Supplementary information for ***Mosaic genome of endobacteria in arbuscular mycorrhizal fungi: trans-kingdom gene transfer in an ancient mycoplasma-fungus association***

Gloria Torres-Cortés, Stefano Ghignone, Paola Bonfante and Arthur Schüßler

Supporting Online Material

Table of Contents

Text S1

1. MATERIAL AND METHODS

1.1 Endobacteria DNA extraction for sequencing

For *Dh*MRE genomic preparation, *Dentiscutata heterogama* spores were crushed in 1 ml extraction buffer (250 mM sucrose, 10 mM MES pH 6.5, 25 mM KCl, 20 mM $MgCl_2$, 1 mM dithiothreitol) at 4°C by using a glass homogenizer. The major spore debris was pelleted by centrifugation at 500 g for 2 min, the supernatant was then again centrifuged at 1,000 g for 2 min to remove nuclei and other high density debris. The newly formed supernatant was filtered through an 8 µm and then a 3 µm polycarbonate filter (Whatman). The resulting bacteria suspension was centrifuged at 25,000 g for 15 min to pellet the bacteria; the pellet was re-suspended in re-suspension buffer (10 mM Tris-HCl pH 8, 250 mM sucrose) and treated for 60 min with DNase at 4ºC to remove free DNA. After inactivating DNase activity by heat treatment, DNA was extracted with the MasterPure Gram-positive DNA purification kit (Epicentre) according to the manufacturer's recommendations.

1.2 Semiquantitative analysis of *Dh*MRE phylotypes I and II abundance

For the semiquantitative analysis of *Dh*MRE phylotypes, DNA was extracted from 20 *D. heterogama* spores and from the suspension resulting after the endobacteria DNA extraction protocol. These DNA samples were used in the construction of each clone library. PCR was performed using MRE specific primers and the Phusion High-Fidelity DNA Polymerase (New England Biolabs). PCR products were TOPO cloned (Invitrogen) and transformed into Top10 chemically competent *Escherichia coli*. Colonies were then PCR-screened for phylotype-specific insert length differences and *Rsa*I digestion (phylotype I); 498 clones were analyzed.

1.3 Illumina sequencing and assembly

Three different sequencing libraries were constructed using the transposon-based Nextera™ DNA Sample Prep Kit (Illumina), with 50 ng of DNA as starting material. One additional library was produced with the transposon-based Nextera™ XT kit (Illumina). After library production, Illumina sequencing was performed using the Illumina MiSeq platform at the Genomics Service Unit of the Ludwig-Maximilian-University Munich Biocenter, generating 41x10^6 paired end 150 bp raw reads.

The paired-end reads were quality trimmed using CLC workbench v5 (CLC Bio), under the default parameters. Trimmed reads were mapped against the main bacterial contaminant genomes, to remove contaminant reads. The remaining, cleaned reads were assembled using the CLC de novo assembly algorithm, with a kmer size of 23, resulting in 3,655 contigs, which were then filtered according to two criteria: all contigs that had a G+C content > 45% and a coverage < 20 were discarded. This resulted in 119 contigs (1.17 Mb; *SI Appendix*, Table S1) used for further analyses. To identify the putative *Dh*MRE sequences from the resulting 119 contigs we performed BLASTX searches against the NCBI database.

To determine possible contamination by sequences from the fungal host, raw reads obtained after *Dh*MRE spore metagenome sequencing were mapped (>60% identity in 0.5 of sequence length) against the published *Rhizophagus irregularis* genome assembly (1), with only 0.18% of the reads mapping to it.

To further validate the CLC assembly, two additional strategies were followed. Firstly, raw data reads were mapped against the scaffolds with 90 % sequence similarity and length coverage as criteria, using CLC Genomics v. 5.2. Areas with low coverage were also included and manually inspected. Secondly, we tracked and visualized the paired-end connections between scaffolds, following the instructions of Albertsen *et al*. (2) and cytoscape for the visualization (3).

To identify the hypothesized existence of divergent genomes we carried out BLAST searches of putative *Dh*MRE contigs against the total of all contigs obtained. We could not identify any contigs with > 75% identity for > 1,500 bp length with the query, indicating that the assembled data are not composed of multiple closely related genomes. To analyze the expected presence of both 16S rRNA gene phylotypes in the raw reads, we mapped the raw data against both major 16S phylotypes known from Sanger sequencing approaches, but we mainly identify one of them (only 24 reads in total mapped the phylotype II; Fig. S4). To exclude the possibility of reads belonging to different phylotypes binding together in the assembly, a QualitySNPng analysis (4) was performed on the contig containing the 16S rRNA gene.

In addition to these analyses, an additional assembly was done with MIRA (5). Using this approach, we obtained the same sequence information as with the CLC assembly, but with a higher fragmentation.

Correlations between Z-scores of tetranucleotide composition were assessed using TETRA (6). For the circular and linear representation of the scaffolds the software DNAPlotter was used (7).

1.4 Phylogenetic analyses

To study *Dh*MRE proteins candidate for HGT by phylogenetic analyses, homologous sequences were selected after BLAST searches. For this, 40 BLASTP hits were selected for each *Dh*MRE query protein, represented by the five best BLASTP hits of the i) non-redundant protein sequences (nr) database from the NCBI ii) -nr database excluding *R. irregularis* iii) -nr database including fungi sequences only, but excluding *R. irregularis* and iv) -nr database including bacteria sequences only. Redundant hits were removed.

1.5 Identification of horizontal gene transfer

BLASTP analyses of the *Dh*MRE proteome against the non-redundant protein sequences (nr) database from the NCBI using the software Blast2GO were conducted to identify protein sequences with similarity to proteins from the AMF *R. irregularis,* which lacks endobacteria. BLASTP affiliation was based on the best-hit with a cut-off value of e^{-03}. Eukaryotic domains were identified by analyzing the results obtained in Interpro and SUPERFAMILY databases integrated in MicroScope platform (8). The presence of genomic islands in the *Dh*MRE genome draft was studied using the software IslandViewer.

2. SUPPLEMENTARY TABLES

Table S1. Metagenome assembly data.

Raw data	Reads (bp)	2x150
	No. of reads	41x10^{6}
	Primary sequence data (Gb)	6
Cleaned data	No. of reads	15.8x10^{6}
	Contigs sequence data Mb-No. of contigs	15/3655
	No. of contigs with G+C < 45%	1494
	No. contigs G+C < 45% and coverage > 20	119
	Cleaned sequence data (Mb)	1.17
***Dh*MRE sequences**	*Dh*MRE sequence data	0.702
	No. of *Dh*MRE contigs	24
	Average *Dh*MRE contig length (bp)	29,244
	*Dh*MRE average coverage	172x
	N50 contig size (bp)	147,306
	Longest scaffold (bp)	222,151

Contigs recovered after removing the main contaminants from the raw data were filtered according to GC content and coverage. All contigs presenting a G+C content higher than 45% and coverage below 20 were discarded. *Dh*MRE contigs were identified from the resulting contigs performing BLASTX searches against NCBI non-redundant protein sequences database.

Table S2. Genome features of *Dh*MRE in comparison to members of the *Tenericutes* and *Firmicutes*.

	***Dh*MRE**			***Tenericutes***					***Firmicutes***	
	Sca.A	**Sca. B**	**Sca. C**	**Mgen**	**Upar**	**Mhyo**	**Mflo**	**CaPhy**	**Linn**	**Saga**
Length (Mb)	0.649	0.0604	0.0038	0.580	0.752	0.840	0.793	0.602	3.01	2.13
G+C ratio	34.06	32.3	34.06	32	25	25.88	27.02	21.39	37.4	35.65
CDS[a]	606	58	6	482	613	663	683	482	3141	2196
Coding region[b]	80.96	83.69	57.13	92.14	93	85.3	93.3	78.75	89.3	86.8
RNA operons	1	0	0	1	1	1	2	2	6	7
tRNAs	35	0	0	36	39	30	29	32	66	80
Lifestyle (Host)[c]		O (F)		P (A)	P (A)	P (A)	FL	P (Pl)	FL	P (A)

[a] Number of protein-coding sequences in the corresponding scaffold/chromosome.
[b] Percentage of coding regions in the total scaffold/chromosome.
[c] Lifestyle of local taxa; O: obligate endosymbiont; P: pathogen; FL: free-living; F: fungi; A: animal; Pl: plants.

Data for Mgen, Upar, Mhyo, Mflo and CaPhy were obtained from the Molligen (9) database and for Linn and Saga from the Microscope database (8) under the following accession numbers: Mgen: *Mycoplasma genitalium* G37 (NC_000908); Upar: *Ureaplasma parvum* serovar 3 ATCC 700970 (NC_002162); Mhyo*: Mycoplasma hyorhinis* HUB-1 (CP002170); Mflo: *Mesoplasma florum* L1 (NC_006055); CaPhy: *Ca*. Phytoplasma mali (NC_011047); Linn: *Listeria innocua* (NC_003212); Saga: *Streptococcus agalactiae* A909 (NC_007432). Sca.A, B and C: Scaffold A, B and C from *Dh*MRE draft genome.

Table S3. Pearson correlation coefficients for Z-score of tetranucleotide frequency.

	***Dh*MRE A**	***Dh*MRE B**	**Mgen**	**Upar**	**Mhyo**	**Mflo**	**CaPhy**	**Linn**	**Saga**	**Rirr 523**	**Rirr 4192**	**Dacid**
***Dh*MRE A**	1											
***Dh*MRE B**	**0.89**	1										
Mgen	0.59	0.52	1									
Upar	0.55	0.50	0.66	1								
Mhyo	0.43	0.38	0.47	0.83	1							
Mflo	0.41	0.35	0.60	0.80	0.83	1						
CaPhy	0.56	0.50	0.65	0.86	0.80	0.80	1					
Linn	0.65	0.61	0.58	0.70	0.69	0.71	0.74	1				
Saga	0.65	0.61	0.68	0.65	0.59	0.71	0.74	0.83	1			
Rirr 523	0.33	0.25	0.34	0.48	0.55	0.53	0.56	0.48	0.43	1		
Rirr 4192	0.43	0.38	0.36	0.62	0.65	0.62	0.64	0.51	0.46	0.65	1	
Dacid	0.06	0.001	0.19	0.35	0.39	0.46	0.33	0.27	0.25	0.34	0.30	1

Abbreviations and GenBank accession numbers: *Dh*MRE, *Dentiscutata heterogama* MRE scaffold; Mgen, *Mycoplasma genitalium* G37 (NC_000908); Upar, *Ureaplasma parvum* serovar 3 ATCC 700970 (NC_002162); Mhyo, *Mycoplasma hyorhinis* HUB-1 (CP002170); Mflo, *Mesoplasma florum* L1 (NC_006055.1); CaPhy, *Candidatus* Phytoplasma mali (NC_011047); Linn, *Listeria innocua* (NC_003212); Saga, *Streptococcus agalactiae* (NC_007432); Rirr, *R. irregularis* scaffold (523 = KE392324; 4192 = KE392320); Daci, *Delftia acidovorax* (NC_010002). The tetranucletotide correlation between the *Dh*MRE scaffolds A and B is marked in bold.

Table S4. Analysis of the proportion of different phylotypes found in *D. heterogama* spores, and in the bacterial suspension resulting after the MRE DNA extraction protocol, through clone libraries.

		Phylotype 1		Phylotype 2		Total Number
		Clone number	**%**	**Clone number**	**%**	
Spores	Spore Clone library 1	74	86	12	14	86
	Spore Clone library 2	68	78	19	22	87
	Spore Clone library 3	74	90	8	10	82
	Spore Clone library 4	61	77	18	23	79
	TOTAL SPORES	**277**	**83**	**57**	**17**	**334**
After DNA extraction protocol	MRE Clone library 1	78	95	4	5	82
	MRE Clone library 2	79	96	3	4	82
	TOTAL MRE preparation	**157**	**96**	**7**	**4**	**164**

Table S5. List of housekeeping genes and species used for the multilocus analysis.

				infB	***rplN***	***rplO***	***rplP***	***rpoB***	***rpsC***	***rpsD***	***rpsE***	***rpsH***	***rpsM***	**Total Length (aa)**
	Phylum	**Class**	**Order**											
***Dh*MRE**	***Tenericutes***	***Mollicutes***	***Mycoplasmatales***	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3224
Ca. Phytoplasma asteris AYWB	*Tenericutes*	*Mollicutes*	*Acholeplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3192
Ca. Phytoplasma asteris OY-M	*Tenericutes*	*Mollicutes*	*Acholeplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3101
Ca. Phytoplasma australiense	*Tenericutes*	*Mollicutes*	*Acholeplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3138
Ca. Phytoplasma mali	*Tenericutes*	*Mollicutes*	*Acholeplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3170
Acholeplasma laidlawii	*Tenericutes*	*Mollicutes*	*Acholeplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3174
Mesoplasma florum	*Tenericutes*	*Mollicutes*	*Entomoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3222
Spiroplasma citri	*Tenericutes*	*Mollicutes*	*Entomoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3231
Spiroplasma melliferum	*Tenericutes*	*Mollicutes*	*Entomoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3232
Mycoplasma bovis	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3118
Mycoplasma synoviae	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3040
Mycoplasma agalactiae	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3118
Mycoplasma mobile	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3115
Mycoplasma haemofelis	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3286

Mycoplasma leachii	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3256
Mycoplasma hyorhinis	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3171
Mycoplasma gallisepticum	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3386
Mycoplasma genitalium	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3368
Mycoplasma hominis	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3142
Mycoplasma arthritidis	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3107
Mycoplasma hyopneumoniae	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3109
Mycoplasma conjunctivae	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3102
Mycoplasma mycoides	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3260
Mycoplasma crocodyli	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3147
Mycoplasma capricolum	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3256
Mycoplasma penetrans	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3512
Mycoplasma pulmonis	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3146
Mycoplasma pneumoniae	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3383
Ureaplasma urealyticum	*Tenericutes*	*Mollicutes*	*Mycoplasmatales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3382
Lactobacillus plantarum	*Firmicutes*	*Bacilli*	*Lactobacillales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3187
Streptococcus pneumoniae	*Firmicutes*	*Bacilli*	*Lactobacillales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3383
Lactobacillus crispatus	*Firmicutes*	*Bacilli*	*Lactobacillales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3185
Listeria monocitogenes	*Firmicutes*	*Bacilli*	*Bacillales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3216

Staphylococcus aureus	*Firmicutes*	*Bacilli*	*Bacillales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3136
Staphylococcus epidermidis	*Firmicutes*	*Bacilli*	*Bacillales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3151
Bacillus pumilus	*Firmicutes*	*Bacilli*	*Bacillales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3132
Ca. Desulforudis audaxviator	*Firmicutes*	*Clostridia*	*Clostridiales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3314
Clostridium kluyveri	*Firmicutes*	*Clostridia*	*Clostridiales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3166
Clostridium botulinum	*Firmicutes*	*Clostridia*	*Clostridiales*	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	3167

Sequences have been retrieved from the Molligen database and from NCBI. Gene abbreviations: *infB*, translation initiation factor IF-2; *rplN*, large subunit ribosomal protein L14; *rplO*, large subunit ribosomal protein L15; *rplP*, large subunit ribosomal protein L16; *rpoB*, DNA-directed RNA polymerase subunit beta; *rpsC*, small subunit ribosomal protein S3; *rpsD*, small subunit ribosomal protein S4; *rpsE*, small subunit ribosomal protein S5; *rpsH*, small subunit ribosomal protein S8; *rpsM*, small subunit ribosomal protein S13. P: sequence available.

Table S6. Proteins from *Dh*MRE showing BLAST hit against proteins from *Rhizophagus irregularis*. Proteins with an identity > 30% and a query coverage > 35% are shown in the table.

Scaffold	***Dh*MRE**				***R. irregularis***				**BLAST result**			**Possible function/Domains**[c]
	Gene number	**GC**[a] **%**	**CAI**[b]	**Prot. lenght (aa)**	**Acc. number**	**GC**[a] **%**	**CAI**[b]	**Prot. lenght (aa)**	**Query cover**	**E value**	**Ident (%)**	
A	0417	31	0.222	381	EXX74175	22	0.194	788	0.55	$4E^{-29}$	38	PUF_L domain like
A	0659	38	0.194	284	EXX58955	27	0.232	1437	0.97	$3E^{-39}$	37	PUF_L domain like
A	0608*	31	0.233	507	EXX62832	28	0.349	571	0.41	$2E^{-27}$	39	PUF_L domain like
A	0487	36	0.260	264	ESA21130	27	0.220	380	0.65	$4E^{-21}$	37	PUF-RNI-like domain
A	0002	38	0.225	391	ESA09032	35	0.247	433	0.92	$2E^{-125}$	50	PUF
A	0032	31	0.234	279	ESA09071	32	0.221	292	0.82	$3E^{-40}$	39	PUF-AIG1 domain
A	0522	30	0.236	564	ESA09071	32	0.221	292	0.45	$2E^{-31}$	34	PUF-AIG1 domain
A	0172	38	0.153	217	EXX59421	39	0.145	140	0.6	$5E\text{-}^{14}$	37	PUF
A	0345	34	0.209	353	EXX67517	36	0.240	219	0.49	$5E\text{-}^{6}$	27	PUF
A	0022	32	0.201	553	ESA03387	24	0.227	331	0.52	$4E^{-20}$	31	PUF
A	0031	30	0.217	1,031	EXX75677	28	0.221	420	0.36	$5E^{-107}$	47	Non-specific protein-tyrosine kinase
A	0091*	33	0.196	812	EXX53579	28	0.189	446	0.47	$8E^{-105}$	45	Non-specific protein-tyrosine kinase
A	0349	32	0.245	787	EXX62449	27	0.224	495	0.44	$3E^{-81}$	41	Non-specific protein-tyrosine kinase
A	0521	.31	0.227	1,119	EXX52799	28	0.196	692	0.44	$6E^{-84}$	34	Non-specific protein-tyrosine kinase
A	0574	32	0.226	766	EXX57629	28	0.223	467	0.52	$5E^{-85}$	40	Non-specific protein-tyrosine kinase
A	0602*	34	0.209	551	EXX75398	27	0.210	477	0.8	$2E^{-125}$	46	Non-specific protein-tyrosine kinase

A	0673	33	0.196	812	EXX53579	28	0.189	446	0.47	$1E^{-104}$	45	Non-specific protein-tyrosine kinase
B	0015	33	0.244	204	ESA20877	34	0.291	434	0.84	$7E^{-33}$	41	Conserved PUF-AIG1
B	0016	34	0.283	448	ESA07576	34	0.244	302	0.67	$8E^{-78}$	47	Conserved PUF-AIG1
B	0018	31	0.210	348	ESA09071	32	0.221	292	0.66	6E-53	44	PUF-AIG1
B	0019	32	0.204	479	EXX77776	30	0.204	844	0.9	$2E^{-57}$	38	PUF-AIG1 domain
B	0026	32	0.247	535	EXX62915	25	0.201	434	0.68	$5E^{-83}$	39	Non-specific protein-tyrosine kinase
B	0032	31	0.237	242	ESA18820	30	0.230	226	0.94	$7E^{-112}$	74	Conserved PUF
B	0035	32	0.184	189	ESA08495	41	0.247	103	0.53	$6E^{-10}$	36	PUF
B	0039	30	0.140	193	EXX54862	26	0.226	623	0.59	$5E^{-07}$	38	PUF-L-Like domain
B	0041	32	0.267	337	EXX54859	26	0.226	740	0.57	$6E^{-29}$	42	PUF-L-Like domain
B	0048	30	0.254	382	EXX54862	26	0.226	623	0.56	$1E^{-31}$	38	PUF-LRR domain

[*] Proteins located in putative genomic islands

[a] The average GC content of the *Dh*MRE genome is 34%

[b] CAI, codon adaptation index. The average CAI for a gene in *Dh*MRE is 0.210. GC content and CAI were calculated using the Mobyle platform at http://mobyle.pasteur.fr/cgi-bin/portal.py#welcome

[c] PUF, protein of unknown function

[d] GI, genes associated to genomic islands.

Table S7. Set of 100 essential COGs conserved in 99% of bacteria (10).

COG	Code	COG's description	*Dh*MRE
COG0563	F	Adenylate kinase and related kinases	P
COG0528	F	Uridylate kinase	A
COG0587	L	DNA polymerase III, alpha subunit	P
COG2812	L	DNA polymerase III, gamma/tau subunits	A
COG0592	L	DNA polymerase sliding clamp subunit (PCNA homologous)	P
COG0358	L	DNA primase (bacterial type)	P
COG0084	L	Mg-dependent DNase	P
COG0305	L	Replicative DNA helicase	A
COG0629	L	Single-stranded DNA-binding protein	P
COG0188	L	Type IIA topoisomerase (DNA gyrase/topo II, topoisomerase IV), A subunit	P
COG0187	L	Type IIA topoisomerase (DNA gyrase/topo II, topoisomerase IV), B subunit	P
COG0202	K	DNA-directed RNA polymerase, alpha subunit/40 kD subunit	P
COG0086	K	DNA-directed RNA polymerase, beta' subunit/160 kD subunit	P
COG0568	K	DNA-directed RNA polymerase, sigma subunit (sigma70/sigma32)	P
COG0571	K	dsRNA-specific ribonuclease	P
COG0250	K	Transcription antiterminator	P
COG0195	K	Transcription elongation factor	P
COG0081	J	Ribosomal protein L1	P
COG0244	J	Ribosomal protein L10	P
COG0080	J	Ribosomal protein L11	P
COG0102	J	Ribosomal protein L13	P
COG0093	J	Ribosomal protein L14	P
COG0200	J	Ribosomal protein L15	P
COG0197	J	Ribosomal protein L16/L10E	P
COG0203	J	Ribosomal protein L17	P
COG0256	J	Ribosomal protein L18	P
COG0335	J	Ribosomal protein L19	P
COG0090	J	Ribosomal protein L2	P
COG0292	J	Ribosomal protein L20	P
COG0091	J	Ribosomal protein L22	P
COG0089	J	Ribosomal protein L23	P
COG0198	J	Ribosomal protein L24	P
COG0211	J	Ribosomal protein L27	P
COG0087	J	Ribosomal protein L3	P
COG0254	J	Ribosomal protein L31	P
COG0088	J	Ribosomal protein L4	P
COG0094	J	Ribosomal protein L5	P
COG0097	J	Ribosomal protein L6P/L9E	P
COG0222	J	Ribosomal protein L7/L12	P
COG0051	J	Ribosomal protein S10	P
COG0100	J	Ribosomal protein S11	P
COG0048	J	Ribosomal protein S12	P
COG0099	J	Ribosomal protein S13	P
COG0184	J	Ribosomal protein S15P/S13E	P
COG0228	J	Ribosomal protein S16	P
COG0186	J	Ribosomal protein S17	P
COG0238	J	Ribosomal protein S18	P
COG0052	J	Ribosomal protein S2	P
COG0268	J	Ribosomal protein S20	P
COG0092	J	Ribosomal protein S3	P
COG0522	J	Ribosomal protein S4 and related proteins	P
COG0098	J	Ribosomal protein S5	P
COG0360	J	Ribosomal protein S6	P
COG0049	J	Ribosomal protein S7	P

COG0096	J	Ribosomal protein S8	P
COG0103	J	Ribosomal protein S9	P
COG0233	J	Ribosome recycling factor	P
COG0858	J	Ribosome-binding factor A	P
COG0013	J	Alanyl-tRNA synthetase	P
COG0018	J	Arginyl-tRNA synthetase	P
COG0215	J	Cysteinyl-tRNA synthetase	P
COG0008	J	Glutamyl- and glutaminyl-tRNA synthetases	P
COG0124	J	Histidyl-tRNA synthetase	P
COG0060	J	Isoleucyl-tRNA synthetase	P
COG0495	J	Leucyl-tRNA synthetase	P
COG0143	J	Methionyl-tRNA synthetase	P
COG0016	J	Phenylalanyl-tRNA synthetase alpha subunit	P
COG0072	J	Phenylalanyl-tRNA synthetase beta subunit	P
COG0193	J	Peptidyl-tRNA hydrolase	P
COG0442	J	Prolyl-tRNA synthetase	P
COG0172	J	Seryl-tRNA synthetase	P
COG0441	J	Threonyl-tRNA synthetase	P
COG0180	J	Tryptophanyl-tRNA synthetase	P
COG0162	J	Tyrosyl-tRNA synthetase	P
COG0024	J	Methionine aminopeptidase	P
COG0336	J	tRNA-(guanine-N1)-methyltransferase	P
COG0030	J	Dimethyladenosine transferase (rRNA methylation)	A
COG0012	J	Predicted GTPase, probable translation factor	P
COG0216	J	Protein chain release factor A	P
COG0050	J	GTPases – translation elongation factors	P
COG0231	J	Translation elongation factor P (EF-P)/translation initiation factor 5A (eIF-5A)	A
COG0264	J	Translation elongation factor Ts	A
COG0480	J	Translation elongation factors (GTPases)	P
COG0361	J	Translation initiation factor 1 (IF-1)	P
COG0532	J	Translation initiation factor 2 (IF-2; GTPase)	P
COG0290	J	Translation initiation factor 3 (IF-3)	P
COG0465	O	ATP-dependent Zn proteases	P
COG0484	O	DnaJ-class molecular chaperone with C-terminal Zn finger domain	P
COG0533	O	Metal-dependent proteases with possible chaperone activity	P
COG0443	O	Molecular chaperone	P
COG0576	O	Molecular chaperone GrpE (heat shock protein)	P
COG0691	O	tmRNA-binding protein	P
COG0653	U	Preprotein translocase subunit SecA (ATPase, RNA helicase)	P
COG0201	U	Preprotein translocase subunit SecY	A
COG0706	U	Preprotein translocase subunit YidC	A
COG0481	M	Membrane GTPase LepA	P
COG0275	M	Predicted S-adenosylmethionine-dependent methyltransferase involved in cell envelope biogenesis	P
COG0536	R	Predicted GTPase	A
COG1160	R	Predicted GTPases	P
COG0319	R	Predicted metal-dependent hydrolase	A

Presence (P) or absence (A) of the COG in the DhMRE genome are indicated.

3. SUPPLEMENTARY FIGURES

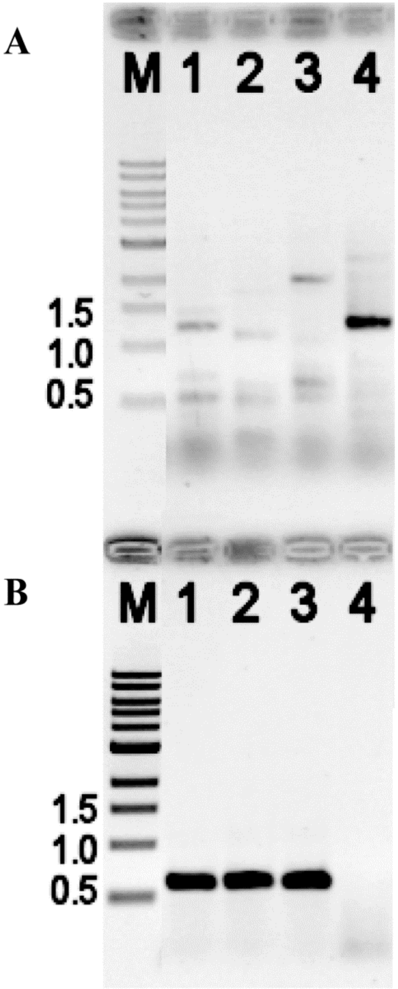

Fig. S1. Figure S1. PCR amplification of bacterial 16S rRNA genes from AMF belonging to the Gigasporaceae. DNA was amplified with MRE specific primers (A; 1.4 kb amplicon expected) and Ca. Glomeribacter gigasporarum (Burkholderia-related) specific primers (B; 0.7 kb amplicon). (M) marker; (1) Gigaspora decipiens AU102, (2, 3) two different Gigaspora margarita isolates, (4) Dentiscutata heterogama FL654. Dentiscutata heterogama spores are free from Ca. Glomeribacter gigasporarum, but contain the Mollicutes-related endobacteria which were target of this study. As the degenerate MRE primers used cause some unspecific products, MRE amplicons were cloned and sequenced to confirm their origin.

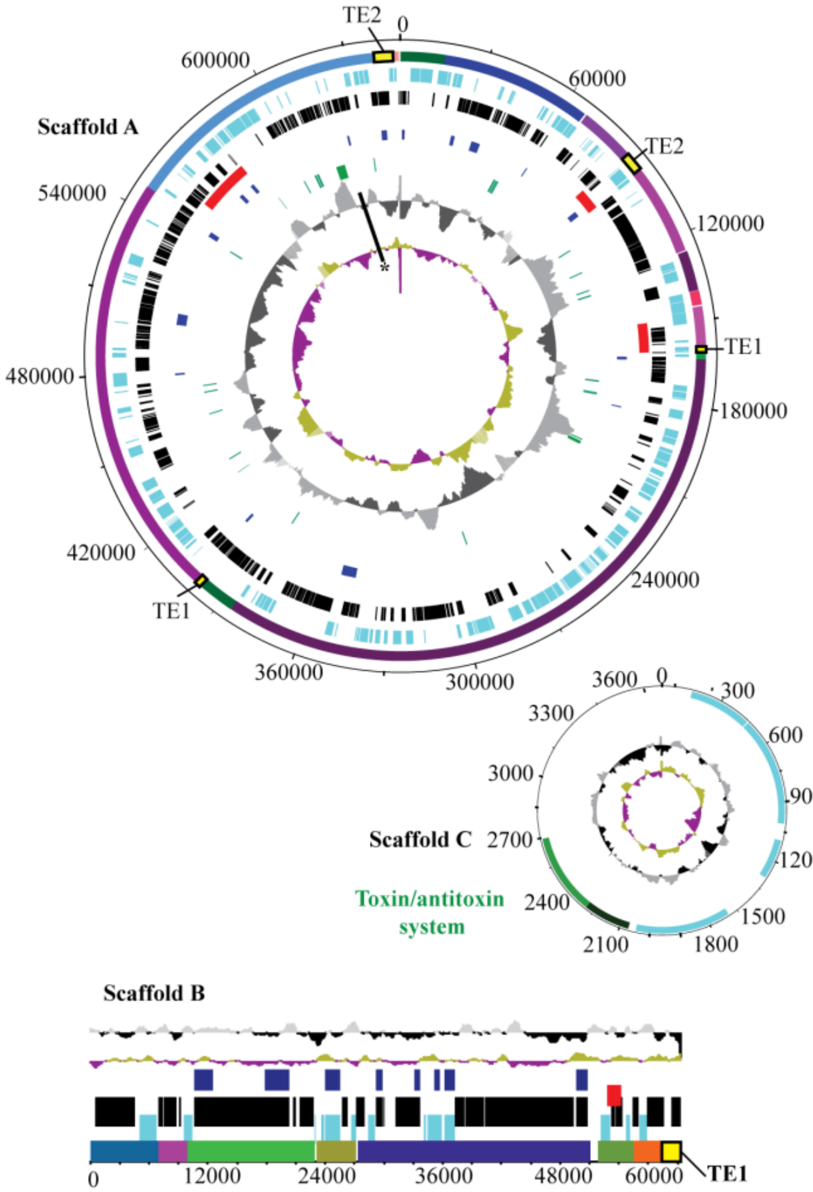

Fig. S2. Visualization of the *Dh*MRE genome draft assembly. Tracks from the outside to the inside represent: i) Contigs constituting the genome draft assembly, ii) location of gene models: in blue forward CDSs and in black reverse CDSs, iii) genomic islands in red, iv) *Dh*MRE candidate-genes for horizontal gene transfer between the AMF host and *Dh*MRE in dark blue, v) rRNA and tRNA genes in green, and in the inner tracks % GC plot and GC skew ([GC]/[G+C]). TE; transposable elements 1 and 2 are marked in yellow. *OriC predicted by GC skew in scaffold A.

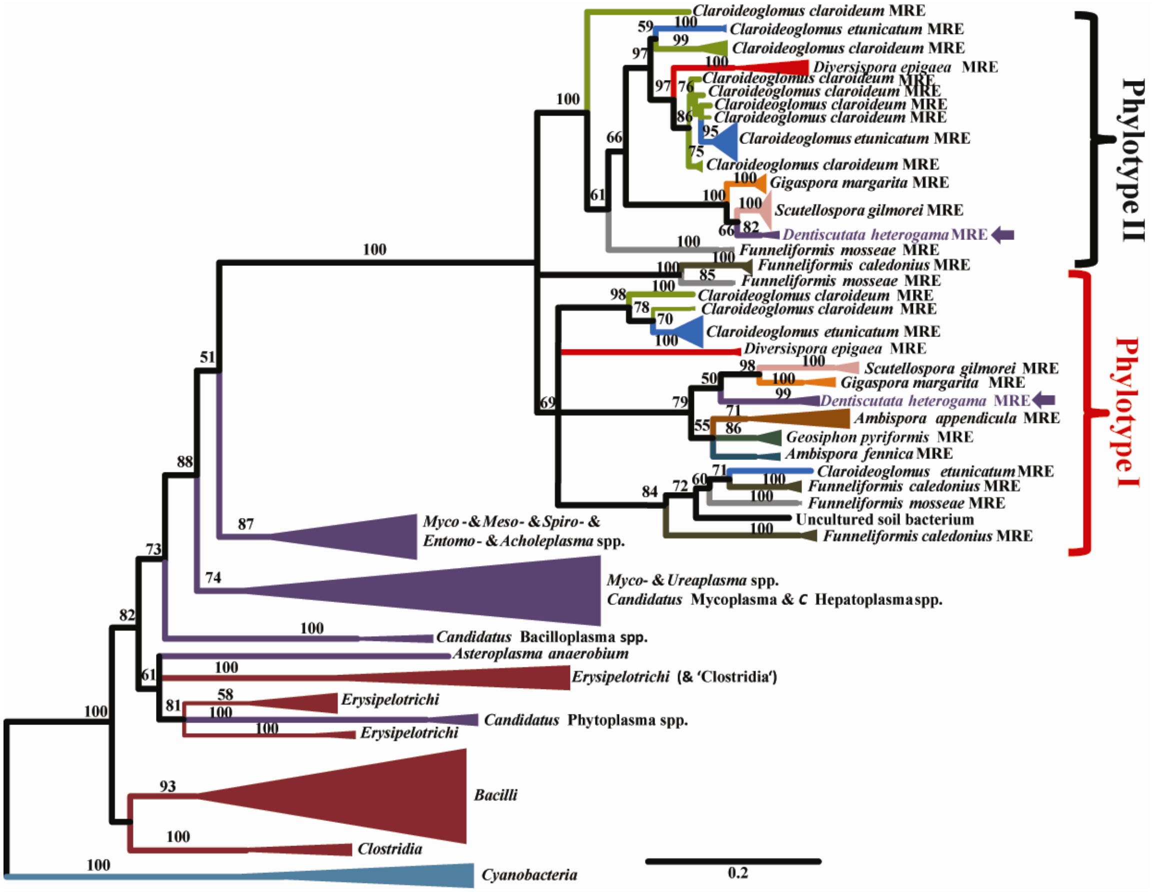
Claroideoglomus claroideum MRE
Claroideoglomus etunicatum MRE
Claroideoglomus claroideum MRE
Diversispora epigaea MRE
Claroideoglomus claroideum MRE
Claroideoglomus claroideum MRE
Claroideoglomus claroideum MRE
Claroideoglomus claroideum MRE
Claroideoglomus etunicatum MRE
Claroideoglomus claroideum MRE
Gigaspora margarita MRE
Scutellospora gilmorei MRE
Dentiscutata heterogama MRE
Funneliformis mosseae MRE
Funneliformis caledonius MRE
Funneliformis mosseae MRE
Claroideoglomus claroideum MRE
Claroideoglomus claroideum MRE
Claroideoglomus etunicatum MRE
Diversispora epigaea MRE
Scutellospora gilmorei MRE
Gigaspora margarita MRE
Dentiscutata heterogama MRE
Ambispora appendicula MRE
Geosiphon pyriformis MRE
Ambispora fennica MRE
Claroideoglomus etunicatum MRE
Funneliformis caledonius MRE
Funneliformis mosseae MRE
Uncultured soil bacterium
Funneliformis caledonius MRE
Phylotype II
Phylotype I
Myco- & Meso- & Spiro- & Entomo- & Acholeplasma spp.
Myco- & Ureaplasma spp.
Candidatus Mycoplasma & C Hepatoplasma spp.
Candidatus Bacilloplasma spp.
Asteroplasma anaerobium
Erysipelotrichi (& 'Clostridia')
Erysipelotrichi
Candidatus Phytoplasma spp.
Erysipelotrichi
Bacilli
Clostridia
Cyanobacteria
0.2

Fig. S3. Phylogeny based on the 16S rRNA gene sequences obtained from AMF spores DNA extracts. *Dh*MRE 16S rRNA gene sequences are related to the *Mollicutes* (purple); *Cyanobacteria* sequences were used as outgroup. The tree shows two main groups of MRE (phylotype I and phylotype II). MRE used in this project are indicated with an arrow and are endosymbionts of *Dentiscutata heterogama* (syn. *Scutellospora heterogama*) FL654.

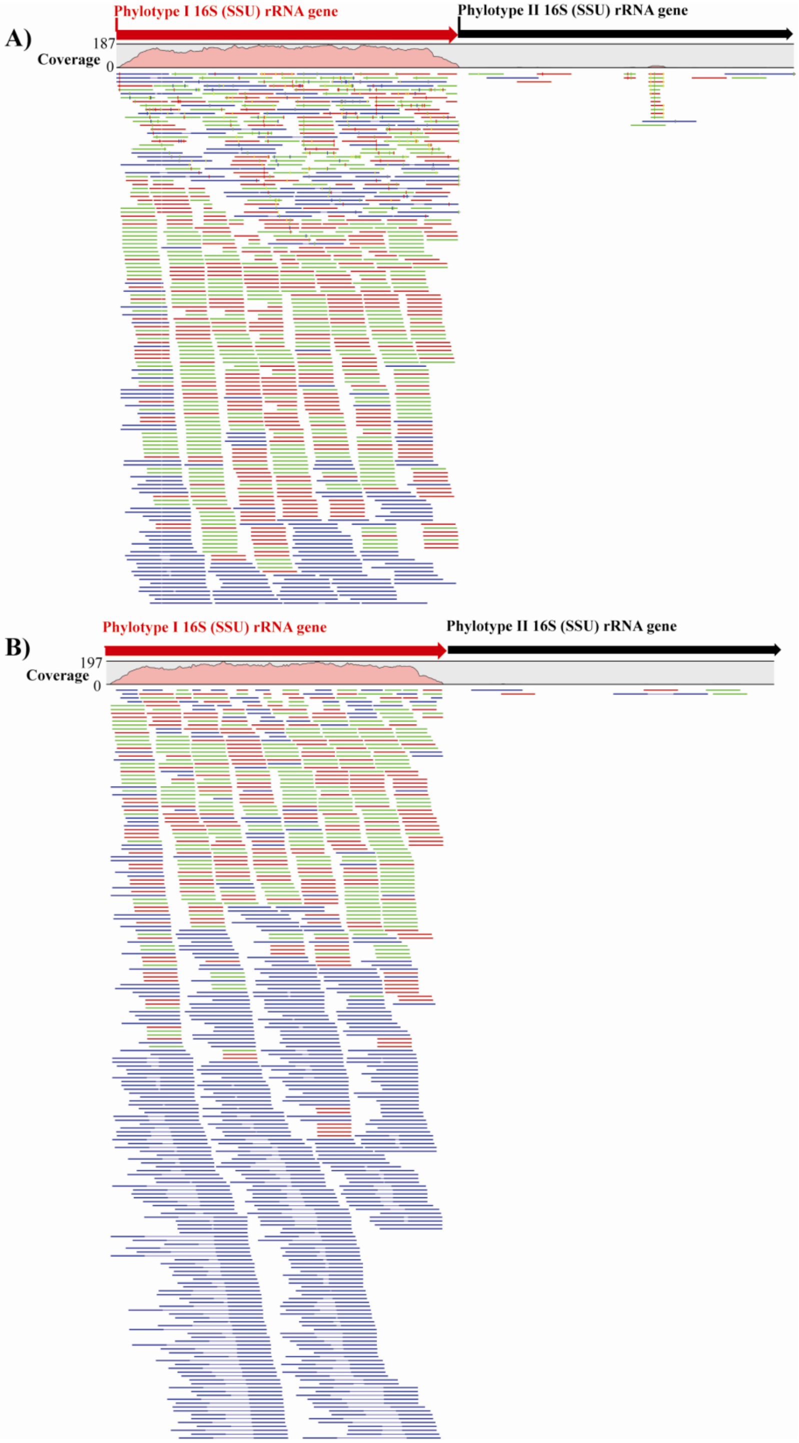

Fig. S4. Raw reads mapping against the 16S rRNA gene sequences of the two *Dh*MRE phylotypes used as a reference. Raw data reads were mapped against the reference sequences with A) 95 % and B) 100 % sequence similarity and length coverage as criteria. Green, red and blue color corresponds to forward, reverse and paired-end reads respectively. Read coverage is shown in the upper part of the figure.

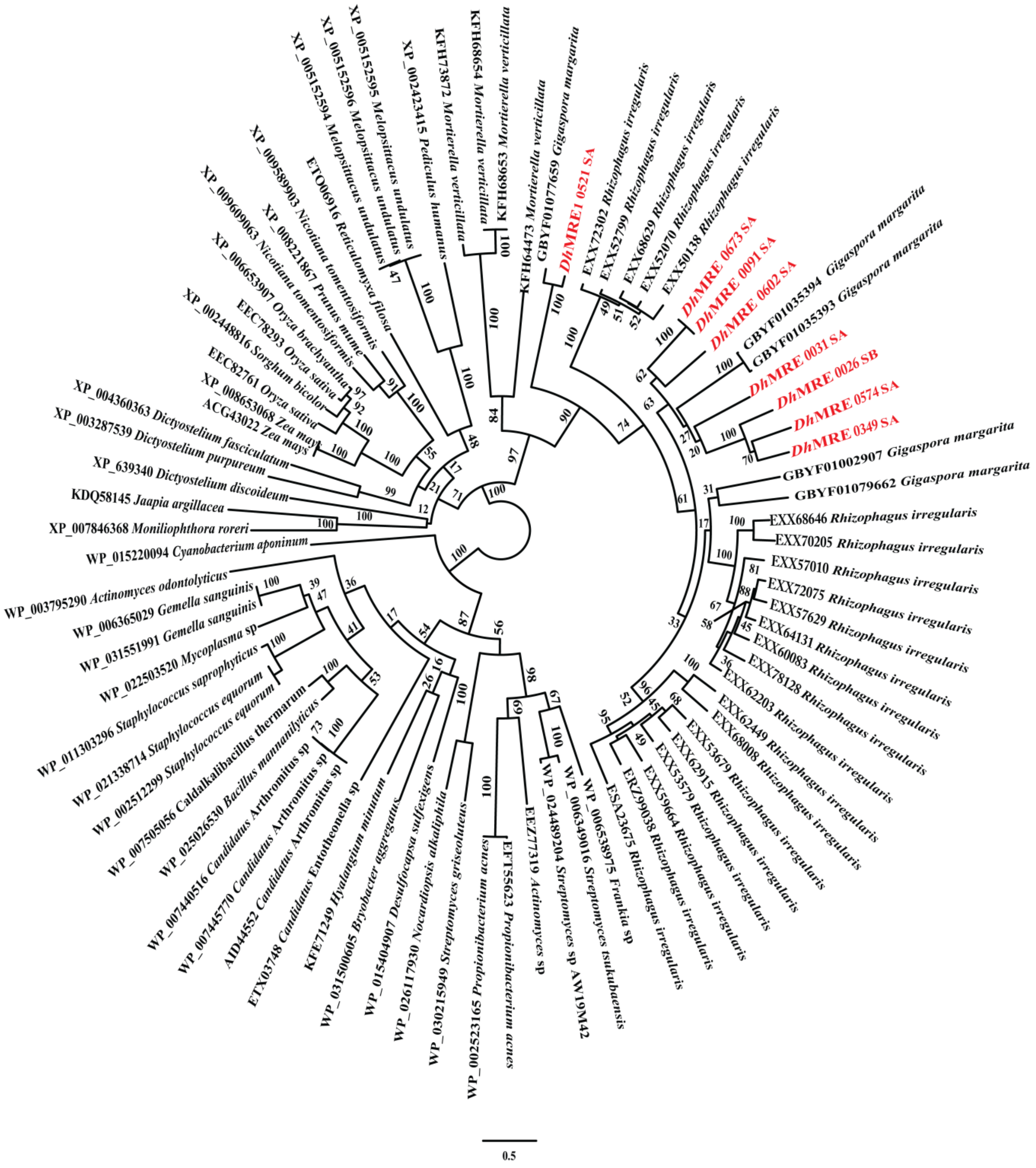
WP_015220094 Cyanobacterium aponinum
XP_007846368 Moniliophthora roreri
KDQ58145 Jaapia argillacea
XP_639340 Dictyostelium discoideum
XP_003287539 Dictyostelium purpureum
XP_004360363 Dictyostelium fasciculatum
ACG43022 Zea mays
XP_008653068 Zea mays
EEC82761 Oryza sativa
XP_002448816 Sorghum bicolor
EEC78293 Oryza sativa
XP_006653907 Oryza brachyantha
XP_009609063 Nicotiana tomentosiformis
XP_008221867 Prunus mume
XP_009589903 Nicotiana tomentosiformis
ETO06916 Reticulomyxa filosa
XP_005152594 Melopsittacus undulatus
XP_005152596 Melopsittacus undulatus
XP_005152595 Melopsittacus undulatus
XP_002423415 Pediculus humanus
KFH73872 Mortierella verticillata
KFH68654 Mortierella verticillata
KFH68653 Mortierella verticillata
KFH64473 Mortierella verticillata
GBYF01077659 Gigaspora margarita
DhMRE1 0521 SA
EXX72302 Rhizophagus irregularis
EXX52799 Rhizophagus irregularis
EXX68629 Rhizophagus irregularis
EXX52070 Rhizophagus irregularis
EXX50138 Rhizophagus irregularis
DhMRE 0673 SA
DhMRE 0091 SA
DhMRE 0602 SA
GBYF01035394 Gigaspora margarita
GBYF01035393 Gigaspora margarita
DhMRE 0031 SA
DhMRE 0026 SB
DhMRE 0574 SA
DhMRE 0349 SA
GBYF01002907 Gigaspora margarita
GBYF01079662 Gigaspora margarita
EXX68646 Rhizophagus irregularis
EXX70205 Rhizophagus irregularis
EXX57010 Rhizophagus irregularis
EXX72075 Rhizophagus irregularis
EXX57629 Rhizophagus irregularis
EXX64131 Rhizophagus irregularis
EXX60083 Rhizophagus irregularis
EXX78128 Rhizophagus irregularis
EXX62203 Rhizophagus irregularis
EXX62449 Rhizophagus irregularis
EXX68008 Rhizophagus irregularis
EXX53679 Rhizophagus irregularis
EXX62915 Rhizophagus irregularis
EXX53579 Rhizophagus irregularis
EXX59664 Rhizophagus irregularis
ERZ99038 Rhizophagus irregularis
ESA23675 Rhizophagus irregularis
WP_006538975 Frankia sp
WP_006349016 Streptomyces tsukubaensis
WP_024489204 Streptomyces sp AW19M42
EEZ77319 Actinomyces sp
EFT55623 Propionibacterium acnes
WP_002523165 Propionibacterium acnes
WP_030215949 Streptomyces griseoluteus
WP_026117930 Nocardiopsis alkaliphila
WP_015404907 Desulfocapsa sulfexigens
WP_031500605 Bryobacter aggregatus
KFE71249 Hyalangium minutum
ETX03748 Candidatus Entotheonella sp
AID44552 Candidatus Arthromitus sp
WP_007445770 Candidatus Arthromitus sp
WP_007440516 Candidatus Arthromitus sp
WP_025026530 Bacillus mannanilyticus
WP_007505056 Caldalkalibacillus thermarum
WP_002512299 Staphylococcus equorum
WP_021338714 Staphylococcus equorum
WP_011303296 Staphylococcus saprophyticus
WP_022503520 Mycoplasma sp
WP_031551991 Gemella sanguinis
WP_006365029 Gemella sanguinis
WP_003795290 Actinomyces odontolyticus
0.5

Fig. S5. Phylogenetic reconstruction based on the tyrosine kinase domain of the protein kinases described in the *Dh*Mre genome draft. Maximum likelihood phylogenetic tree computed using 100 bootstraps. SA, scaffold A; SB, scaffold B.

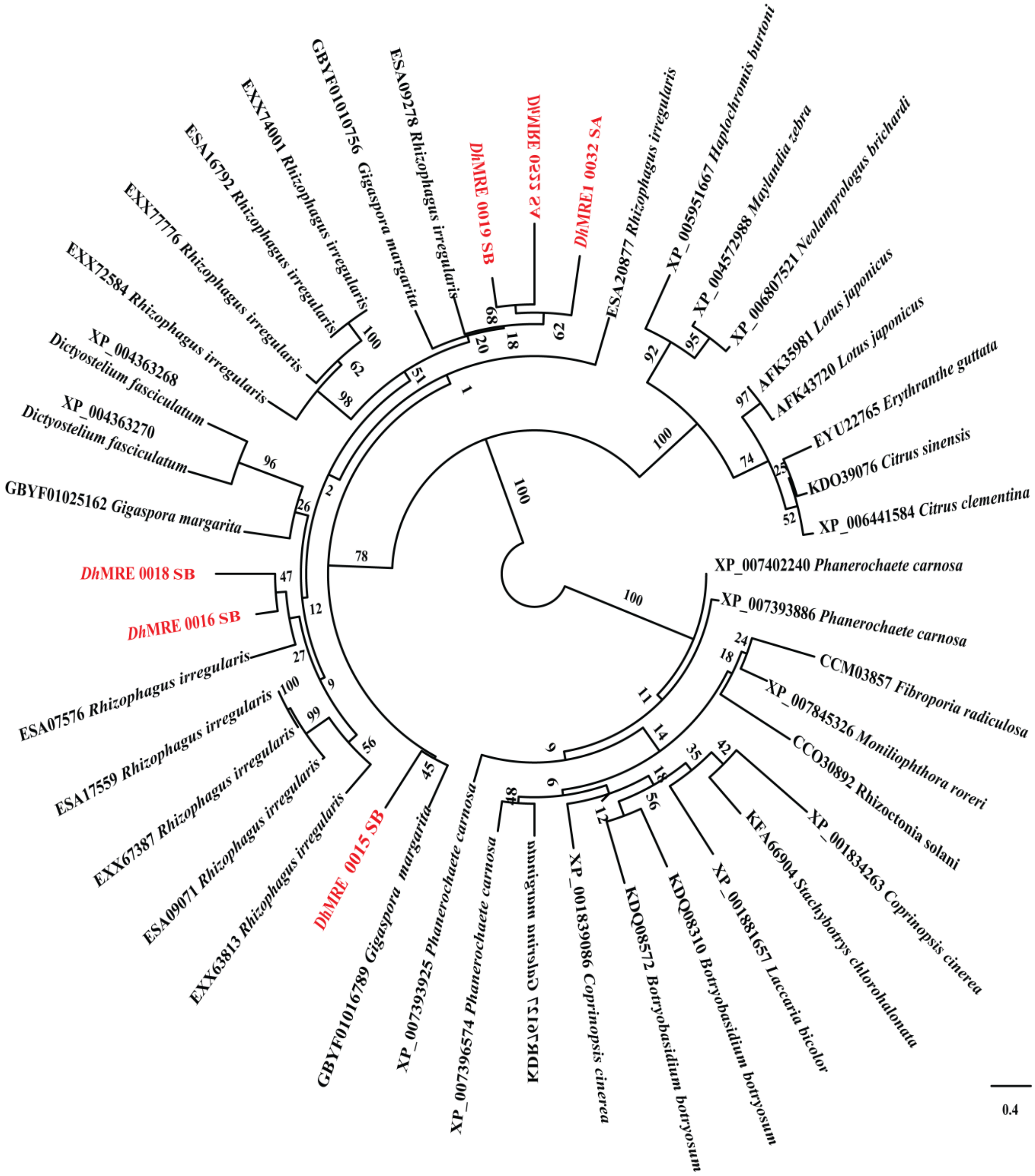
GBYF01010756 Gigaspora margarita
ESA09278 Rhizophagus irregularis
DhMRE 0019 SB
DhMRE 0022 SA
DhMRE1 0032 SA
ESA20877 Rhizophagus irregularis
XP_005951667 Haplochromis burtoni
XP_004572988 Maylandia zebra
XP_006807521 Neolamprologus brichardi
AFK35981 Lotus japonicus
AFK43720 Lotus japonicus
EYU22765 Erythranthe guttata
KDO39076 Citrus sinensis
XP_006441584 Citrus clementina
XP_007402240 Phanerochaete carnosa
XP_007393886 Phanerochaete carnosa
CCM03857 Fibroporia radiculosa
XP_007845326 Moniliophthora roreri
CCO30892 Rhizoctonia solani
XP_001834263 Coprinopsis cinerea
KFA66904 Stachybotrys chlorohalonata
XP_001881657 Laccaria bicolor
KDQ08310 Botryobasidium botryosum
KDQ08572 Botryobasidium botryosum
XP_001839086 Coprinopsis cinerea
KDR79127 Galerina marginata
XP_007396574 Phanerochaete carnosa
XP_007393925 Phanerochaete carnosa
GBYF01016789 Gigaspora margarita
DhMRE 0015 SB
EXX63813 Rhizophagus irregularis
ESA09071 Rhizophagus irregularis
EXX67387 Rhizophagus irregularis
ESA17559 Rhizophagus irregularis
ESA07576 Rhizophagus irregularis
DhMRE 0016 SB
DhMRE 0018 SB
GBYF01025162 Gigaspora margarita
XP_004363270 Dictyostelium fasciculatum
XP_004363268 Dictyostelium fasciculatum
EXX72584 Rhizophagus irregularis
EXX77776 Rhizophagus irregularis
ESA16792 Rhizophagus irregularis
EXX74001 Rhizophagus irregularis
100
68
18
62
20
51
1
62
98
100
96
26
2
78
100
92
95
100
97
74
25
52
100
24
18
11
14
35
42
9
9
18
12
56
48
6
47
12
27
9
100
99
56
45
0.4

Fig. S6. Phylogenetic reconstruction based on the AIG domain of the proteins described in the *Dh*Mre genome draft. Maximum likelihood phylogenetic tree computed using 100 bootstraps. SA, scaffold A; SB, scaffold B.

Scaffold	Gene number	Predicted function	Length (aa)	Structure
A	0149	exported PUF	417	
A	0093	PUF	322	
A	0256	PUF	336	
A	0265	PUF	381	
A	0490	PUF	136	
A	0606	PUF	794	
A	0608	PUF	507	
A	0065	PUF	430	
A	0133	PUF	315	
B	0046	PUF	770	
B	0047	PUF	2107	
B	0048	PUF	381	
A	0134	PUF	329	
A	0136	PUF	406	
A	0417	PUF	381	
A	0659	PUF	248	
B	0039	PUF	192	
B	0041	PUF	336	
B	0015	PUF	1758	
A	0487	PUF	264	
B	0015	conserved PUF	203	
B	0016	conserved PUF	447	
B	0018	conserved PUF	347	
B	0019	PUF	478	
A	0091	Non-specific PTK	812	
A	0623	PUF	107	
A	0580	Peptidase SUMO	753	
A	0603	exported Sentrin-specific protease	254	
B	0044	PUF	666	

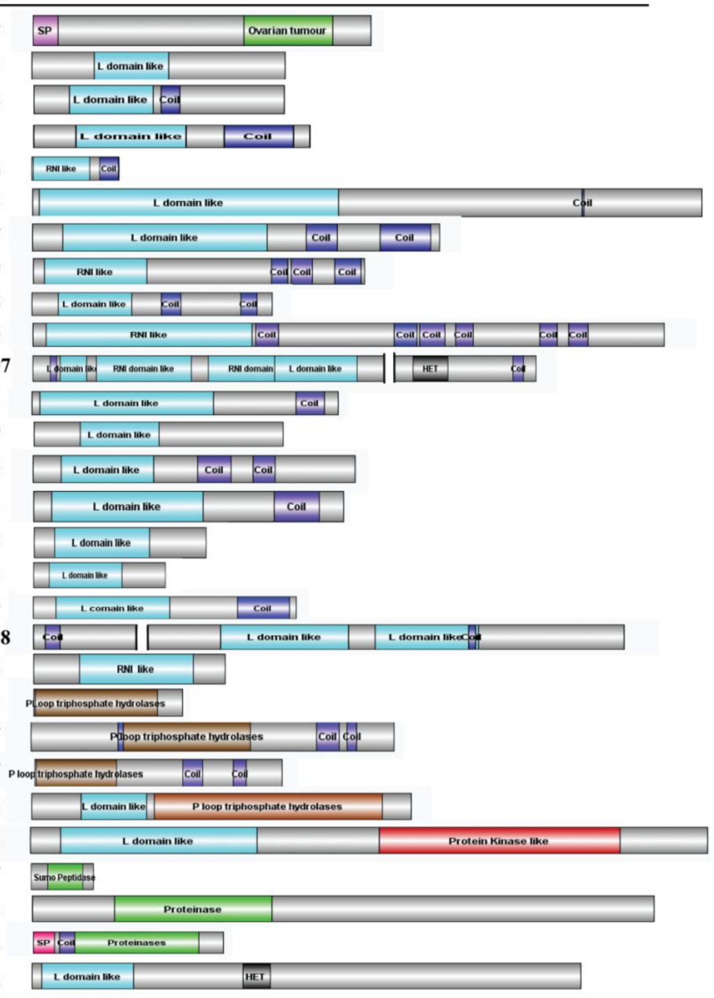

Fig. S7. Proteins with eukaryotic domains in *Dh*MRE genome. The structure of each protein was inferred from the domain description of the SUPERFAMILY database (11).

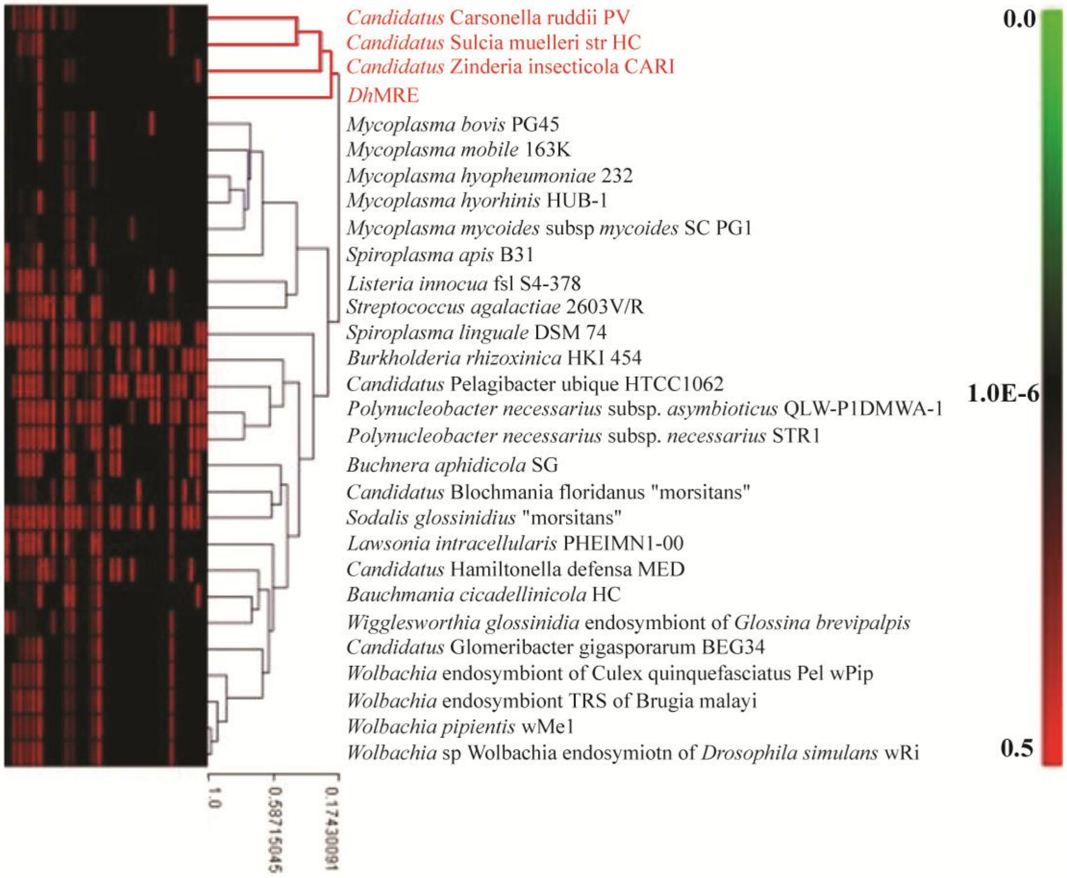

Fig. S8. Hierarchical clustering of KEGG metabolic pathways calculated for *Dh*MRE and other 29 bacteria. *Dh*MRE clusters with obligate endosymbionts of insects with reduced metabolic capacities due to its very low pathway completion values, similar to *Ca.* Carsonella rudii, *Ca.* Dulcia muelleri and *Ca.* Zinderia insecticola.

4. REFERENCES

1. Tisserant E*, et al.* (2013) Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proc Natl Acad Sci USA* 110 (50):20117-20122.
2. Albertsen M*, et al.* (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* 31(6):533-538.
3. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27(3):431-432.
4. Nijveen H, van Kaauwen M, Esselink DG, Hoegen B, Vosman B (2013) QualitySNPng: a user-friendly SNP detection and visualization tool. *Nucleic Acids Res* 41(Web Server issue):W587-590.
5. Chevreux B*, et al.* (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* 14(6):1147-1159.
6. Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner F (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* 5(1):163.
7. Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J (2009) DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* 25(1):119-120.
8. Vallenet D*, et al.* (2013) MicroScope--an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res* 41(Database issue):636-647.
9. Barré A, de Daruvar A, Blanchard A (2004) MolliGen, a database dedicated to the comparative genomics of *Mollicutes*. *Nucleic Acids Res* 32(suppl 1):307-310.
10. Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D (2009) Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biology Direct* 4(1):13.
11. Wilson D, Madera M, Vogel C, Chothia C, Gough J (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res* 35(suppl 1):308-313.