Table S1: Presence of bacterial markers in the trimmed sequencing reads, as determined by Metaphlan3. Numbers represent percentage of found markers that correspond to a certain species. Samples MK001 and MK024 are not shown as no markers were detected.

Sample name	*Propionibacterium namnetense*	*Staphylococcus epidermidis*	*Staphylococcus capitis*	*Bacillus aryabhattai*	*Bacillus megaterium*	*Cutibacterium acnes*	*Orrella dioscoreae*
Herb1	0.00	0.00	0.00	0.00	0.00	0.00	100
Herb2	0.00	0.00	0.00	0.00	0.00	0.00	100
Herb3	0.00	0.00	0.00	0.00	0.00	0.00	100
Herb4	0.00	0.00	0.00	0.00	0.00	0.00	100
Herb5	0.00	0.00	0.00	0.00	0.00	0.00	100
Herb6	0.00	0.00	0.00	0.00	0.00	0.00	100
Herb7	0.00	0.00	0.00	0.00	0.00	0.00	100
Herb8	0.00	0.00	0.00	0.00	0.00	0.00	100
Herb9	0.00	0.00	0.00	0.00	0.00	0.00	100
Herb10	0.00	0.00	0.00	0.00	0.00	0.00	100
MK002	0.00	0.00	0.00	0.00	0.00	0.00	100
MK003	0.00	0.00	0.00	0.00	0.00	0.00	100
MK004	0.00	0.00	0.00	0.00	0.00	0.00	100
MK005	0.00	0.00	0.00	0.00	0.00	0.00	100
MK006	0.00	0.00	0.00	0.00	0.00	0.00	100
MK007	0.00	0.00	0.00	0.00	0.00	0.00	100
MK008	0.00	0.00	0.00	0.00	0.00	0.00	100
MK009	0.00	0.00	0.00	0.00	0.00	0.00	100
MK010	0.99	1.38	5.04	4.70	7.32	41.47	27
MK011	0.00	0.00	0.00	0.00	0.00	0.00	100
MK012	0.00	0.00	0.00	0.00	0.00	0.00	100
MK013	0.00	0.00	0.00	0.00	0.00	0.00	100
MK014	0.00	0.00	0.00	0.00	0.00	0.00	100
MK015	0.00	0.00	0.00	0.00	0.00	0.00	100
MK016	0.00	0.00	0.00	0.00	0.00	0.00	100
MK017	0.00	0.00	0.00	0.00	0.00	0.00	99
MK018	2.19	0.00	0.00	0.23	0.00	16.61	80
MK019	0.00	0.00	0.00	0.00	0.00	0.00	100
MK020	0.00	0.00	0.00	0.00	0.00	0.00	100
MK022	0.00	0.00	0.00	0.00	0.00	0.00	100
MK023	0.00	0.00	0.00	0.00	0.00	0.00	100
MK025	0.00	0.00	0.00	0.00	0.00	0.00	100
MK026	0.00	0.00	0.00	0.00	0.00	0.00	100

FIGURE S1: Manual binning of contigs likely derived from *O. dioscoreae*. Contigs are plotted based on their % G+C and coverage. As the symbiont has higher % G+C and higher coverage, they can be visually separated from contaminating contigs. Coloring of contigs is based on classification using Kraken. The red box represent the contigs that are selected for constructing the whole genome assembly (see methods).

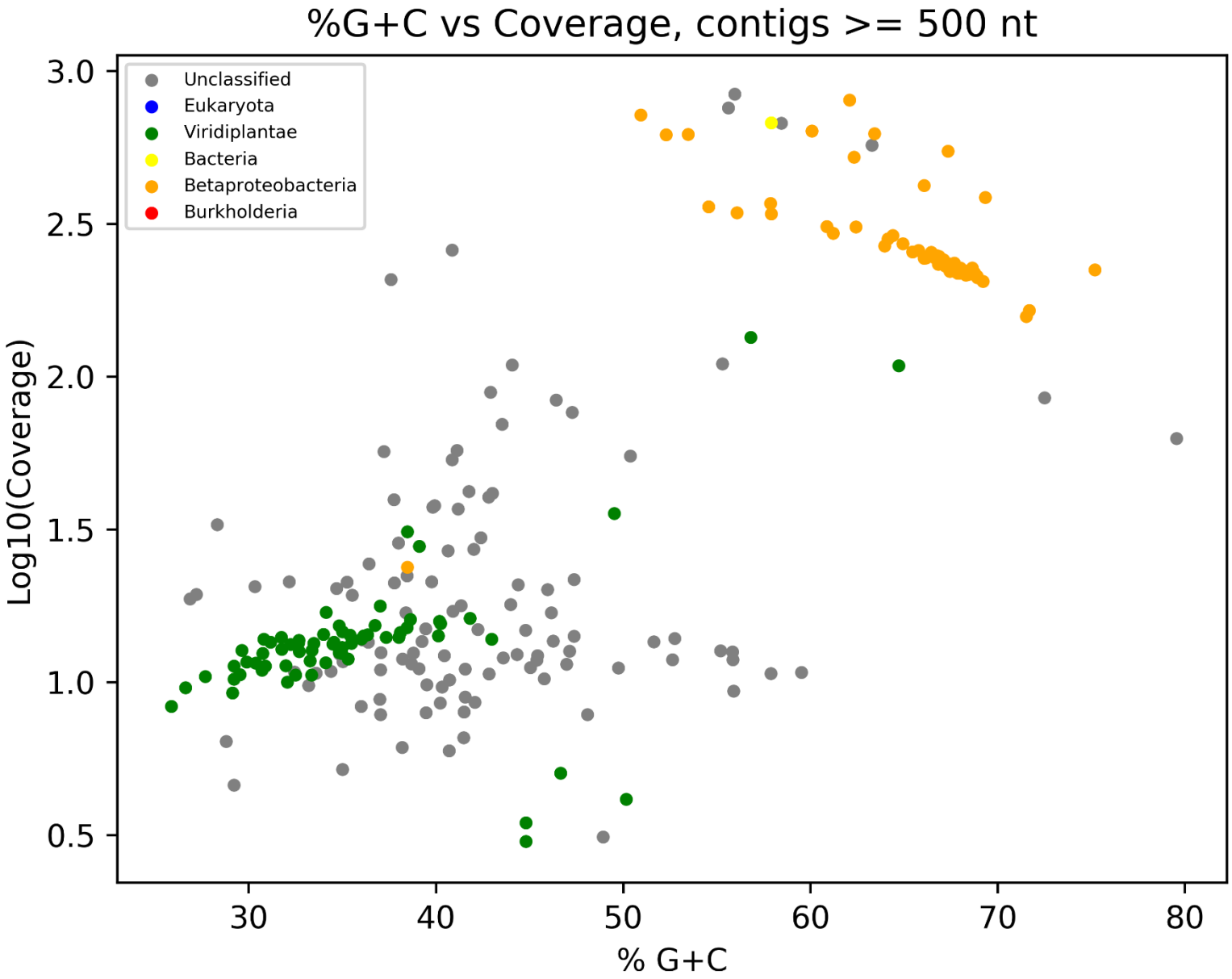
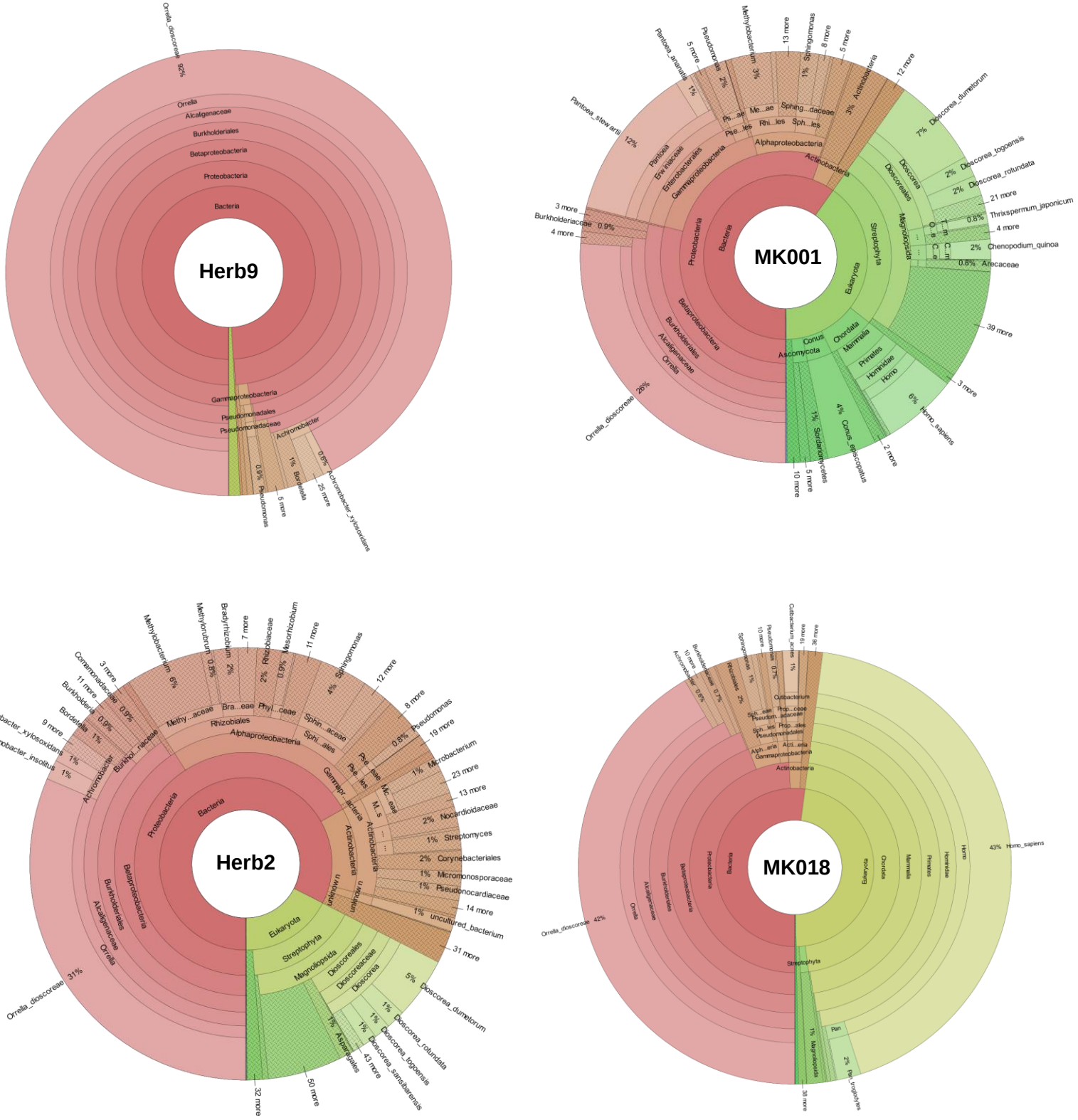


FIGURE S2: Taxonomic classification of reads in samples with low *O. dioscoreae* coverage. Classification based on best blast hit of every read in the NCBI nr nucleotide database. Herb 9 is used as reference, as it represents an average sample, with > 90% of reads classified as *O. dioscoreae.* Remark: Colours are not representative for the same species across samples.



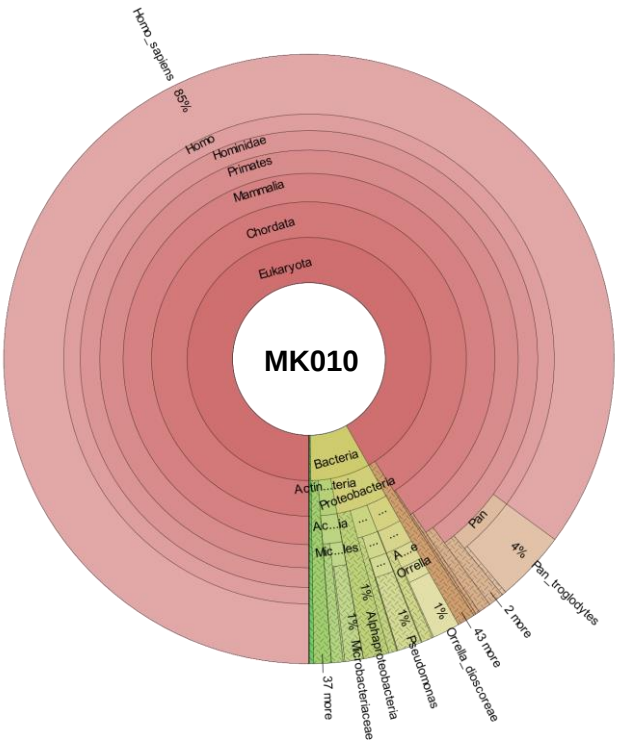

MK010
Homo_sapiens 85%
Homo
Hominidae
Primates
Mammalia
Chordata
Eukaryota
Bacteria
Actin...teria
Proteobacteria
Ac...ia
Mc...les
A...e
Orrelia
Pan
4% Pan_troglodytes
2 more
43 more
Orrella_dioscoreae
1% Pseudomonas
1% Alphaproteobacteria
1% Microbacteriaceae
37 more

FIGURE S3: Differences in C-to-T conversions on the first base, and relative increase of purines between chloroplast and *O. dioscoreae*. (A) Difference in number of C-to-T mismatches between chloroplast and *O. discoreae* DNA. *Orrella dioscoreae* DNA shows a slight significantly higher percentage of mismatches compared to the chloroplast (2.64% vs. 2.25%, paired student t-test *p*-value < 0.05). (B-D) Differences in relative increase of total purines (B), adenosine (C) and guanine (D) preceding strand breaks in chloroplast compared to *O. dioscoreae*. Difference in total purines (chloroplast 1.30x increase, *O. dioscoreae* 1.45x increase, Wilcoxon signed rank sum test *p*-value < 0.01), and adenosines (chloroplast 1.23x increase, *O. dioscoreae* 1.56x increase, Wilcoxon signed rank sum test *p*-value < 0.001), while the difference in guanine was not significant (chloroplast 1.43x increase, *O. dioscoreae* 1.39x increase, Wilcoxon signed rank sum test *p*-value > 0.5)

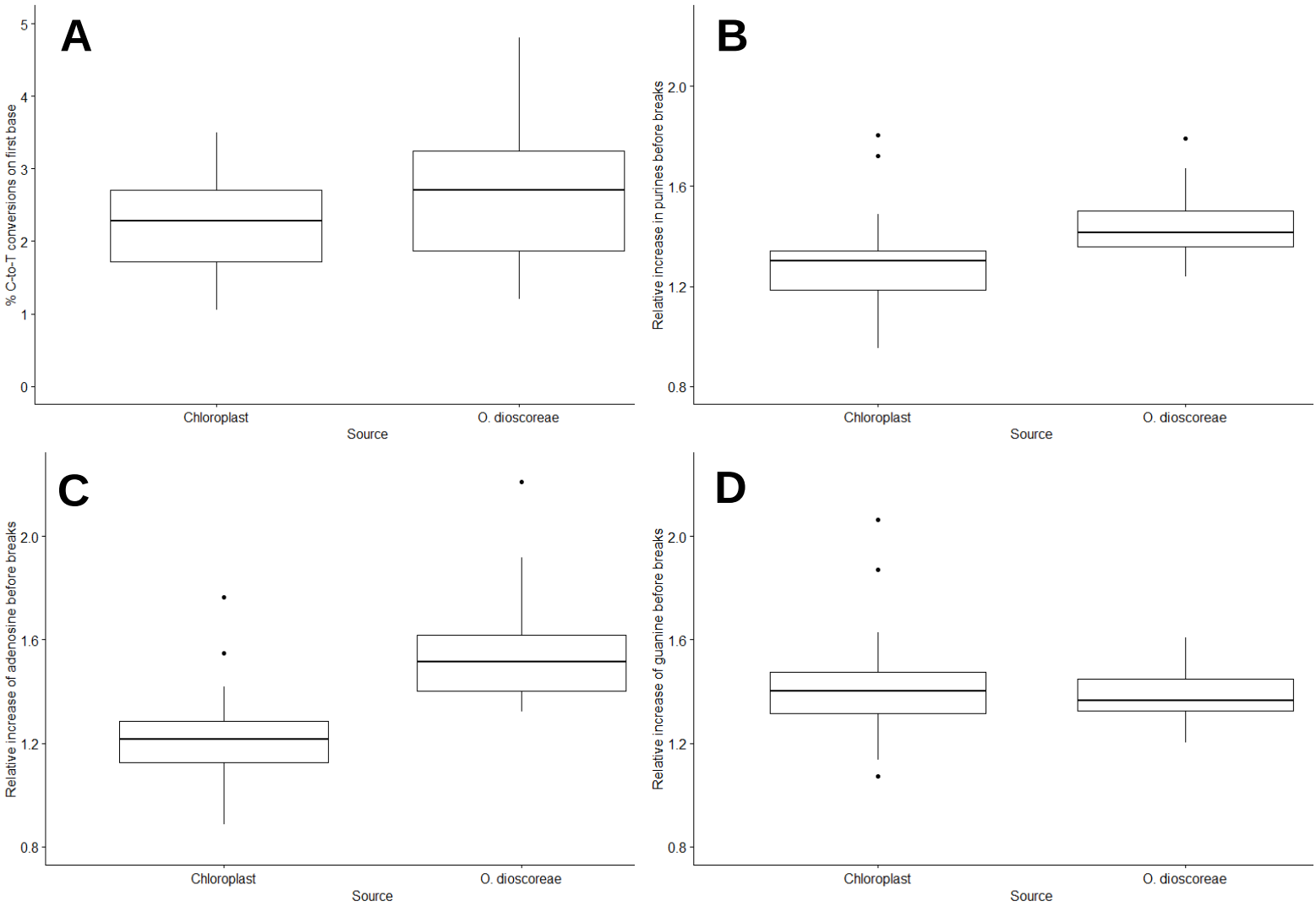


FIGURE S4: Correlation of specimen age with the amount of C-to-T conversions, and with relative purine increase at strand breaks. (A-B) Correlation between number of C-to-T mismatches and specimen age in the *Dioscorea sansibarensis* chloroplast (A) and in *Orrella dioscoreae* (B). Both correlations were statistically significant (Pearsson correlations 0.86, *p*-value < 0.001 (A) and 0.73, *p*-value < 0.01 (B). (C-D) Correlation between the relative increase of purine bases before strand breaks and specimen age. Neither of these correlations was significant (Pearsson correlations 0.40, *p*-value 0.2 (C) and 0.17, *p*-value 0.6 (D)

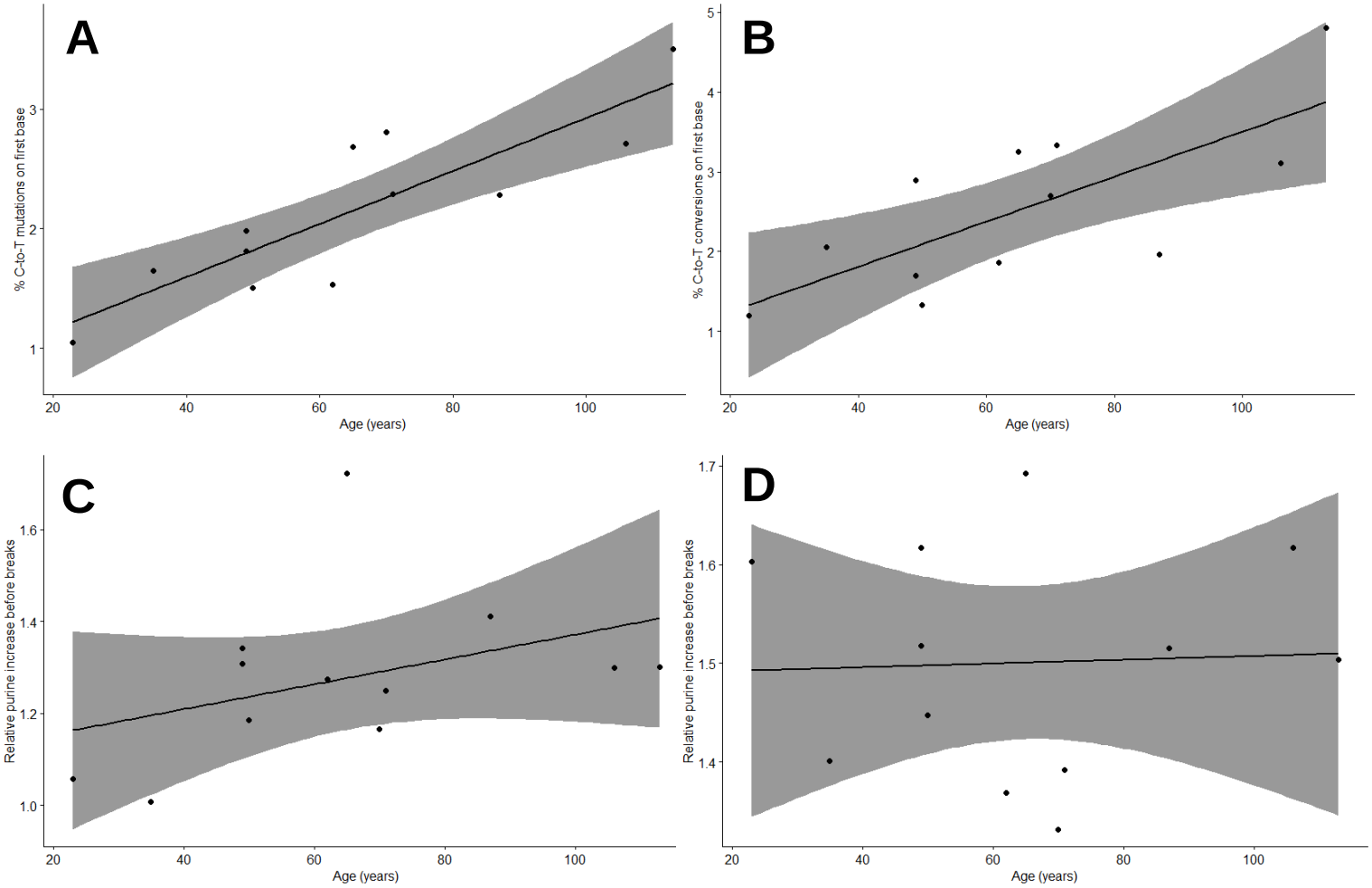


FIGURE S5: Effect on phylogenetic distance on C-to-T conversions. Mutation patterns on the first 25 bases of reads in samples MK026 (A-B) and MK011 (C-D). MK026 is more closely related to the LMG 29303^{T} reference strain, as can be seen from the lower general mutation frequency (A vs C). In this sample, C-to-T mutation frequency is similar between *O. dioscoreae* and *D. sansibarensis* chloroplast (A vs B). In MK011, a sample phylogenetically more distant from the reference, C-to-T mutation frequency is higher in *O. dioscoreae* compared to the *D. sansibarensis* chloroplast (C vs D).

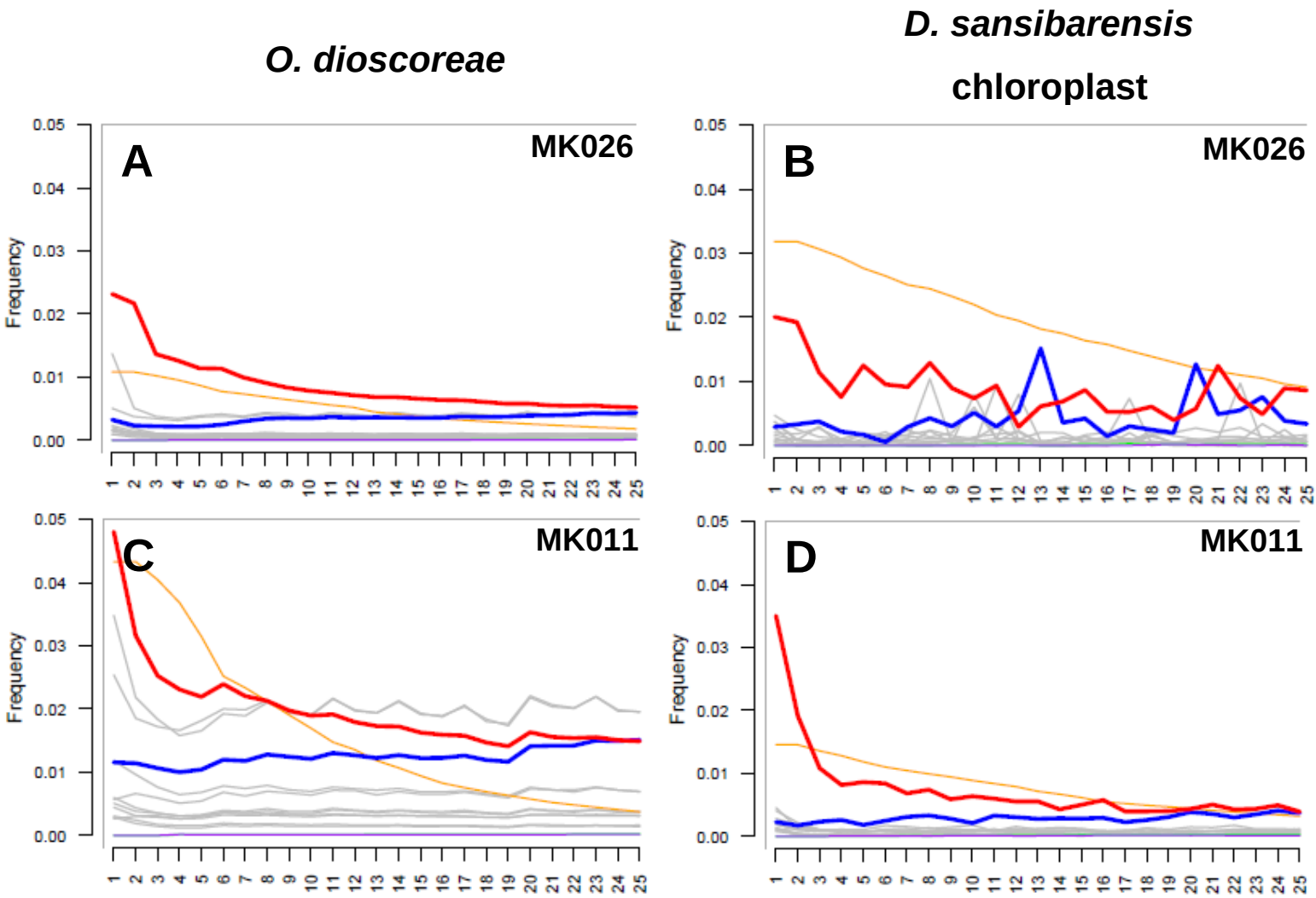


FIGURE S6: Effect of cytosine methylation on DNA breakage. Average number of 5' read ends mapping to the *Orrella dioscoreae* reference genome after a given base in the vicinity of methylated cytosine residues (m4C). Numbers on the x-axis represent the distance from the m4C residue, y-axis represents the average amount of 5' ends mapping after a certain base.

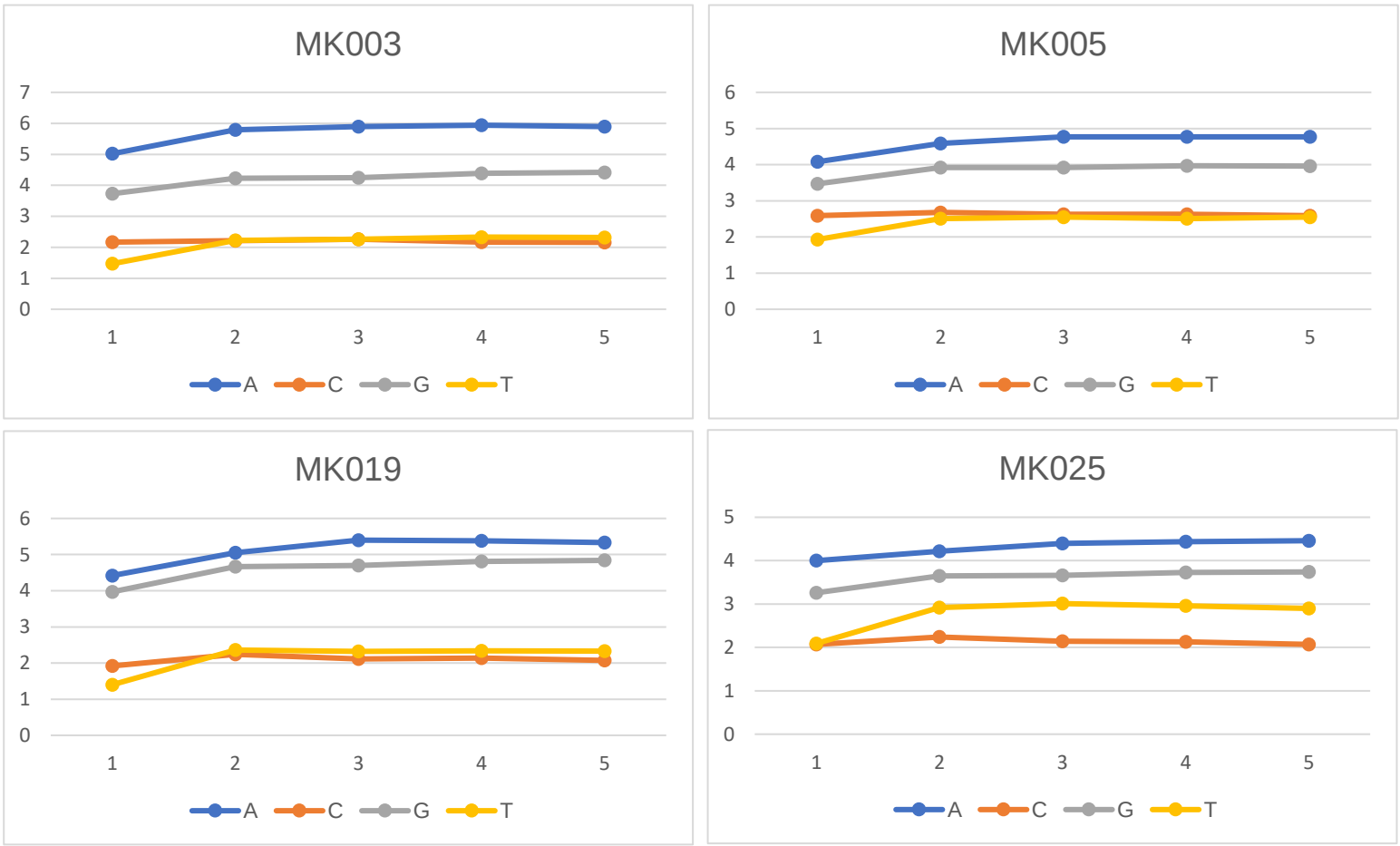


FIGURE S7: Prevalence of orthogroups with certain number of genomes. Histogram of number of orthogroups vs. the number of genomes they include. In total, 28 genomes were included, so the bar at 28 represents the core genome.

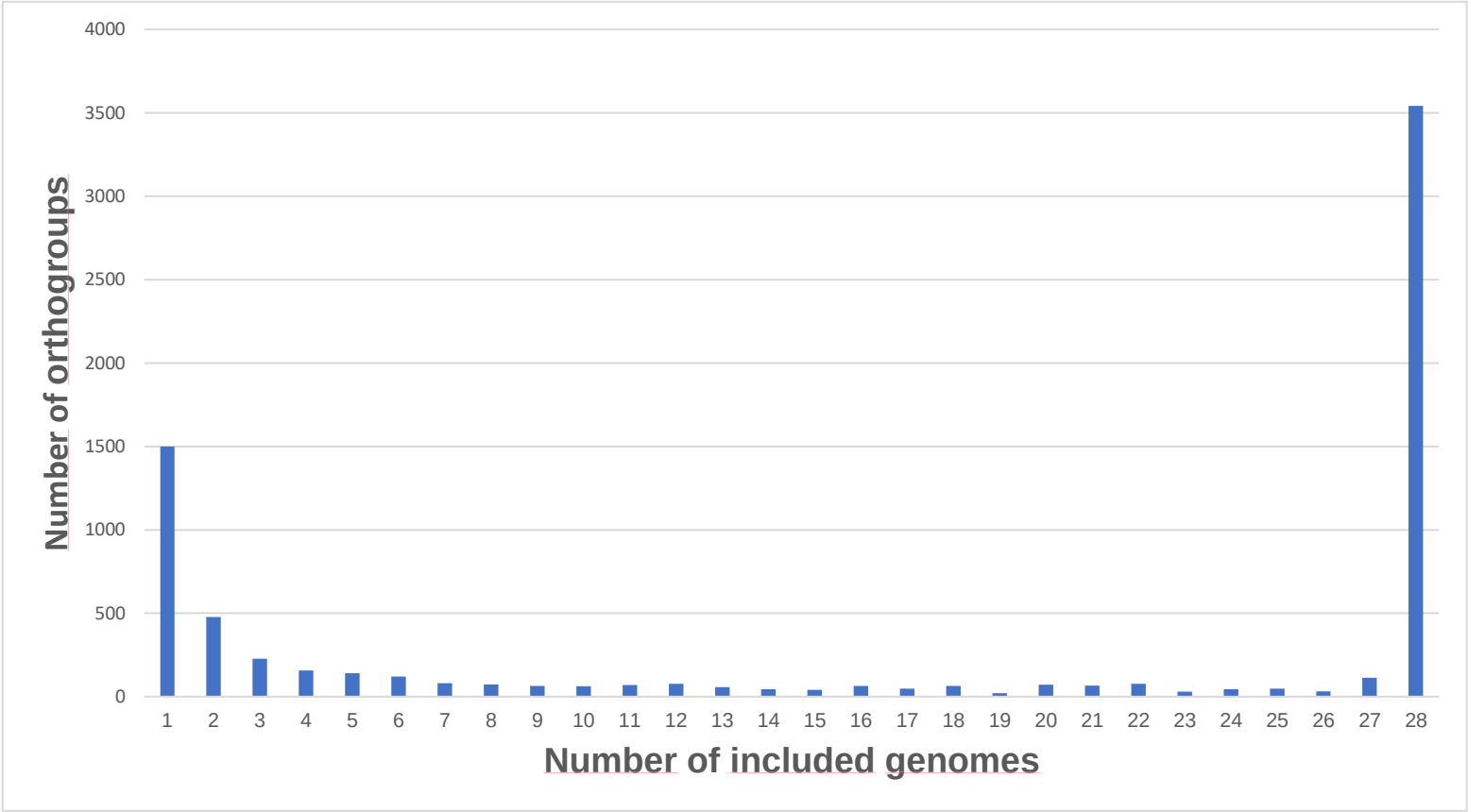


FIGURE S8: Presence of Type III Secretion system in different genomes. Stars indicate genomes where the Type III secretion system cluster is present. Only non-redundant genomes are shown (see methods).

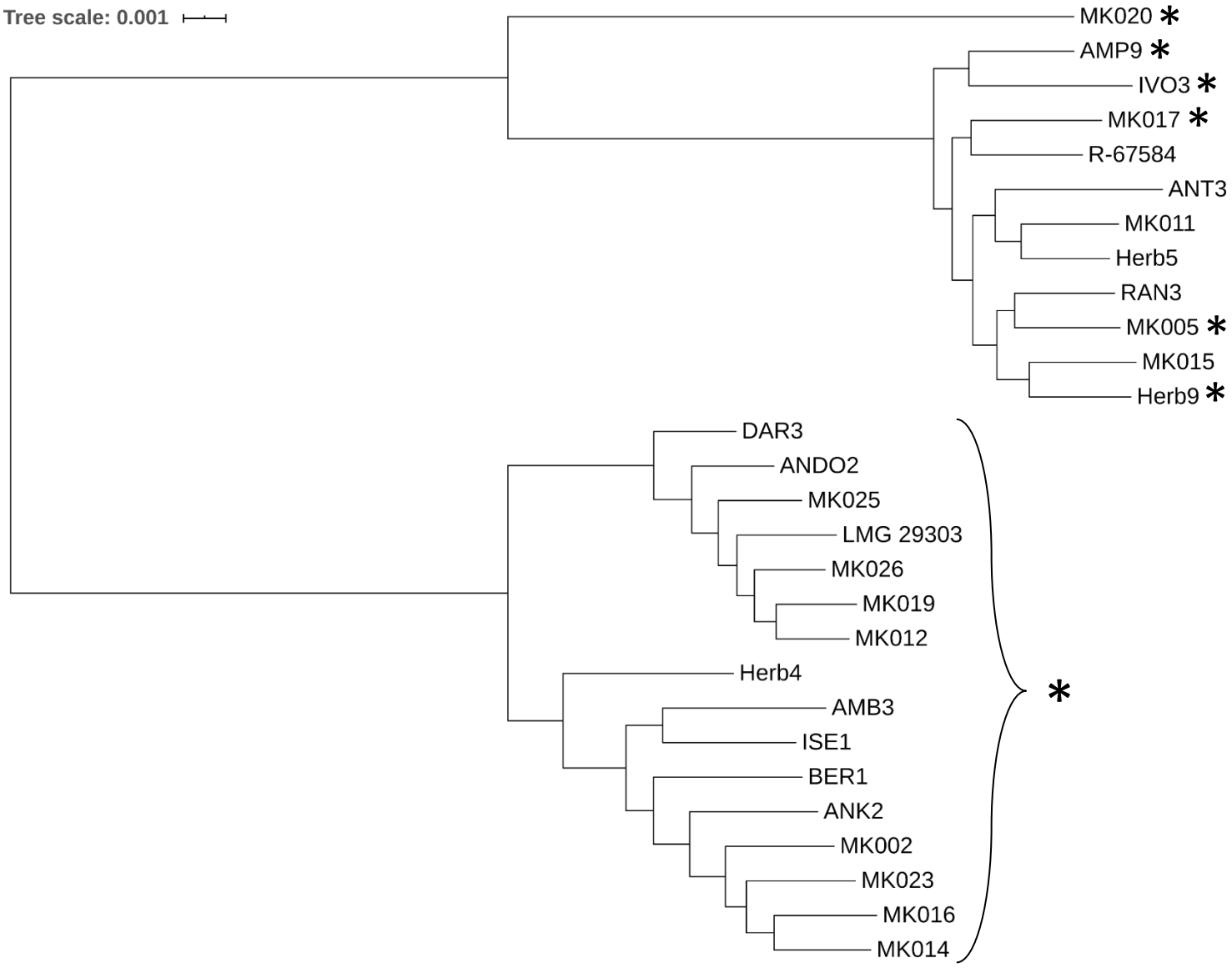


FIGURE S9: Haplotype network and map of *Dioscorea sansibarensis* plastid haplotypes. Letters correspond to haplotypes detected by TCS (see methods). Haplotypes groups are separated per country (fltr: São Tomé & Príncipe, DR Congo, Tanzania, Madagascar), and per country closely connected haplotypes are grouped together (see colours). Lines represent haplotypes groups which are directly connected in the network.