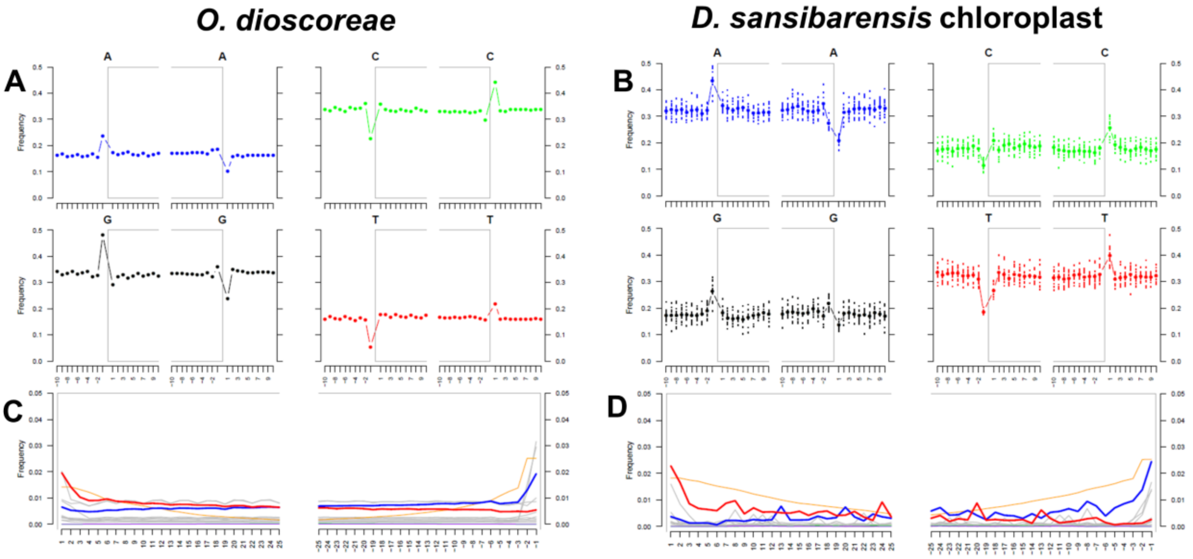

FIGURE 2: DNA Damage patterns. Output from MapDamage 2.0 of sample MK023, showing different DNA damage patterns. (A-B) Frequency of bases around read ends (grey brackets) mapped to the *Orrella dioscoreae* (A) and *Dioscorea sansibarensis* chloroplast (B) reference genomes. Numbers on the x-axis represent 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; Orange: Soft-masked bases; Grey: Other mismatches. The chloroplast lines are more irregular due to the lower sequencing coverage.

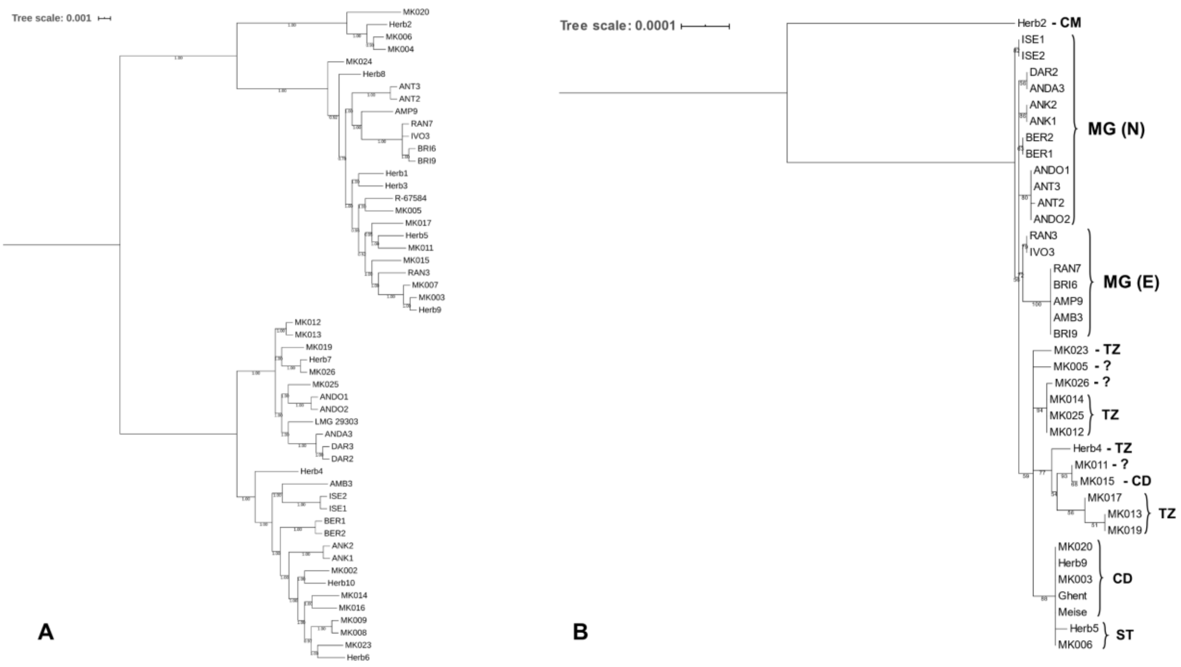

FIGURE 3: SNP-based phylogenies of *D. sansibarensis* chloroplast (A) and *O. dioscoreae* (B). Bootstrap values (for chloroplast), and Shimodaira-Hasegawa local support values (for *O. dioscoreae*) are displayed on branches. Branches with support < 50% were collapsed. Abbreviations next to the chloroplast tree correspond to where the specimens were collected originally. CM: Cameroon; MG: Madagascar (North (N) and East (E)); TZ: Tanzania; CD: Democratic Republic of Congo; ST: São Tomé & Príncipe. Plants from the botanical gardens of Meise and Ghent were originally collected in DR Congo, and are annotated as such on the tree.

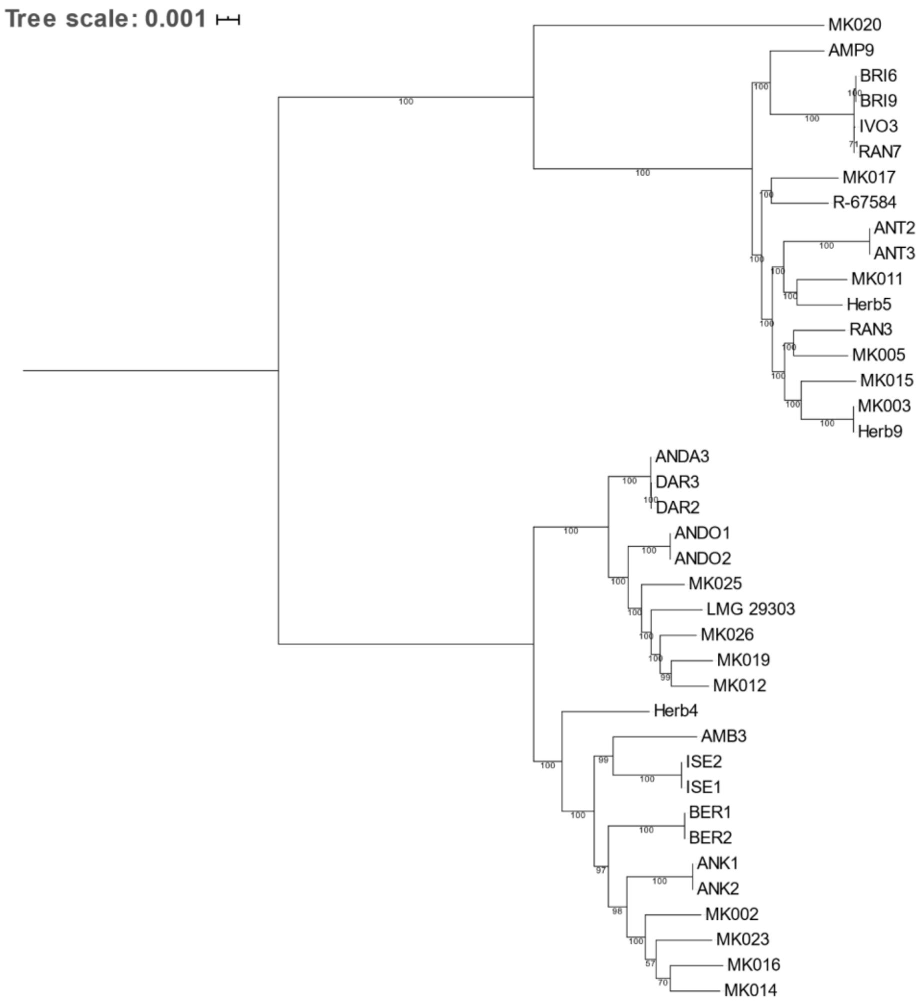

FIGURE 4: Core genome phylogeny of *Orrella dioscoreae*. Core genome phylogeny based on 3247 single-copy core genes. Numbers on the branches represent bootstrap values based on 100 replications.

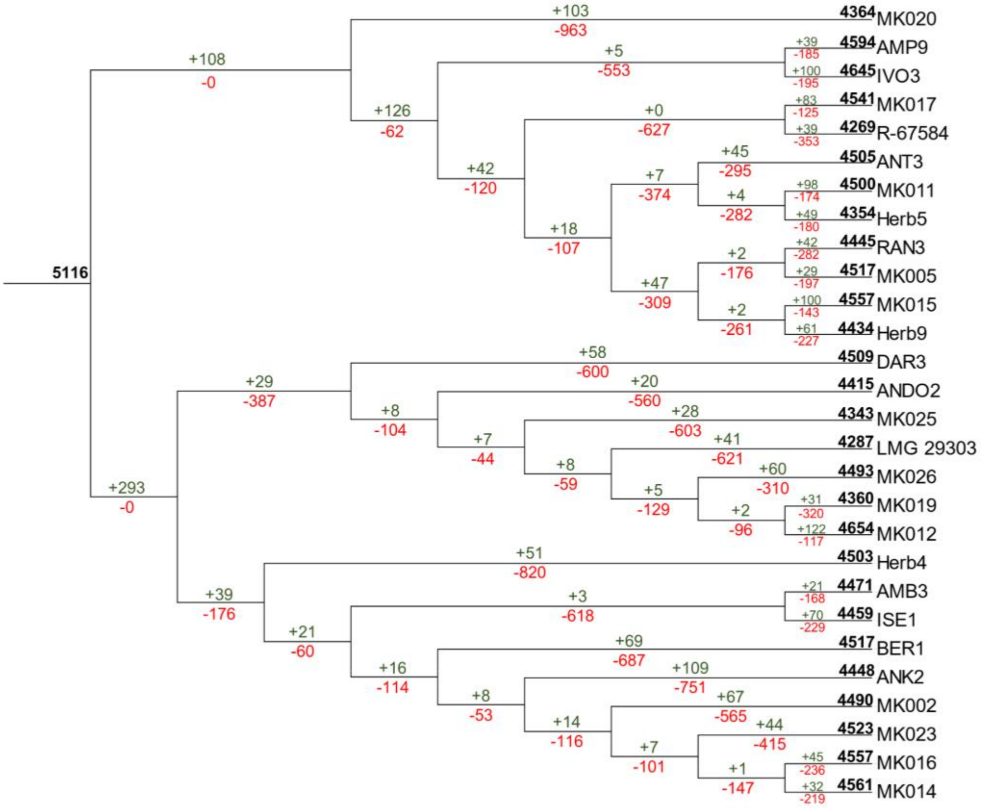

FIGURE 5: Gene gains and losses in the *Orrella dioscoreae* genome. Reconstruction of gene gain and loss, based on Dollo's parsimony principle. Numbers on branches represent gained (+) and lost (-) genes. Bold numbers represent the estimated size of the ancestral gene pool (left), or samples represent the current number of genes in a certain genome (right).