Research article

Comprehensive phylogenomics of *Methylobacterium* reveals four evolutionary distinct groups and underappreciated phyllosphere diversity.

Jean-Baptiste Leducq*[1,2], David Sneddon[2], Malia Santos[2], Domitille Condrain-Morel[3], Geneviève Bourret[3], N. Cecilia Martinez-Gomez[4], Jessica A. Lee[5], James A. Foster[2], Sergey Stolyar[2], B. Jesse Shapiro[6,7], Steven W. Kembel[3], Jack M Sullivan[2], Christopher J. Marx[2].

1 – Université Laval – Quebec City (QC) Canada
2 - University of Idaho – Moscow (ID) US
3 - Université du Québec à Montréal – Montreal (QC) Canada
4 – UC, Berkeley – Berkeley (CA) US
5 – NASA Ames Research Center – Mountain View (CA) US
6 – Université de Montréal – Montreal (QC) Canada
7 – McGill University – Montreal (QC) Canada
*Corresponding author

Abstract – *Methylobacterium* is a group of methylotrophic microbes associated with soil, fresh water, and particularly the phyllosphere, the aerial part of plants that has been well-studied in terms of physiology but whose evolutionary history and taxonomy are unclear. Recent work has suggested that *Methylobacterium* is much more diverse than thought previously, questioning its status as an ecologically and phylogenetically coherent taxonomic genus. However, taxonomic and evolutionary studies of *Methylobacterium* have mostly been restricted to model species, often isolated from habitats other than the phyllosphere, and have yet to utilize comprehensive phylogenomic methods to examine gene trees, gene content, or synteny. By analyzing 189 *Methylobacterium* genomes from a wide range of habitats, including the phyllosphere, we inferred a robust phylogenetic tree while explicitly accounting for the impact of horizontal gene transfer. We showed that *Methylobacterium* contains four evolutionarily distinct groups of bacteria (namely A, B, C, D), characterized by different genome size, GC content, gene content and genome architecture, revealing the

dynamic nature of *Methylobacterium* genomes. In addition to recovering 59 described species, we identified 45 candidate species, mostly phyllosphere-associated, stressing the significance of plants as a reservoir of *Methylobacterium* diversity. We inferred an ancient transition from a free-living lifestyle to association with plant roots in *Methylobacteriaceae* ancestor, followed by phyllosphere association of three of the major groups (A, B, D), whose early branching in *Methylobacterium* history has been heavily obscured by HGT. Together, our work lays the foundations for a thorough redefinition of *Methylobacterium* taxonomy, beginning with the abandonment of *Methylorubrum*.

Key words – *Methylobacterium*, *Methylorubrum*, species concept in bacteria, horizontal gene transfers, genome architecture, core genome, lineage tree, species tree, phyllosphere

Significance Statement

Methylobacterium is an important group of plant-associated bacteria and a model organism in microbiology. Ironically, *Methylobacterium* diversity and evolution have mostly been studied outside plants. Here, we present the first comprehensive reconstruction of *Methylobacterium* evolutionary history accounting for gene exchanges typical of Bacteria, and for diversity with known plant association. We demonstrate that *Methylobacterium* contains four evolutionarily divergent groups of bacteria, also distinguishable by their genome architecture and composition, questioning *Methylobacterium* taxonomy. We identified 104 *Methylobacterium* species, of which a large proportion is as of yet undescribed and mostly plant-associated. We also infer an ancient transition in *Methylobacterium* lifestyle from soil and plant roots, to plant leaves, stressing the significance of plants in *Methylobacterium* evolution and diversity.

Introduction

For billions of years, bacteria have evolved rapidly through vertical and horizontal gene transmission, mutation, selection, diversification, and extinction. These evolutionary processes allowed bacteria to conquer every biome and living host on Earth and, at the same

time, resulted in blurring most traces of their ancient history (Louca et al. 2018). In the past thousands of years, humans have increasingly imposed new selective pressures on bacterial evolution, through bacterial host domestication and ecosystem perturbations (Gillings & Paulsen 2014). Ironically, the human perception of microbial life was until recently limited to the diversity we could "see" (through cultivation) and "use" (through domestication), representing only an infinitesimal proportion of bacterial diversity in nature (Hugenholtz 2002). As a result, bacterial diversity, evolution and speciation concepts remain fuzzy and largely biased (Shapiro et al. 2016). Yet, the advent of high-throughput sequencing technologies, and our awakening to the essential role of bacteria in every living system has spurred research into the evolutionary processes shaping the microbial world (Koonin et al. 2021).

Methylobacterium is a well-studied group of bacteria that are abundant and widespread in every plant microbiome (Corpe & Rheem 1989; Keppler et al. 2006). *Methylobacterium* is part of *Methylobacteriaceae* (class: *Alphaproteobacteria*; order: *Hyphomycrobiales syn. Rhizobiales* (Hördt et al. 2020)), a family including three other genera, mostly isolated from aquifers and soils, sometimes in association with plant roots: *Microvirga* (Kanso & Patel 2003), *Enterovirga* (Chen et al. 2017) and *Psychroglaciecola* (Qu et al. 2014). Easy to isolate and to cultivate, thanks to a pink coloration due to carotenoids and their ability to use methanol as sole carbon source (Clarke 1983; Anthony 1991; Keppler et al. 2006), *Methylobacterium* are also essential players in plant functions, like growth stimulation (Ivanova et al. 2001; Madhaiyan et al. 2005, 2007), heavy metal sequestration (Madhaiyan et al. 2007), protection against phytopathogens and nitrogen fixation (Dourado et al. 2015), sparking increasing interest in their use in plant biotechnology applications (Ryu et al. 2006; Lee et al. 2006; Madhaiyan et al. 2007).

Recently, Green and Ardley (2018) questioned the taxonomy of *Methylobacterium*, noticing a *"greater degree of phenotypic and genotypic heterogeneity than would normally be expected for a single genus*." Accordingly, these authors proposed to split the genus in three distinct taxa corresponding to monophyletic groups in the 16S rRNA ribosomal gene phylogeny (groups A, B and C). Group A, containing the *Methylobacterium* type species *M.*

organophilum, was retained as *Methylobacterium*. For group B, which included the model species *M. extorquens*, the authors proposed a new genus: *Methylorubrum*. Finally, the authors suggested that group C, including *M. aquaticum* and *M. nodulans*, should constitute a distinct genus, pending future genetic and phenotypic investigations. The *Methylobacterium* reclassification has been pointed out as problematic, because of the low phylogenetic resolution of the 16S rRNA gene, and because no genus name was proposed for strains that were not retained in *Methylorubrum* or *Methylobacterium*, which could potentially render either new genus as paraphyletic (Hördt *et al.*, 2020; Leducq *et al.*, 2022). Accordingly, the taxonomy of *Methylobacterium* was reexamined by coupling genome-wide DNA-DNA hybridization and phenotypic information for 63 strains, each representative of a described species (Alessa et al. 2021). Alessa *et al.* confirmed Green and Ardley's (2018) observation that group C was phenotypically and genetically distinct from other groups, but they also showed that *Methylorubrum* (group B) was embedded within *Methylobacterium* (group A), forming a homogeneous group, and proposed to merge *Methylobacterium* and *Methylorubrum* back into a single genus.

The evolutionary history of *Methylobacterium* remains poorly resolved for several reasons. First, phylogenetic relationships among and within groups are often inconsistent depending upon the chosen marker gene (Green and Ardley, 2018; Leducq *et al.*, 2022). Such inconsistent phylogenetic signals suggest that these marker genes had different evolutionary histories, perhaps due to horizontal gene transfer (HGT) or incomplete lineage sorting (ILS), illustrating the dynamic nature of bacterial genome evolution and the limitations of bacterial taxonomy based on a limited number of gene phylogenies (Castillo-Ramírez & González 2008; Creevey et al. 2011). Second, Alessa *et al.* (2021) based their *Methylobacterium* taxonomy on DNA-DNA hybridization methods, which are widely used to classify prokaryotic species, but are not phylogenetic methods *per se*, as they do not account for ancestry. They also validated their taxonomy using a phylogenetic tree based on concatenated protein sequences of core genes but did not present evaluations of the uncertainty in the resulting tree. Finally, phylogenies based on concatenated gene alignments assume the same tree for each gene, and thus do not take into account potential ILS and HGT affecting topology and branch lengths differentially in each individual gene trees. With the

onset of genomics in evolutionary studies, several coalescent methods have been developed to reconstruct the phylogeny and solve the taxonomy of organisms with complex evolutionary history like bacteria (Davidson et al. 2015). For instance, coalescent-based phylogenetic methods like ASTRAL-III (Mirarab et al. 2014; Zhang et al. 2018) and SVDquartets (Chifman & Kubatko 2014) allow the reconstruction of a consensus tree (the lineage tree) taking into account different levels of ILS and HGT among individual gene trees.

Although more than 60 *Methylobacterium* species have been described so far (Green & Ardley 2018; Chen et al. 2019; Feng et al. 2020; Jia et al. 2020; Kim, Chhetri, Kim, Lee, et al. 2020; Kim, Chhetri, Kim, Kim, et al. 2020; Ten et al. 2020; Jiang et al. 2020; Pascual et al. 2020; Alessa et al. 2021), available genomic and phenotypic information was until recently limited to a few model species, mostly from groups B and C, and mostly isolated from anthropogenically impacted environments, and in rare cases from plants (Marx et al. 2012; Tani et al. 2015; Minami et al. 2016; Morohoshi & Ikeda 2016; Belkhelfa et al. 2018). Surveys of *Methylobacterium* diversity associated with plants have mainly focused on the rhizosphere, especially in crop species (Sy et al. 2001; Jourand et al. 2004; Grossi et al. 2020). Recent studies however revealed that the phyllosphere of model plant species like *A. thaliana* (Helfrich et al. 2018), of wheat (Zervas et al. 2019), and of natural temperate forests (Leducq et al. 2022) are major reservoirs of undescribed *Methylobacterium* diversity, most of which belongs to group A (Leducq et al. 2022).

Here, we explored *Methylobacterium* diversity from an evolutionary genomic perspective. We *de novo* annotated 189 *Methylobacterium* genomes, including 62 strains isolated from temperate forest, wheat, and *Arabidopsis* phyllosphere, and 127 additional genomes that represent the remainder of the *Methylobacterium* species described so far. Using different phylogenomic approaches, we reconstructed the *Methylobacterium* evolutionary tree from 384 *Methylobacteriaceae* core genes and showed that the genus is consistently constituted of four monophyletic groups: A, B, C and D. Gene content and especially the highly dynamic core genome architecture predicted the four *Methylobacterium* groups remarkably well. We estimated that *Methylobacterium* includes at least 104 species, of which only 59 were

previously described. Most of the undescribed species were assigned to groups A and D and were isolated from plant leaves, stressing the significance of the phyllosphere as a reservoir of *Methylobacterium* diversity. Our inferences of the *Methylobacterium* evolutionary tree also suggest an ancient transition from a free-living lifestyle to association with plant roots in *Methylobacteriaceae* ancestor, followed by phyllosphere association of three of the major groups (A, B, D), whose early branching in *Methylobacterium* history was heavily obscured by HGT. Finally, our comprehensive phylogenetic analysis of *Methylobacterium* lays the foundation for a profound redefinition of its taxonomy, beginning with the abandonment of *Methylorubrum*.

Results

Definition *of the* Methylobacteriaceae *core genome.*

We assembled a collection of 213 *Methylobacteriaceae* genomes, including 189 *Methylobacterium* and 24 genomes from related genera as outgroups (*Microvirga*: n=22; *Enterovirga*: n=2). Most *Methylobacterium* (*n*=98) and all outgroup genomes (n=24) came from distinct studies (Dataset S1). We included 29 genomes from *Methylobacterium* type strains recently sequenced (Alessa et al. 2021; Bijlani et al. 2021), hence covering most *Methylobacterium* species described so far. We also included 38 genomes available from two large surveys of the *Arabidopsis* and wheat phyllospheres (Helfrich et al. 2018; Zervas et al. 2019), and sequenced 24 additional genomes of isolates from a large survey of the temperate forest phyllosphere (Leducq et al. 2022), hence extending our dataset to the leaf-associated *Methylobacterium* diversity. The 24 newly assembled genomes had 41 to 405 scaffolds (depth: 188-304x) for a total size (5-7Mb) and average GC content (67-70%) in the expected range for *Methylobacterium* genomes (Dataset S2). We annotated 184 genomes *de novo*, excluding 29 genomes that were not published at the time of the analysis (Alessa et al. 2021; Bijlani et al. 2021) through the same pipeline (RAST) and after excluding hypothetical proteins, repeat and mobile elements, we identified 9,970 unique gene annotations (i.e., regardless of copy number: Dataset S3), with on average 2637 (SD: 210) unique gene annotations per genome. We identified 893 candidate core genes, i.e., genes that were

present in a single copy in at least 90% of *Methylobacteriaceae* genomes. After filtering for missing data and false duplications attributable to large variations among genome assembly qualities (Figures S1, S2), we identified 384 *Methylobacteriaceae* core genes (Dataset S4) for which the complete nucleotide sequences could be retrieved for at least 181 out of 184 genomes. We repeated the RAST annotation for recently sequenced genomes from 29 *Methylobacterium* species type strains that were not available during our initial survey (Alessa et al. 2021; Bijlani et al. 2021). Doing this, we slightly extended the number of unique gene annotation in *Methylobacteriaceae* (n = 10,190). We confirmed that the 384 previously identified genes were part of the *Methylobacteriaceae* core genome and retrieved each core gene nucleotide sequence for at least 26 out of these 29 genomes. Our final dataset consisted of 213 genomes for which we retrieved 327 to 384 core genes nucleotide sequences (average, SD: 381 ± 6).

Inference of the Methylobacteriaceae *lineage tree*

We reconstructed the lineage tree of *Methylobacteriaceae* from 213 genomes from the 384 core gene nucleotide sequences using three complementary approaches in order to assess the effect of ILS and HGT in the evolutionary history of *Methylobacterium*. First, we used RAxML to determine a maximum-likelihood tree (512 replicated tree; bootstraps) from concatenated alignments of the core gene nucleotide sequences, assuming 57 groups of genes (partitions) with different substitution models (partitions determined in IQ-tree; GTRCAT model of substitution), but the same evolutionary tree for all genes, hence not accounting for ILS or HGT (Figure 1a). Second, we used ASTRAL, a coalescent-based method combining Maximum-Likelihood (ML) trees determined for each core gene independently (RAxML, GTRGAMMA model, 1,000 replicated tree), accounting for ILS among genes (Figure 1b, Figure S3a). Third, we used SVDquartets, a coalescent-based method estimating the tree for each possible combination of four genomes and assuming all nucleotide sites are unlinked in the concatenated alignment of 384 genes, hence accounting for ILS and HGT both within and among genes (Figure 1c, Figure S3b). In all lineage trees rooted on *Microvirga* and *Enterovirga*, *Methylobacterium* was monophyletic and consisted of four groups of genomes, consistently monophyletic and strongly supported, regardless of the method used (nodal

support: 100% in RAxML and SVDquartets trees; local posterior probability: 1.0 in the ASTRAL tree; Figure 1a,b,c). Group C always formed the most basal group of *Methylobacterium*, confirming previous observations (Green & Ardley 2018; Alessa et al. 2021; Leducq et al. 2022). Group B regrouped clades B, formerly *Methylorubrum* (Green & Ardley 2018) and B2 (Alessa *et al.* 2021; corresponding to clade A4 in Leducq *et al.* 2022). Most strains previously assigned to clade A (Green & Ardley 2018) were distributed across two distinct monophyletic groups that we named A and D (Figure 1a). Group A included clades A2, A3, A4 and A5 described by Alessa *et al.* (2021) and corresponded to clades A5, A10, A19 and A7+A8 described by Leducq *et al.* (2022), respectively. Group D corresponded to clade A1 proposed by Alessa *et al.* (2021) and clades A1, A2 and A3 proposed by Leducq *et al.* (2022).

Groups A, B and D consistently formed a monophyletic group (nodal support: 100% in RAxML and SVDquartets trees; local posterior probability: 1.0 in the ASTRAL tree); however, phylogenetic relationships among groups A, B and D were more challenging to assess. Groups A, B, and D could not be resolved with the RAxML tree (nodal support = 9%; Figure 1a). Group D was sister to groups A and B according to ASTRAL (local posterior probability: 0.8; Figure 1b) and SVDquartets trees (nodal support: 100%; Figure 1c). We evaluated differences between the three lineage tree topologies using the Robinson-Foulds (RF) distance metric in PAUP (Wilgenbusch & Swofford 2003). RAxML and ASTRAL lineage tree topologies were more similar to each other (RF = 0.181) than with the SVDquartets tree (RF = 0.225 and 0.289, respectively; Figure S4). In order to determine whether the difference between the RAxML and other trees was higher than expected by chance, we estimated the distribution of RF distance between each replicate tree of the RAxML search for the lineage tree (512 replicates). The normalized RF value ranges from 0.028 to 0.113 (RF= 0.069±0.015), indicating that the differences observed between lineage trees were larger than expected by chance (Figure S4), and suggesting that ILS and HGT among core genes had a significant impact on the *Methylobacteriaceae* lineage tree. The larger difference between the SVDquartets tree and other trees also suggested that recombination within core genes also occurred during *Methylobacteriaceae* evolution, although without affecting the relationship among the four major groups (C/D/(A,B)).

Inference of the Methylobacteriaceae *taxonomy and species tree*

We classified *Methylobacteriaceae* genomes into 124 species using a 97% threshold on percentage nucleotide similarity (*PNS*; analogous to average nucleotide identity; (Mende et al. 2013; Chun & Rainey 2014); Dataset S5) on the core genome (concatenated alignments of 384 core genes; 361,403 bp). In the outgroups, we identified 2 *Enterovirga* species and 18 *Microvirga* species. We identified 104 *Methylobacterium* species (1 to 9 genomes per species), of which 59 included the type strain for at least one described species (Table 1; Dataset S5). *M. extorquens*, *M. chloromethanicum* and *M. dichloromethanicum* type strains were assigned to the same species (*PNS* range: 97.61-99.68%), as previously reported (Alessa et al. 2021). *M. populi* and *M. thiocyanatum* type strains were assigned to the same species (*PNS* range: 98.97%-99.08%), as previously reported (Alessa et al. 2021). *M. phyllosphaerae*, *M. ozyzae* and *M. fujisawaense* type strains were assigned to the same species (99.23-100%), as previously reported (Alessa et al. 2021). We identified 45 candidate species that included no type strain, and thus corresponded to new candidate *Methylobacterium* species (Table 1; Dataset S5). We numbered these candidate species from *Methylobacterium sp.* 001 to 045. We used the 124 identified species to infer the *Methylobacteriaceae* species trees with SVDquartets (Figure S5a) and ASTRAL (Figure S5b). Although the two species trees were not strictly identical (normalized RF distance = 0.234), the monophyly and relationships among the four main groups was consistent between ASTRAL and SVDquartets species trees (C/D/(AB); Figures S5a,b), and with ASTRAL and SVDquartets lineage trees (Figure 1b,c). Each group of genomes assigned to the same species was also monophyletic and strongly supported in lineage trees (Figure 1).

In summary, the *Methylobacterium* species are distributed across four groups, each of which with somewhat distinct environmental sources of isolation (plant phyllosphere and rhizosphere, water and sediments, soils, others), as well as the proportion of strain isolated from anthropogenic environments (Table 1, Figure 1d). Group A contained 62 genomes which fell into 41 species, including 17 new species (*Methylobacterium sp.* 018 to 034). Group B contained 41 genomes which fell into 21 species, including 7 candidate species

(*Methylobacterium sp.* 035 to 041). Group C contained 25 genomes which fell into 19 species, including 4 new candidate species (*Methylobacterium sp.* 042 to 045; Table 1). Group D contained 42 genomes which fell into 23 species, including 17 new candidate species (*Methylobacterium sp.* 001 to 017). Species from *Microvirga* and *Enterovirga* were mostly isolated from soil samples (65% of species; corrected by the number of genomes per species), often in association with plant roots (Rhizosphere; 30%). Species from *Methylobacterium* groups B and C were isolated from plants (40 and 31% of genomes, respectively), soil samples (13 and 32%), sediments or water samples (18 and 21%), often in association with anthropogenic environments (29 and 49%). Species from groups A and D were mostly isolated from plants (62 and 75% of species, respectively), especially the phyllosphere (51 and 67%). Of the 45 new candidate *Methylobacterium* species, most were assigned to groups A (n=17) and D (n=17); the majority (81%) was isolated from plants, and especially the phyllosphere (66%; Table 1; Figure 1d).

Genome comparison across Methylobacterium *groups*

The four main *Methylobacterium* groups have consistently contrasting genome characteristics (Figure S6, Table 2). These four groups have significantly different genome sizes (Tukey test, $p<0.001$), with group D having smaller genomes (4.99 ± 0.35 Mb; Average ± SD), than groups B (5.58 ± 0.49 Mb), A (6.21 ± 0.59 Mb) and C (7.15 ± 0.66 Mb). Groups D and B had a smaller number of annotated genes (5,224 ± 476 and 5,766 ± 509, respectively) than groups A and C (6,907 ± 821 and 7,670 ± 956, respectively; $p<0.001$). The average number of gene annotation copies per genome was significantly different among groups ($p<0.001$) and was smaller for group D (1.31 ± 0.04 copies per annotation) than for group B (1.37 ± 0.05), A (1.46 ± 0.07) and C (1.54 ± 0.07). GC content was significantly lower in groups D and B (68.8 ± 1.1 and 69.1 ± 0.8 %, respectively) than in group A (70.1 ± 0.8 %; $p<0.001$) or group C (71.1 ± 0.7 %; $p<0.001$; Figure S6, Table 2). Although the number of mobile elements per genome was slightly lower in group D (42 ± 21) compared to A (60 ± 34), B (57 ± 31) and C (71 ± 49), these differences were not significant (Table 2; Tukey test, $p>0.05$). We compared the abundance of 10,187 gene annotations (excluding hypothetical proteins, repeat elements and mobile elements) across the four

311 *Methylobacterium* groups and outgroups (Figure 2). *Methylobacterium* genomes clustered
312 according to their gene content and abundance and matched the ASTRAL species tree
313 (Figure 2a). As observed for other genome characteristics, group D had the smaller pan
314 genome size (n = 4,217 ± 70; estimation assuming rarefaction of 15 species per group, mean
315 and standard deviation over 100 replicates; Figure S7a), followed by group B (n = 4,973 ±
316 137), group A (n = 4,974 ± 132) and group D (n = 5,636 ± 91 genes; Figure 2b). On the
317 contrary, group D had a larger core genome size (i.e. gene present in a single copy in all
318 species; n = 1,103 ± 29 core genes) than groups A (n=845 ± 79), B (n=924 ± 65) and C
319 (n=843 ± 39; Figure 2c; Figure S7b). Venn diagrams on shared annotations indicate a limited
320 overlap of gene content among groups, with only 2,863 ± 38 pan genes shared among the
321 four groups (Figure 2b) and 350 ± 32 core genes (Figure 2c).
322
323 *Gene content comparison across* Methylobacterium *groups*
324
325 We next asked to what extent gene content evolved concordantly along the core genome
326 phylogeny. We used the Bray-Curtis index to measure the pairwise dissimilarity among
327 genomes based on their gene annotation abundance (*BC*; Hellinger normalization of gene
328 abundance; Figure S8). The dissimilarity matrix in gene content among species matched the
329 species tree (Figure S8). Gene content was more similar among genomes from the same
330 *Methylobacterium* species (*BC* range: 0.044 ± 0.017 - 0.080 ± 0.023) than among species
331 within *Methylobacterium* groups (*BC* range: 0.159 ± 0.031- 0.197 ± 0.044) or than among
332 *Methylobacterium* groups (*BC* range: 0.238 ± 0.025 - 0.339 ± 0.019; Figure S8, Table 3). We
333 determined the relationships among *Methylobacteriaceae* members upon their gene content
334 using a ML phylogeny based on the occurrence of the 10,187 gene annotations across 213
335 genomes (RAxML assuming a BINCAT model; 1,001 replicate trees; Figures 2d, detailed
336 tree in Figure S9a). The gene content tree supported each of the 124 *Methylobacteriaceae*
337 species, as well as the monophyly of groups B, C and D (nodal support: 99, 94 and 87%,
338 respectively). Groups A, B and D formed a monophyletic group (nodal support: 100%),
339 making group C the most basal group, as observed for lineage (Figure 1) and species trees
340 (Figure S5). Most of the species assigned to group A clustered together (nodal support: 77%)
341 but five species formerly assigned to clade A2 (*M. planium*, *M. soli*, *M. oxalidis*, *M. durans*,

M. segetis; Alessa *et al.*, 2021) and *M. jeotgali* were more similar to groups B and D, which altogether formed a monophyletic group (nodal support: 97%). The normalized RF value between the gene content tree and lineage trees (Figure 1) ranged from 0.429 to 0.469. As a comparison, normalized RF values between the best gene content tree and its 1,001 replicate trees ranged from 0.085 to 0.249 (RF= 0.169±0.026), indicating that the gene content tree had significantly different topology than lineage trees.

Core genome architecture comparison (synteny) across Methylobacteriaceae *genomes*

We next evaluated the level of conservation in the architecture of the *Methylobacteriaceae* core genome to assess the extent of chromosomal rearrangement during *Methylobacterium* evolution. Most genomes (177 out of 213) were not fully assembled (i.e., the chromosome consisted of more than one scaffold), and we thus inferred the order or the 384 core genes along the chromosome of draft genomes by aligning their scaffolds to the chromosomes of 36 completely assembled *Methylobacteriaceae* genomes, while conserving the order of core genes within scaffolds. We compared the order of core genes among genomes using a synteny index (*SI*) calculated as the proportion of pairs of core genes (links) that were neighbors in two genomes, ranging from 0 (no link conserved) to 1 (fully conserved synteny; Figure S8, Table 4). The matrix of synteny among species was generally concordant with the species tree (Figure S8). In the 213 *Methylobacteriaceae* genomes, we observed 6,109 different links among the 384 core genes. Core genome architecture was well conserved among genomes from the same *Methylobacterium* species (*SI* range: 0.914 ± 0.064 - 0.995 ± 0.007) but was highly reshuffled among species within *Methylobacterium* groups (*SI* range: 0.608 ± 0.118 - 0.769 ± 0.207; Figure S8, Table 4). As a comparison, the core genome architecture among *Microvirga* species was remarkably well conserved (*SI* = 0.913 ± 0.048). Average synteny among *Methylobacterium* groups A, B, C and D was low (*SI* range: 0.433 ± 0.025 - 0.528 ± 0.049) and in the same order of magnitude as synteny between *Methylobacterium* and *Microvirga* genomes (*SI* range: 0.458 ± 0.010 - 0.525 ± 0.020; Figure S8, Table 4). We identified *M. planium* (strain YIM132548, group A) as the species having, on average, the highest core genome synteny with other *Methylobacterium* genomes. Accordingly, we used *M. planium* as a reference to visualize the conservation of the 384

links identified in its genome across *Methylobacterium* species (Figure 3a; Figure S10). We identified 150 links (involving 231 genes; 60.2% of core genes) that were mostly conserved among *Methylobacteriaceae* genomes. With the exception of a remarkably well-conserved cluster of 26 genes that included ribosomal genes and gene *rpoB* (Figure 3a; Figure S10), most of the 150 conserved links were scattered across the *M. planium* chromosome. We determined the relationships among *Methylobacteriaceae* members in their core genome architecture using a ML phylogeny based on the occurrence of 6,109 links identified across 213 genomes (RAxML assuming a BINCAT model; 1,001 replicate trees; Figure 3b, detailed tree in Figure S9b). The synteny tree supported the monophyly of the four major *Methylobacterium* groups (nodal support = 100%). Groups A, B and D formed a monophyletic group (nodal support: 83%), making group C the most basal group, as observed for lineage trees (Figure 1), species trees (Figure S5) and the gene content tree (Figure 2d). The normalized RF value between the synteny tree and lineage trees ranged from 0.589 to 0.638. As a comparison, normalized RF values between the best synteny tree and its 1,001 replicate trees ranged from 0.235 to 0.390 (RF= 0.310 ± 0.026), indicating that the synteny tree had a significantly different topology than lineage trees. Interestingly, although *M. planium* and related species previously assigned to clade A2 (*M. soli*, *M. oxalidis*, *M. segetis*, *M. durans*; Alessa *et al.* (2021)) as well as *M. jeotgali* and *M. trifolii* were assigned to clade A in the ML synteny tree (Figure 3b), these species had on average higher synteny with species from group D (*SI* = 0.651 ± 0.045) than with other species from group A (*SI*= 0.556 ± 0.031; Figure 3c). Accordingly, we identified 29 links involving 54 core genes that were more often conserved between groups A and D than with other *Methylobacterium* groups. These links, however, were scattered along the *M. planium* chromosome (Figures 3a, S10).

Discussion

Methylobacterium *consists of four evolutionarily divergent groups of bacteria*

Recent work has suggested that *Methylobacterium* is much more diverse than thought previously, questioning its genus status (Green & Ardley 2018; Hördt et al. 2020; Alessa et

al. 2021; Leducq et al. 2022). Here, we used a comprehensive phylogenomic approach to provide unprecedented insight on the taxonomic diversity of *Methylobacterium*. Our reconstructions of the *Methylobacteriaceae* lineage tree based on the core genome confirmed previous comparative genomic and phenotypic studies that group C, including *M. nodulans* and *M. aquaticum*, form a distinct and cohesive group at the root of the *Methylobacterium* phylogeny (Green & Ardley 2018; Hördt et al. 2020; Alessa et al. 2021). On the contrary, we demonstrated that Group B, including the model species *M. extorquens*, and previously amended as a distinct genus, *Methylorubrum* (Green & Ardley 2018), formed a monophyletic group with the *Methylobacterium* type species *M. organophilum* and other species formerly assigned to group A (e.g. *M. oxalidis* and *M. planium*) (Green & Ardley 2018). Our analyses hence support the proposal to extend group B to *M. organophilum*, *M. oxalidis*, *M. planium*, and relatives (Alessa et al. 2021). Although the newly defined group B was monophyletic according to our different inferences of the *Methylobacteriaceae* species, it was still embedded within former group A (Green & Ardley 2018), making the later paraphyletic, and confirming that *Methylorubrum* cannot be considered as a distinct genus without breaking apart *Methylobacterium* (Hördt et al. 2020; Alessa et al. 2021). Accordingly, we support the proposal to abandon "*Methylorubrum*" as a designation for group B, and to split group A into two monophyletic groups distinct from group B: group A (including *M. brachiatum*, *M. komagatae*, *M. cerastii*, *M. jeotgali*, *M. trifolii*, *M. planium* and relatives) and group D (including *M. bullatum*, *M. gossipicola*, *M. goesingense*, *M. iners* and relatives).

We observed that the newly defined monophyletic groups (A, B, C and D) were characterized by distinct genome sizes and GC content, two metrics that were highly correlated with each other in *Methylobacterium*, as observed in other bacteria (Nishida 2012). With the exception of a few species from group A (including, *M. trifolii*, *M. jeotgali*, *M. planium* and relatives), the four groups could also be distinguished upon their gene content. GC content, genome size and gene content are widely accepted as criteria for taxonomic definition in prokaryotes (Rosselló-Mora & Amann 2001; Coenye et al. 2005). We also demonstrated that core gene order was highly reshuffled among the four *Methylobacterium* groups. For instance, we observed the same level of rearrangement in core

gene order among *Methylobacterium* groups, as between *Microvirga* and *Methylobacterium*, and the same level of core gene order conservation within *Methylobacterium* groups as within *Microvirga*. Core gene order has recently been proposed as a complementary criterion to define bacteria genus and species taxonomy (Chung et al. 2018). The fact that the four groups were monophyletic, regardless of whether we used a concatenated or a coalescent-based approach to infer the *Methylobacteriaceae* lineage tree and could be consistently distinguished from each other upon different genome characteristics (gene content, core genome architecture, GC content, genome size), supports considering them as distinct genera.

Role of HGT and ILS in the early divergence of groups A, B and D

The evolution of bacteria is marked by recurrent HGT, gene duplication and loss events, making the reconstruction of bacterial phylogenies challenging. Given that each gene potentially has its own evolutionary history, marked by exchanges among divergent taxa, the evolutionary tree of most bacteria is quite reticulate (Shapiro et al. 2016). The reconstruction of a consensus phylogenetic tree (lineage tree) thus remains highly conceptual in bacteria and could only be achieved by considering a pool of genes assumed to be representative of the prevailing evolutionary history of the considered taxa: the core genome (Sakoparnig et al. 2021). Therefore, HGT and ILS must be considered when attempting to reconstruct bacterial phylogeny. Accordingly, we showed that the concatenated-based reconstruction of *Methylobacterium* lineage tree, assuming the same evolutionary history for each core gene, significantly differed in its topology from lineages tree reconstructions accounting for ILS and/or HGT among core genes (ASTRAL and SVDquartets lineage trees), indicating that both processes were major drivers of *Methylobacterium* evolution. While the concatenated tree suggested that groups A and D formed a monophyletic group, coalescent-based estimations from ASTRAL (ILS + HGT among genes) and SVDquartets (ILS + HGT among sites) rather indicated the earlier divergence of group D from the A/B/D group.

A possible explanation of the divergent topology in the concatenated lineage tree is that shared polymorphism was retained by ILS and/or HGT between groups A and D after the

A/B divergence. Interestingly, although supporting the four *Methylobacterium* groups, our phylogeny reconstructed from core genome architecture suggested the closer relationship between groups A and D, in agreement with the concatenated lineage tree, hence supporting the hypothesis of horizontal core gene exchanges having occurred between groups A and D after the A/B divergence. Accordingly, we observed syntenic islands (groups of neighbor core genes) shared between group D and some basal species of group A (*M. jeotgali*, *M. trifolii*, *M. planium*, *M. oxalidis* and relatives). These islands were scattered across the *Methylobacterium* chromosome, either suggesting that extensive chromosomal rearrangements occurred after HGT between A and D, or that HGT occurred multiple times during their evolutionary history, potentially among divergent lineages. According to a phylogeny reconstructed from gene occurrence in *Methylobacterium*, *M. jeotgalii*, *M. planium*, *M. oxalidis* and relatives, belonged to different lineages branching at the root of groups A, B and D, supporting the hypothesis of multiple and independent gene exchanges among distinct *Methylobacterium* lineages after the divergence of the three groups, blurring their phylogenetic relationships.

Outstanding Methylobacterium *diversity: the role of the phyllosphere?*

Methylobacterium is frequently associated with the phyllosphere, yet taxonomic and phylogenomic surveys of its diversity have mostly focused on human-impacted environments such as food factories, contaminated soils, air conditioning systems or even the International Space Station. Here we presented the first comprehensive genomic survey of *Methylobacterium* diversity in the phyllosphere. By including genomes of strains isolated from the phyllosphere of wheat (Zervas et al. 2019), of the model plant *A. thaliana* (Helfrich et al. 2018), and of trees from natural temperate forests (Leducq et al. 2022), our phylogenomic analysis of *Methylobacterium* revealed that its evolutionary and taxonomic diversity was larger than previously thought. In addition to recovering the 59 previously described species (Alessa et al. 2021), we identified 45 new (candidate) *Methylobacterium* species, of which a majority belonged to groups A and D, and were mostly isolated from the phyllosphere. Beyond taxonomic considerations, this result reveals a profound bias in our understanding of natural processes underlying the existing diversity of *Methylobacterium*,

497 and more generally, of bacteria. For example, the evolutionary distinction between groups A
498 and D, and their importance in *Methylobacterium* diversity, could not have been revealed
499 without a thorough investigation of diversity in the phyllosphere, from which the majority of
500 candidate species from groups A, B and D were isolated. A recent survey of
501 *Methylobacterium* in metagenomes from various biomes (Lee et al. 2022) also suggested the
502 association of groups A (represented by *M. pseudosasicola* and *M. radiotolerans* in Lee *et*
503 *al.*, 2022 study), B and especially D (represented by *M. gossipiicola* and *Methylobacterium*
504 *sp.* Leaf 88) with the aerial part of plants. Similarly, we recently showed that groups A and D
505 were the dominant *Methylobacterium* groups in the phyllosphere of trees from temperate
506 forests (Leducq et al. 2022). On the contrary, groups B and C included most
507 *Methylobacterium* model species frequently used in the lab and isolated from anthropogenic
508 environments. While group B is occasionally identified on and isolated from the surface of
509 leaves (Leducq et al. 2022; Lee et al. 2022), group C is rarely, if ever, found in the
510 phyllosphere, and seems to be more widespread in soil and in aquatic environments, often in
511 association with plant roots (Lee et al. 2022). Interestingly, authors from a recent study
512 estimated that *Rhizobiales* common ancestor likely had a free-living lifestyle, while
513 *Methylobacterium* groups A, B and D's common ancestor likely had a plant-associated
514 lifestyle (node 1 in Figure 1 from Wang *et al.* study (Wang et al. 2020)). The ancestral
515 lifestyle of *Methylobacterium*, and more widely, of *Methylobacteriaceae*, is more unclear.
516 The isolation source of group C genomes, as well as the two sister genera of
517 *Methylobacterium*, *Enterovirga* and *Microvirga*, and their survey in metagenomes (Lee et al.
518 2022) indicate that these three groups are mostly found with soils, sometimes in association
519 with the rhizosphere. These observations suggest that *Methylobacteriaceae* and
520 *Methylobacterium*'s ancestors inhabited soils, and were occasionally associated with plants,
521 for instance in the rhizosphere, and that *Methylobacterium* groups A/B/D's association with
522 the phyllosphere occurred after divergence from group C. The exact origin and nature of this
523 association is an open question, but the smaller genome size, gene copy number, GC content
524 and to a lower extent, mobile element number, we observed in group A/B/D in comparison
525 with group C could be the genomic signatures of a progressive specialization to life on plants
526 (Nishida 2012; Levy et al. 2018), among other things through the evolution of metabolic
527 pathways in response to contrasted nutrient availability between the soil and the phyllosphere

(Lee et al. 2022; Alessa et al. 2021). For instance, some genes involved in the metabolism of aromatic compounds resulting from lignin degradation are present in *Microvirga* and *Methylobacterium* group C, but absent from other *Methylobacterium* groups (Lee et al. 2022), suggesting that these functions essential for ground lifestyle were lost in A/B/D group after they divergence with group C. Inversely, *Methylobacterium* from group A/B/D arbor a larger panoply of genes allowing the use of methanol, available in the phyllosphere, than group C, while most of these pathways are absent from *Microvirga* (Alessa et al. 2021), suggesting that the transition from soil to phyllosphere lifestyle in *Methylobacterium* also coincided with the acquisition and diversification of methylotrophic pathways.

According to our phylogenomic analyses of group A/B/D, group D diverged first, and, like group A, was mostly isolated from the phyllosphere, suggesting that the A/B/D ancestor inhabited the surface of plant leaves. The fact that our analyses support horizontal gene exchanges between groups A, B and D is also consistent with the hypothesis that these groups lived in the same habitat during their divergence. One can speculate that some horizontally transferred, yet to be discovered, genes may have had shared roles in *Methylobacterium* adaptation to the phyllosphere. For instance, strains from groups A and D were often identified in the same studies, sometimes isolated from the same plants, indicating that strains from these two groups likely share the same microhabitats on the surface of plant leaves, hence favoring gene exchanges among them and the maintenance of similar molecular pathways and functions. Further identifications of genes exchanges among these groups and the characterization of their functions will be critical to understand evolutionary mechanisms underlying the adaptive role and radiation of *Methylobacterium* in the phyllosphere.

Conclusion - Our unprecedented phylogenomic analysis of *Methylobacterium* revealed the outstanding diversity within this taxon, and the role of HGT in its early evolutionary history. Future genomic and functional studies will be needed to characterize the evolutionary and functional features of *Methylobacterium* adaptation to the phyllosphere. Finally, our work lays the foundation for a thorough taxonomic redefinition of this genus.

Methods

Methylobacteriaceae *genome collection*

We assembled a collection of 213 complete and draft *Methylobacteriaceae* genomes, including 189 *Methylobacterium* and 24 genomes from related genera as outgroups (*Microvirga*: n=22; *Enterovirga*: n=2). Most *Methylobacterium* (*n*=98) and all outgroup genomes (n=24) came from distinct studies (see references in Leducq *et al.* (2022)) and corresponded to genomes publicly available in October 2020 on NCBI. We included 29 genomes from *Methylobacterium* type strains recently published (Alessa et al. 2021; Bijlani et al. 2021) in order to cover most *Methylobacterium* species described so far. We also included 38 genomes available from two large surveys of the *Arabidopsis* and wheat phyllospheres (Helfrich et al. 2018; Zervas et al. 2019) and sequenced 24 additional genomes (**see next section**) of isolates from a large survey of the temperate forest phyllosphere (Leducq et al. 2022), hence extending our dataset to the leaf-associated *Methylobacterium* diversity.

Library preparation and genome assembly of 24 Methylobacterium *strains*

We performed genome sequencing and *de novo* assembly of 24 *Methylobacterium* strains representative of the diversity previously found in the phyllosphere of two temperate forests in the province of Québec, Canada (Leducq *et al.* (2022); Dataset S2). DNA extraction was performed from culture stocks frozen at -80 °C directly after isolation and identification (Leducq et al. 2022) and thawed 30 min on ice. About 750 µl of cell culture were used for DNA extraction with DNeasy PowerSoil Pro Kit (Qiagen) according to the manufacturer protocol, with the following modification: final elution was repeated twice in 25 µl (total volume: 50 µl). 300 bp paired-end shotgun libraries were prepared from 35 ng genomic DNA with QIAseq FX DNA Library Kit (Qiagen) and protocol was adjusted to target DNA fragments in the range 400-1000 bp. Genomes were assembled from libraries with MEGAHIT (Li et al. 2015) with default parameters. Genome assemblies had 7050-24785 contigs with average depth in the range 188-304x and a total size in the range 7.2-17.1 Mb.

After removing contigs with depth <10x, we obtained 82-411 contigs per genome. Most assemblies had total size (5-7 Mb) and average GC content (67-70%) in the expected range for *Methylobacterium* genomes (Dataset S2). For three out of twenty-four genomes, GC content and depth distribution were clearly bimodal, and total size was much higher (9.5-11.9 Mb) suggesting that these assemblies contained genomes from at least two evolutionary distinct taxa. For these three heterogeneous assemblies, we divided contigs into two pools based on median depth value between two modes (threshold range: 100-150x). For each heterogeneous assembly, the pool with highest average depth (174-241x) had average GC content (67-68%) and total size (5.6-5.8 Mb) in the ranges expected for *Methylobacterium*. Contigs with lower depth were considered as contaminants and discarded from assemblies.

Gene annotation

Methylobacteriaceae genomes (n=213) were individually annotated using RAST (https://rast.nmpdr.org/rast.cgi) (Aziz et al. 2008) with following parameters: genetic code=11; Annotation scheme=RASTtk; Preserve gene calls=no; Automatically fix errors=yes; Fix frameshifts=no; Backfill gaps=yes. Annotation output from each genome was retrieved separately as Spreadsheet (GFF file in tab-separated text format). Core genome definition was conducted in R (R-Developement-Core-Team 2011). For each genome, we retrieved the abundance of gene annotations (column *function* in RAST output), excluding *Hypothetical proteins*, *repeat regions* and *Mobile element proteins* (Dataset S3).

Methylobacteriaceae *core genome definition*

We first defined the *Methylobacteriaceae* core genome from 184 genomes, excluding 29 genomes that were not yet published nor annotated at the time of the analysis (Alessa et al. 2021; Bijlani et al. 2021). In these 184 genomes, we identified 9,970 unique gene annotations (i.e., regardless copy number: Dataset S3), with on average 2637 ± 210 unique gene annotations per genome. We defined candidate core genes as genes present in one copy in at least 90% of the 184 genomes, resulting in 893 candidate core genes, for which we retrieved the nucleotide sequence (column *nucleotide_sequence* in RAST output). In order to

correct for false gene duplication events that increased consistently with assembly incompleteness (Figure S1) and to estimate the actual copy number of each candidate core gene, we used 36 complete *Methylobacteriaceae* genomes as references (defined as genomes with $N50 > 3\times10^6$ Mb). For each candidate core gene, we calculated the average expected nucleotide sequence size observed among 36 complete genomes. Then, for each genome (n=184) and each candidate core gene, we retrieved all nucleotide sequences (0-10 per gene and genome) and calculated their average size normalized (divided) by the average nucleotide sequence size observed in complete genomes. By doing so, we could distinguish between duplication caused by genome incompleteness (single copy genes divided between different scaffolds) and real duplication events (Figure S2). We considered 398 genes for which at least one genome had more than one copy with normalized size >0.75 as true duplicates and removed them from candidate core genes. For the 495 remaining candidate core genes, we considered single-copy genes with normalized size >1.3 and gene copies with normalized size <0.7 (regardless copy number) as missing data in the considered genome. After this filter, we removed 111 candidate core genes that were missing in at least 4 genomes, resulting in 384 core genes for which a single full-length copy could be retrieved for at least 181 genomes (out of 184; Dataset S4). Subsequently, we included recently sequenced genomes from 29 *Methylobacterium* species type strains that were missing from our survey (Alessa et al. 2021; Bijlani et al. 2021). By doing this, we slightly extended the number of unique gene annotations in *Methylobacteriaceae* (n = 10,190). We confirmed that the 384 previously identified core genes were part of the *Methylobacteriaceae* core genome and retrieved each core gene nucleotide sequence for at least 26 out of 29 genomes. Our final dataset consisted of 213 genomes for which we retrieved 327 to 384 core genes nucleotide sequences (381 ± 6; mean, SD; Dataset S1).

Core gene nucleotide sequence alignments

We performed an alignment for each core gene. For each genome (n = 184 + 29 = 213), we first extracted nucleotide sequences of the 384 core genes (when not missing data for the considered genome; column *nucleotide_sequence* in RAST output) and converted them in sequence fasta files using R package *seqinr()*. We then performed an alignment for each core

gene using R packages *seqinr()* and *msa()*. For each gene, nucleotide sequences were translated (function *getTrans()*) and alignments of amino-acid sequences were performed using ClustalW with default parameters in function *msa()*. Sequences were converted back in nucleotides (stop codons excluded) and 5' and 3' end codons with more than 90% of missing data (gaps of "Ns") were trimmed. We also constructed an alignment of concatenated core genes nucleotide sequence alignments. In the concatenated alignment, sequences of genes missing for at least one of the 213 genomes (0-6 genomes missing per gene) were replaced by strings of "Ns".

Inferences of the Methylobacteriaceae *lineage trees*

We reconstructed the lineage tree of *Methylobacteriaceae* from 213 genomes from the 384 core gene nucleotide sequences using three complementary approaches in order to assess the effect of ILS and HGT in the evolutionary history of *Methylobacterium*.

First, we used RAxML v. 8.2.8 (Stamatakis 2014) to determine a maximum-likelihood (ML) lineage tree from concatenated alignments of the core 384 gene nucleotide sequences assuming a different substitution model for each gene but the same evolutionary tree for all genes (and hence not accounting for ILS or HGT). We used PartitionFinder implemented in IQ-tree2 (Minh et al. 2020) to determine an appropriate bipartitioning scheme allowing us to merge genes evolving under similar nucleotide substitution models (Lanfear et al. 2012). The best-fit partition scheme was determined using TESTMERGERONLY model (option *–m*) to avoid tree reconstruction, and using the relaxed hierarchical clustering algorithm to reduce the computation burden (Lanfear et al. 2014) by only examining the top 10% partition merging schemes (option *–rcluster*). We then inferred the *Methylobacteriaceae* lineage tree from the 384 core gene alignment with RAxML v. 8.2.8 (Stamatakis 2014), using the IQ-tree2 best-scheme output file as partition file (option *–q* in RAxML). The program performed 512 replicate (bootstrap) searches from independent starting trees with a GTRCAT model of substitution, estimating parameters for each partition separately. Of the 512 trees, the one with the highest ML score (the best-scoring tree) was retained as the lineage tree.

683 Second, we used ASTRAL-III (Zhang et al. 2018), a coalescent-based method inferring the
684 lineage and the species trees by combining individual core gene trees, hence accounting for
685 ILS and HGT among genes. For each core gene, a gene tree was first inferred from
686 nucleotide sequence alignments with RAxML v. 8.2.8 (Stamatakis 2014). Briefly, for each
687 gene, the program performed 1,000 replicate (bootstrap) searches from independent starting
688 trees assuming a GTRgamma model of nucleotide substitution. Each gene tree in Newick
689 format, including branch length (*L*: nucleotide substitution per site) and node label (*N*: nodal
690 support representing the proportion of replicated supporting nodes), was imported in R as a
691 vector. The gene tree in RAxML format: *((():L1[N1]):L2[N2])* was rewritten so that it could
692 be readable in R (package *ape* (Paradis & Schliep 2019)) and ASTRAL-III:
693 *((():L1)N1:L2)N2*. The tree was then reopened in R with function read.tree (package *ape*)
694 and nodes with < 10% support were collapsed using function *collapseUnsupportedEdges*
695 (package *ips*), to optimize accuracy in estimating the lineage and species tree (Zhang et al.
696 2018). All reformatted gene trees were written in a single file (*multiPhylo* object), which was
697 used to infer the lineage and the species tree in ASTRAL-III v5.7.7, with default parameters.
698 In ASTRAL trees, branch lengths were measured in coalescent units and nodal support
699 represented local posterior probability (Sayyari & Mirarab 2016).
700
701 Third, we used SVDquartets (Chifman & Kubatko 2014) as implemented in PAUP* v4.0a
702 (build 169) (Wilgenbusch & Swofford 2003), a coalescent-based method estimating the tree
703 for each possible combination of four genomes and assuming all sites unlinked in the
704 concatenated alignment of 384 genes. We estimated the lineage tree from the concatenated
705 213 *Methylobacteriaceae* core genes by evaluating 2,000,000 random quartets for 100
706 bootstrap replicates. Phylogenies were estimated under the multispecies coalescent model
707 accounting for incomplete lineage sorting (ILS) and assessing all sites independently to
708 account for recombination within and among loci.
709
710 Lineage trees were displayed in Figtree v1.4.4 and rooted on *Microvirga* and *Enterovirga.*
711
712 Methylobacteriaceae *species definition and lineage tree inferences*
713

We classified *Methylobacteriaceae* genomes in species using percentage nucleotide similarity (*PNS*) on the core genome (concatenated alignments on 384 core genes; 361,403 bp), similar to average nucleotide identity (Mende et al. 2013; Chun & Rainey 2014). Briefly, *PNS* between two genomes was calculated in R as the proportion of conserved nucleotide positions, gaps and "Ns" excluded (Dataset S4). Two genomes were considered from the same species when their *PNS* was higher or equal to 97%, a threshold similar to what is typically used for bacterial species (96.5% based on nucleotides sequences of 40 marker genes; (Mende et al. 2013; Chun & Rainey 2014)). We inferred the *Methylobacteriaceae* species tree using both ASTRAL-III (Zhang et al. 2018) and SVDquartets (Chifman & Kubatko 2014), as described above, using individual assignment to species determined from PNS. Species trees were displayed in Figtree v1.4.4 and rooted on *Microvirga* and *Enterovirga.*

Test for HGT and ILS severity in lineage and species tree inferences

We tested for the severity of HGT and ILS in our dataset by measuring the differences in tree topologies estimated using different assumptions. To quantify differences between the topologies obtained under different assumptions, we calculated normalized Robinson-Foulds (RF) distances (Robinson & Foulds 1981), which evaluates the pairwise proportion of unique nodes between tree topologies, between all three lineage trees (RAxML from concatenated core gene alignments, SVDquartets, and ASTRAL) as well as between both species trees (ASTRAL and SVDquartets). RF distances were estimated using the treedist function implemented within PAUP* v4.0 (build 169) (Wilgenbusch & Swofford 2003) using final phylogenies in NEWICK format as input. We also calculated the distribution of RF distances between our best RAxML tree from concatenated core gene alignments and all 512 RAxML bootstrap replicates. We then compared our RF distances between each inference method to this distribution of distances to assess whether discordant topologies among lineage trees are due to different assumptions of methods, or due to phylogenetic uncertainty.

Methylobacteriaceae *genome characteristics and gene content*

745 We analyzed *Methylobacteriaceae* genome characteristics (size, GC content and gene
746 content) of the coding sequence for each. We first calculated the number of gene annotations,
747 their total nucleotide size (coding genome size), their GC content, the number of unique
748 annotations (excluding hypothetical proteins, repeat elements and mobile elements), the
749 number of mobile elements, and the average copy number of annotations (Dataset S1, Table
750 2). For each statistic, we compared *Methylobacterium* groups (as defined by lineage trees)
751 and outgroups (*Microvirga*, *Enterovirga*) using a Tukey test.
752
753 In a heatmap, we displayed the average abundance of 10,187 gene annotations (excluding
754 hypothetical proteins, repeat elements and mobile elements) per *Methylobacteriaceae*
755 species, per *Methylobacterium* group and outgroup, ordered according to the ASTRAL
756 species tree (Figure S5a). For gene abundance per species, we calculated the average
757 occurrence (n) of each gene annotation across genomes assigned to the same species,
758 rounded to 0 ($n<0.5$), 1 ($0.5\leq n<1.5$) or 2 copies ($n\geq 2$). For gene abundance per group, we
759 calculated the average occurrence of each gene annotation across species assigned to the
760 same group, using the same principle as for species.
761
762 We estimated pan genome and core genome sizes per *Methylobacterium* group (unknown
763 proteins, repeat and mobile elements excluded; Figure S7). To deal with biases in size
764 estimations due to the variable number of genomes per group, we applied rarefaction on gene
765 number estimates by randomly sampling 1 to N genomes per group (Park et al. 2019) and by
766 forcing resampling of one genome per species. For each N value and each group, we
767 calculated the average and standard deviation in core genome size (genes in 1:1 copy in each
768 genome of a given group) and in pan genome (any gene present in at least one copy in at
769 least one genome of a given group) over 100 replicates. As pan genome size estimations
770 increased with the number of sampled species per group (Figure S7a) and core genome size
771 estimations decreased (Figure S7b), curves of estimates per group did not cross each other,
772 nor reached a plateau, indicating that sizes were either under-estimated (pan genomes) or
773 over-estimated (core genome) but could still be compared among groups. Accordingly, we
774 compared pan and core genome sizes among *Methylobacterium* groups in a Venn diagram,
775 assuming 15 species per group (Figures 2b,c).

776

777 We constructed a phylogeny of *Methylobacteriaceae* based on gene content (Figure 2d).
778 First, we constructed a matrix of gene occurrence in each genome (0 for absence and 1 for
779 presence) and converted it into a fasta file (one sequence per genome). We inferred an
780 evolutionary tree of based on gene content using with RAxML v. 8.2.8 (Stamatakis 2014).
781 The program performed 1,000 replicate (bootstrap) searches from independent starting trees
782 with a BINCAT model of substitution assuming gene presence of absence as binary data. Of
783 the 1,000 replicate trees, the one with the highest Maximum-likelihood (ML) score (the best-
784 scoring tree) was considered as the best tree. The tree was displayed in Figtree v1.4.4 and
785 rooted on *Microvirga* and *Enterovirga.*
786

787 In order to compare *Methylobacteriaceae* genomes based upon their gene content, we
788 calculated an index of dissimilarity among each pair of genomes from their gene abundance
789 (Table 3, Figure S8b). As no index was available for this purpose, we used the Bray-Curtis
790 (*BC*) index of dissimilarity, initially developed in ecology for the comparison of
791 communities based on their species abundance (Bray & Curtis 1957). To minimize the effect
792 of higher copy number due to false gene duplications due to genome incompleteness, we
793 applied a normalization on gene abundances (Hellinger normalization; function *decostand* in
794 R package *vegan*). We calculated pairwise *BC* indexes of dissimilarity among normalized
795 gene abundances, using function *vegdist* in R package *vegan*.
796

797 *Core genome architecture (synteny)*
798

799 We evaluated the level of conservation in core gene organization (synteny) between
800 *Methylobacteriaceae* genomes. All analyses were performed in R (R-Developement-Core-
801 Team 2011). For each genome, we retrieved core gene coordinates (scaffold name and
802 coordinates in the scaffold). For complete genomes consisting of a single linear scaffold
803 (n=36), each core gene was paired with its two immediate neighbors, based on shortest
804 distance between gene *start* and *stop* coordinates, and core genes located on scaffold edges
805 were also paired together, assuming genome circularity. Hence, for each complete genome,
806 each of the 384 core genes was involved in two links (pairs of neighbor core genes), for a

807 total of *N*=384 links per genome. For the 177 draft genomes, core genes were located on
808 different scaffolds, so we predicted scaffold order and orientation using complete genomes as
809 references. The draft genome with the highest completeness was reorganized first. Briefly,
810 for each comparison with a reference genome, and for each scaffold of the draft genome, the
811 list of embedded core genes was retrieved, and a score based on gene average order in the
812 reference genome was calculated. Scaffolds of the draft genome were reordered according to
813 these scores (one per scaffold). Then, each scaffold was eventually reoriented (without
814 affecting gene order within scaffolds) to optimize pairing of edge genes (genes located at the
815 edge of a scaffold) as compared to the reference genome. We repeated the operation for
816 comparisons with the 36 reference genomes. Finally, for each of the 36 new configurations
817 of the draft genome, we calculated a synteny conservation index (*SI*) with each reference
818 genome, as the proportion of links conserved between two genomes. *SI* ranged from 0 (no
819 link conserved) to 1 (fully conserved synteny). The draft genome configuration with the
820 highest *SI* value was conserved for further analyses and added to the list of reference
821 genomes. We repeated this operation for each draft genome, ranked according to their
822 decreasing completeness, hence increasing the number of reference genomes and possible
823 configurations for highly fragmented genomes. Finally, we calculated *SI* for all possible
824 pairwise comparisons between genomes (Figure S8b), and average and standard deviation
825 values within and among *Methylobacteriaceae* species and groups (Table 4).
826
827 In order to visualize the spatial organization of core genes along the *Methylobacteriaceae*
828 chromosome, we used two approaches. First, we realized a heatmap of link conservation per
829 species, along a reference genome (Figure S10). We choose as reference the genome having
830 the highest average *SI* value with other *Methylobacterium* genomes. In the heatmap, we
831 displayed the 384 links identified in the reference genome, ordered according to core gene
832 order along its chromosome, and highlighted them when also present in other
833 *Methylobacteriaceae* species. For each species, we also reported the average *SI* value with
834 the reference genome. Finally, for each link in the reference genome, we calculated its
835 frequency in each *Methylobacterium* group. Second, we drew a consensus map of the
836 *Methylobacterium* core genome architecture, as well as major rearrangements within and
837 among *Methylobacterium* groups, as a network in Cytoscape v.3.4.0 (Shannon et al. 2003)

(Figure 3a). In this network, we represented the 384 core genes as nodes, ordered according to *M. planium YIM132548* core genome, and links among neighbor core genes as edges. The network was drawn using 389 links observed in a majority (>50%) of species from a given *Methylobacterium* group (5,720 links discarded).

In order to reconstruct the evolution of *Methylobacteriaceae* core genome based on its architecture, we constructed a matrix of occurrence of each possible link observed among genomes (0 for absence and 1 for presence) and converted it into a fasta file. We inferred an ML evolutionary tree of *Methylobacteriaceae* based on synteny using with RAxML v. 8.2.8 (Stamatakis 2014) with a BINCAT model of substitution assuming pair of core genes presence of absence as binary data, as described for annotations (Figure 3b).

Data availability statement

Draft genomes for 24 *Methylobacterium* strains corresponding to Bioproject PRJNA730554 (Biosamples listed in Dataset S1) were deposited on NCBI under accession numbers JAKSXU000000000 - JAKSYR000000000. R code and related data were deposited on Github (https://github.com/JBLED/methylobacterium-phylogenomics.git).

Acknowledgments

This research received funding from the National Science Foundation to C.J.M., N.C.M.-G., J.A.F, J.M.S., and J.A.L (DEB-1831838).

We thank Benjamin Oswald and the IBEST Computational Resources Core at University of Idaho for their technical support, and Alex Alleman, as well as two anonymous reviewers for their helpful suggestions on the manuscript.

J.-B. L., D. S., M. S. and J.M.S. performed the bioinformatic analyses, J.-B. L., D.C.-M., and G.B. performed the experiments, N. C. M.-G., J. A. L., J. A. F. and S. S. provided discussion at the early stage of the study, J.-B. L., B. J. S., S. W. K., J. M. S. and C. J. M. drafted the manuscript, with help from N. C. M.-G., J. A. L. and S. S.

Table 1: Description and isolation source of 104 *Methylobacterium* species distributed in the four main phylogenetic groups. For each group, the number of species, of genomes, and the proportion of genomes isolated from each main category of environment, are given for described and candidate species (numbered from *Methylobacterium sp* 001 to 045), separately. Proportions were corrected by the number of genomes per species. Anthropogenic environments include several other isolations sources.

Group	Description		Species	Genomes	Isolation source						
					Plant (phyllosphere)	Plant (rhizosphere)	Plant (other)	Water, sediments	Soil	Other	Anthropogenic environments*
B	Described species (*M. aminovorans ; M. brachythecii ; M. extorquens M. gnaphalii ; M. haplocladii ;M. organophilum M. podarium ; M. rhodesianum ; M. rhodinum M. salsuginis ; M. suomiense ; M. thiocyanatum M. thuringiense ; M. zatmanii*)		14	33	30%	1%	1%	13%	19%	35%	29%
	Candidate species (*Methylobacterium sp 035 to 041*)		7	8	50%	-	7%	29%	-	14%	29%
	All species		21	41	37%	1%	3%	18%	13%	28%	29%
A	Described species (*M. aerolatum ; M. brachiatum ; M. cerastii M. dankookense ; M. durans ; M. fujisawaense M. gregans ; M. hispanicum ; M. jeotgali M. komagatae ; M. longum ; M. mesophilicum M. oxalidis ; M. persicinum ; M. phyllostachyos M. planium ; M. pseudosasicola ; M. radiodurans M. radiotolerans ; M. segetis ; M. soli M. symbioticum ; M. tardum ; M. trifolii*)		24	45	30%	1%	4%	28%	17%	20%	40%
	Candidate species (*Methylobacterium sp 018 to 034*)		17	36	80%	10%	10%	-	-	-	-
	All species		41	81	51%	4%	7%	16%	10%	12%	23%
D	Described species (*M. adhaesivum ; M. bullatum ; M. goesingense M. gossipiicola ; M. iners ; M. marchantiae*)		6	10	47%	-	17%	20%	-	17%	17%
	Candidate species (*Methylobacterium sp 001 to 017*)		17	32	74%	6%	-	21%	-	-	-
	All species		23	42	67%	4%	4%	20%	-	4%	4%
C	Described species (*M. ajmalii ; M. aquaticum ; M. crusticola ; M. currus M. frigidaeris ; M. indicum ; M. isbiliense M. nodulans ; M. nonmethylotrophicum M. oryzihabitans ; M. platani ; M. tarhaniae M. terrae ; M. terricola ; M. variabile*)		15	21	8%	10%	7%	27%	33%	15%	55%
	Candidate species (*Methylobacterium sp 042 to 045*)		4	4	-	25%	25%	-	25%	25%	25%
	All species		19	25	7%	13%	11%	21%	32%	17%	49%
Methylobacterium		**All described species**	59	109	26%	3%	5%	23%	20%	22%	39%
		All candidate species	45	80	66%	8%	7%	12%	2%	4%	7%
		All species	104	189	43%	5%	6%	18%	12%	14%	25%
***Microvirga* (all species)**			18	24	-	33%	6%	17%	33%	11%	6%
***Enterovirga* (all species)**			2	2	-	-	-	-	50%	50%	-

Table 2: *Methylobacteriaceae* genome characteristics (average and standard deviation per group). GC content was estimated from coding sequences. Hypothetical protein, mobile and repeat elements were excluded from unique annotation counts and estimated copy numbers.

Group	Genomes	Species	Size (Mb)	Annotations	Unique annotations	Estimated copy number	Mobile elements	% GC
A	81	41	6.21 ± 0.59	6907 ± 821	2696 ± 134	1.457 ± 0.067	60 ± 34	70.1 ± 0.8
B	41	21	5.58 ± 0.49	5766 ± 509	2706 ± 173	1.365 ± 0.048	57 ± 31	69.1 ± 0.8
C	25	19	7.15 ± 0.66	7670 ± 956	2899 ± 122	1.542 ± 0.066	71 ± 49	71.1 ± 0.7
D	42	23	4.99 ± 0.35	5224 ± 476	2421 ± 82	1.312 ± 0.042	42 ± 21	68.8 ± 1.1
Enterovirga	2	2	4.91 ± 0.36	5128 ± 182	2321 ± 14	1.414 ± 0.019	14 ± 10	68.8 ± 0.1
Microvirga	22	18	5.92 ± 1.74	6834 ± 2929	2495 ± 251	1.471 ± 0.173	128 ± 147	63.9 ± 1.6

Table 3: Average and standard deviation in gene content dissimilarity (*BC* index, Hellinger transformation on gene occurrence per genome) per and among *Methylobacterium* group and outgroups.

Group	*BC* within species	*BC* among species within groups	*BC* among groups				
			A	***B***	***C***	***D***	***Enterovirga***
A	0.06 ± 0.02	0.19 ± 0.05					
B	0.08 ± 0.02	0.17 ± 0.05	0.28 ± 0.02				
C	0.07 ± 0.03	0.20 ± 0.04	0.29 ± 0.02	0.34 ± 0.02			
D	0.04 ± 0.02	0.16 ± 0.03	0.24 ± 0.03	0.24 ± 0.03	0.32 ± 0.02		
Enterovirga	-	0.28	0.39 ± 0.01	0.41 ± 0.02	0.36 ± 0.02	0.37 ± 0.01	
Microvirga	0.03 ± 0.04	0.26 ± 0.04	0.41 ± 0.02	0.44 ± 0.02	0.38 ± 0.02	0.40 ± 0.02	0.36 ± 0.03

Table 4: Average and standard deviation in core genome synteny (*SI*) per and among *Methylobacterium* group and outgroups.

Group	*SI* within species	*SI* among species within groups	*SI* among groups				
			A	***B***	***C***	***D***	***Enterovirga***
A	0.98 ± 0.03	0.70 ± 0.17					
B	0.99 ± 0.01	0.77 ± 0.21	0.46 ± 0.03				
C	0.91 ± 0.06	0.61 ± 0.12	0.43 ± 0.03	0.44 ± 0.01			
D	0.98 ± 0.02	0.73 ± 0.11	0.53 ± 0.05	0.51 ± 0.02	0.48 ± 0.01		
Enterovirga	-	0.47	0.39 ± 0.03	0.39 ± 0.02	0.38 ± 0.02	0.42 ± 0.03	
Microvirga	1.00 ± 0.00	0.91 ± 0.05	0.49 ± 0.03	0.48 ± 0.01	0.46 ± 0.01	0.52 ± 0.02	0.52 ± 0.05

889

890

Figure Legends

Figure 1: *Methylobacteriaceae* lineage trees inferred from 213 genomes. a) Best tree from RAxML ML search on the concatenated alignments of 384 core gene nucleotide sequences (GTRCAT model, 512 replicated trees), rooted on *Microvirga* and *Enterovirga* (grey). Colors indicate the four major *Methylobacterium* groups: A (red), B (purple), C (green) and D (blue). Correspondence with clades described by previous studies is indicated. b) ASTRAL tree inferred from 384 core gene ML trees. Each gene ML tree was inferred assuming a GTRgamma model (1,000 replicated trees; nodes with less than 10% of support collapsed) and combined in ASTRAL-III. Branch lengths are in coalescent units. Nodal support values represent local posterior probability. c) SVD quartet tree inferred from the concatenated alignments of 384 core gene nucleotide sequences. Nodes supported by less than 75% of quartets were collapsed. d) Main isolation sources of species from *Methylobacterium* group and *Microvirga* (see Table 1). For each group, ordered according to a consensus tree (see panels b and c), the number of species is indicated in parenthesis.

Figure 2: Gene content comparison among the four main *Methylobacterium* groups. a) Occurrence of 10,187 gene annotations (rows) in 124 *Methylobacteriaceae* species (average occurrence per species; column, ordered according to the ASTRAL species tree, left) and in four *Methylobacterium* groups and two outgroups (mean occurrence among species within groups; legend in bottom right) are shown. b-c) Venn diagrams showing the overlap of pan genomes (b) and core genome (c) among four groups. Pan and core genome sizes were estimated assuming 15 species per group (mean and standard deviation over 100 random resampling of 15 species per group). d) RAxML ML best tree based on annotation occurrence per genome (best ML tree, BINCAT model, 1,001 replicate trees). Main groups are shown and are monophyletic in the gene content tree, but group A: clade A2 (Alessa et al. 2021) and *M. jeotgali* branched out of group A.

Figure 3: Core genome architecture comparison (synteny) among *Methylobacteriaceae* genomes. a) Consensus map of the *Methylobacterium* core genome architecture, and major rearrangements within and among *Methylobacterium* groups, using *M. planium YIM132548*

core genome as a reference. The map was drawn as a network using 384 core genes as nodes, and links among neighbor core genes as edges. Only 389 links that were observed in a majority (>50%) of species from a given *Methylobacterium* group are shown (Venn diagram on top right; 5,720 links discarded). Bold lines indicate links mostly conserved in group A, colored according to their dominance in other groups (legend on bottom right). Thick lines indicate links mostly absent in group A but dominant in other groups. A syntenic island conserved in most *Methylobacterium* genomes and containing ribosomal genes and gene *rpoB* is indicated (dotted frame). b) RAxML ML best tree based on link occurrence per genome (6,109 links; best ML tree, BINCAT model, 1,001 replicate trees). Main groups are shown and are monophyletic in the synteny tree. c) Detailed synteny plot for the comparison of core genome architecture between seven species from group A and six species from group D (best assembled genome per species). For each pairwise comparison, core gene (black points) are ordered according to their relative position in species 1 genome (x-axis) and are compared with their relative positions in species 2 genome (y-axis). Each plot is colored according to the *SI* value between species 1 and 2 (scale on top right).

References

Alessa O et al. 2021. Comprehensive Comparative Genomics and Phenotyping of *Methylobacterium* Species. Frontiers in Microbiology. 12:2852. doi: 10.3389/fmicb.2021.740610.

Anthony C. 1991. Assimilation of carbon by methylotrophs. Biotechnology. 18:79–109. doi: 10.1016/b978-0-7506-9188-8.50011-5.

Aziz RK et al. 2008. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. 9:75. doi: 10.1186/1471-2164-9-75.

Belkhelfa S et al. 2018. Complete Genome Sequence of the Facultative Methylotroph *Methylobacterium extorquens* TK 0001 Isolated from Soil in Poland. Genome Announc. 6. doi: 10.1128/genomeA.00018-18.

Bijlani S et al. 2021. *Methylobacterium ajmalii* sp. nov., Isolated From the International Space Station. Frontiers in Microbiology. 12:534. doi: 10.3389/fmicb.2021.639396.

Bray JR, Curtis JT. 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecological Monographs. 27:325–349. doi: 10.2307/1942268.

Castillo-Ramírez S, González V. 2008. Factors affecting the concordance between orthologous gene trees and species tree in bacteria. BMC Evol Biol. 8:300. doi: 10.1186/1471-2148-8-300.

Chen W-M, Cai C-Y, Li Z-H, Young C-C, Sheu S-Y. 2019. *Methylobacterium oryzihabitans* sp. nov., isolated from water sampled from a rice paddy field. International Journal of Systematic and Evolutionary Microbiology,. 69:3843–3850. doi: 10.1099/ijsem.0.003693.

Chen X et al. 2017. *Enterovirga rhinocerotis* gen. nov., sp. nov., isolated from *Rhinoceros unicornis* faeces. Antonie Van Leeuwenhoek. 110:553–562. doi: 10.1007/s10482-016-0823-1.

Chifman J, Kubatko L. 2014. Quartet Inference from SNP Data Under the Coalescent Model. Bioinformatics. 30:3317–3324. doi: 10.1093/bioinformatics/btu530.

Chun J, Rainey FA. 2014. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. International Journal of Systematic and Evolutionary Microbiology. 64:316–324. doi: 10.1099/ijs.0.054171-0.

Chung M, Munro JB, Tettelin H, Hotopp JCD. 2018. Using Core Genome Alignments To

970 Assign Bacterial Species. mSystems. doi: 10.1128/mSystems.00236-18.
971 Clarke PH. 1983. The biochemistry of Methylotrophs by C. Anthony Academic Press;
972 London, New York, 1982 xvi + 432 pages. £24.00, $49.50. FEBS Letters. 160:303–
973 303. doi: 10.1016/0014-5793(83)80989-2.
974 Coenye T, Gevers D, de Peer YV, Vandamme P, Swings J. 2005. Towards a prokaryotic
975 genomic taxonomy⋆. FEMS Microbiology Reviews. 29:147–167. doi:
976 10.1016/j.fmrre.2004.11.004.
977 Corpe WA, Rheem S. 1989. Ecology of the methylotrophic bacteria on living leaf surfaces.
978 FEMS Microbiol Ecol. 5:243–249. doi: 10.1111/j.1574-6968.1989.tb03698.x.
979 Creevey CJ, Doerks T, Fitzpatrick DA, Raes J, Bork P. 2011. Universally Distributed Single-
980 Copy Genes Indicate a Constant Rate of Horizontal Transfer. PLOS ONE. 6:e22099.
981 doi: 10.1371/journal.pone.0022099.
982 Davidson R, Vachaspati P, Mirarab S, Warnow T. 2015. Phylogenomic species tree
983 estimation in the presence of incomplete lineage sorting and horizontal gene transfer.
984 BMC Genomics. 16:S1. doi: 10.1186/1471-2164-16-S10-S1.
985 Dourado MN, Aparecida Camargo Neves A, Santos DS, Araújo WL. 2015. Biotechnological
986 and Agronomic Potential of Endophytic Pink-Pigmented Methylotrophic
987 Methylobacterium spp. Biomed Res Int. 2015. doi: 10.1155/2015/909016.
988 Feng G-D et al. 2020. *Methylobacterium nonmethylotrophicum* sp. nov., isolated from
989 tungsten mine tailing. International Journal of Systematic and Evolutionary
990 Microbiology,. 70:2867–2872. doi: 10.1099/ijsem.0.004112.
991 Gillings MR, Paulsen IT. 2014. Microbiology of the Anthropocene. Anthropocene. 5:1–8.
992 doi: 10.1016/j.ancene.2014.06.004.
993 Green PN, Ardley JK. 2018. Review of the genus *Methylobacterium* and closely related
994 organisms: a proposal that some *Methylobacterium* species be reclassified into a new
995 genus, *Methylorubrum* gen. nov. International Journal of Systematic and Evolutionary
996 Microbiology. 68:2727–2748. doi: 10.1099/ijsem.0.002856.
997 Grossi CEM et al. 2020. *Methylobacterium sp.* 2A Is a Plant Growth-Promoting
998 Rhizobacteria That Has the Potential to Improve Potato Crop Yield Under Adverse
999 Conditions. Front. Plant Sci. 11. doi: 10.3389/fpls.2020.00071.
1000 Helfrich EJN et al. 2018. Bipartite interactions, antibiotic production and biosynthetic

potential of the Arabidopsis leaf microbiome. Nat Microbiol. 3:909–919. doi: 10.1038/s41564-018-0200-0.

Hördt A et al. 2020. Analysis of 1,000+ Type-Strain Genomes Substantially Improves Taxonomic Classification of *Alphaproteobacteria*. Frontiers in Microbiology. 11:468. doi: 10.3389/fmicb.2020.00468.

Hugenholtz P. 2002. Exploring prokaryotic diversity in the genomic era. Genome Biol. 3:reviews0003.1. doi: 10.1186/gb-2002-3-2-reviews0003.

Ivanova EG, Doronina NV, Trotsenko YuA. 2001. Aerobic Methylobacteria Are Capable of Synthesizing Auxins. Microbiology. 70:392–397. doi: 10.1023/A:1010469708107.

Jia L juan et al. 2020. *Methylobacterium crusticola* sp. nov., isolated from biological soil crusts. International Journal of Systematic and Evolutionary Microbiology,. 70:2089–2095. doi: 10.1099/ijsem.0.004020.

Jiang L et al. 2020. *Methylobacterium planium* sp. nov., isolated from a lichen sample. Arch Microbiol. 202:1709–1715. doi: 10.1007/s00203-020-01881-4.

Jourand P et al. 2004. *Methylobacterium nodulans* sp. nov., for a group of aerobic, facultatively methylotrophic, legume root-nodule-forming and nitrogen-fixing bacteria. International Journal of Systematic and Evolutionary Microbiology,. 54:2269–2273. doi: 10.1099/ijs.0.02902-0.

Kanso S, Patel BKC. 2003. *Microvirga subterranea* gen. nov., sp. nov., a moderate thermophile from a deep subsurface Australian thermal aquifer. Int J Syst Evol Microbiol. 53:401–406. doi: 10.1099/ijs.0.02348-0.

Keppler F, Hamilton JTG, Braß M, Röckmann T. 2006. Methane emissions from terrestrial plants under aerobic conditions. Nature. 439:187–191. doi: 10.1038/nature04420.

Kim J, Chhetri G, Kim I, Lee B, et al. 2020. *Methylobacterium terricola* sp. nov., a gamma radiation-resistant bacterium isolated from gamma ray-irradiated soil. International Journal of Systematic and Evolutionary Microbiology,. 70:2449–2456. doi: 10.1099/ijsem.0.004054.

Kim J, Chhetri G, Kim I, Kim MK, Seo T. 2020. *Methylobacterium durans* sp. nov., a radiation-resistant bacterium isolated from gamma ray-irradiated soil. Antonie van Leeuwenhoek. 113:211–220. doi: 10.1007/s10482-019-01331-2.

Koonin EV, Makarova KS, Wolf YI. 2021. Evolution of Microbial Genomics: Conceptual

1032 Shifts over a Quarter Century. Trends in Microbiology. 29:582–592. doi:
1033 10.1016/j.tim.2021.01.005.
1034 Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. PartitionFinder: Combined Selection of
1035 Partitioning Schemes and Substitution Models for Phylogenetic Analyses. Molecular
1036 Biology and Evolution. 29:1695–1701. doi: 10.1093/molbev/mss020.
1037 Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. 2014. Selecting optimal partitioning
1038 schemes for phylogenomic datasets. BMC Evolutionary Biology. 14:82. doi:
1039 10.1186/1471-2148-14-82.
1040 Leducq J-B et al. 2022. Fine-Scale Adaptations to Environmental Variation and Growth
1041 Strategies Drive Phyllosphere Methylobacterium Diversity. mBio. doi:
1042 10.1128/mbio.03175-21.
1043 Lee HS et al. 2006. Physiological enhancement of early growth of rice seedlings (Oryza
1044 sativa L.) by production of phytohormone of N2-fixing methylotrophic isolates. Biol
1045 Fertil Soils. 42:402–408. doi: 10.1007/s00374-006-0083-8.
1046 Lee JA, Stolyar S, Marx CJ. 2022. Aerobic Methoxydotrophy: Growth on Methoxylated
1047 Aromatic Compounds by *Methylobacteriaceae*. Frontiers in Microbiology. 13.
1048 https://www.frontiersin.org/article/10.3389/fmicb.2022.849573 (Accessed June 1,
1049 2022).
1050 Levy A et al. 2018. Genomic features of bacterial adaptation to plants. Nat Genet. 50:138–
1051 150. doi: 10.1038/s41588-017-0012-9.
1052 Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node
1053 solution for large and complex metagenomics assembly via succinct de Bruijn graph.
1054 Bioinformatics. 31:1674–1676. doi: 10.1093/bioinformatics/btv033.
1055 Louca S et al. 2018. Bacterial diversification through geological time. Nat Ecol Evol.
1056 2:1458–1467. doi: 10.1038/s41559-018-0625-0.
1057 Madhaiyan M et al. 2005. Pink-pigmented facultative methylotrophic bacteria accelerate
1058 germination, growth and yield of sugarcane clone Co86032 (Saccharum officinarum
1059 L.). Biol Fertil Soils. 41:350–358. doi: 10.1007/s00374-005-0838-7.
1060 Madhaiyan M, Poonguzhali S, Sa T. 2007. Metal tolerating methylotrophic bacteria reduces
1061 nickel and cadmium toxicity and promotes plant growth of tomato (*Lycopersicon*
1062 *esculentum* L.). Chemosphere. 69:220–228. doi: 10.1016/j.chemosphere.2007.04.017.

1063 Marx CJ et al. 2012. Complete Genome Sequences of Six Strains of the Genus
1064 *Methylobacterium*. J Bacteriol. 194:4746–4748. doi: 10.1128/JB.01009-12.
1065 Mende DR, Sunagawa S, Zeller G, Bork P. 2013. Accurate and universal delineation of
1066 prokaryotic species. Nat Methods. 10:881–884. doi: 10.1038/nmeth.2575.
1067 Minami T et al. 2016. Complete Genome Sequence of *Methylobacterium* sp. Strain AMS5,
1068 an Isolate from a Soybean Stem. Genome Announc. 4. doi: 10.1128/genomeA.00144-
1069 16.
1070 Minh BQ et al. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic
1071 Inference in the Genomic Era. Molecular Biology and Evolution. 37:1530–1534. doi:
1072 10.1093/molbev/msaa015.
1073 Mirarab S et al. 2014. ASTRAL: genome-scale coalescent-based species tree estimation.
1074 Bioinformatics. 30:i541–i548. doi: 10.1093/bioinformatics/btu462.
1075 Morohoshi T, Ikeda T. 2016. Complete Genome Sequence of *Methylobacterium popul*i P-
1076 1M, Isolated from Pink-Pigmented Household Biofilm. Genome Announc. 4. doi:
1077 10.1128/genomeA.00458-16.
1078 Nishida H. 2012. Evolution of genome base composition and genome size in bacteria.
1079 Frontiers in Microbiology. 3.
1080 https://www.frontiersin.org/article/10.3389/fmicb.2012.00420 (Accessed January 17,
1081 2022).
1082 Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and
1083 evolutionary analyses in R. Bioinformatics. 35:526–528. doi:
1084 10.1093/bioinformatics/bty633.
1085 Park S-C, Lee K, Kim YO, Won S, Chun J. 2019. Large-Scale Genomics Reveals the
1086 Genetic Characteristics of Seven Species and Importance of Phylogenetic Distance for
1087 Estimating Pan-Genome Size. Frontiers in Microbiology. 10.
1088 https://www.frontiersin.org/article/10.3389/fmicb.2019.00834 (Accessed January 27,
1089 2022).
1090 Pascual JA et al. 2020. *Methylobacterium symbioticum* sp. nov., a new species isolated from
1091 spores of *Glomus iranicum* var. *tenuihypharum*. Curr Microbiol. 77:2031–2041. doi:
1092 10.1007/s00284-020-02101-4.
1093 Qu Z et al. 2014. *Psychroglaciecola arctica* gen. nov., sp. nov., isolated from Arctic glacial

foreland soil. International Journal of Systematic and Evolutionary Microbiology,. 64:1817–1824. doi: 10.1099/ijs.0.060913-0.

R-Developement-Core-Team. 2011. *R: A language and environment for statistical computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013. url: http://www. R-project. org.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Mathematical Biosciences. 53:131–147. doi: 10.1016/0025-5564(81)90043-2.

Rosselló-Mora R, Amann R. 2001. The species concept for prokaryotes. FEMS Microbiology Reviews. 25:39–67. doi: 10.1111/j.1574-6976.2001.tb00571.x.

Ryu JH (Chungbuk NU et al. 2006. Plant Growth Substances Produced by *Methylobacterium spp*. and Their Effect on Tomato (*Lycopersicon esculentum* L.) and Red Pepper (*Capsicum annuum* L.) Growth. Journal of Microbiology and Biotechnology. https://agris.fao.org/agris-search/search.do?recordID=KR2007002803 (Accessed September 12, 2020).

Sakoparnig T, Field C, van Nimwegen E. 2021. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species Nourmohammad, A & Walczak, AM, editors. eLife. 10:e65366. doi: 10.7554/eLife.65366.

Sayyari E, Mirarab S. 2016. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. Molecular Biology and Evolution. 33:1654–1668. doi: 10.1093/molbev/msw079.

Shannon P et al. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 13:2498–2504. doi: 10.1101/gr.1239303.

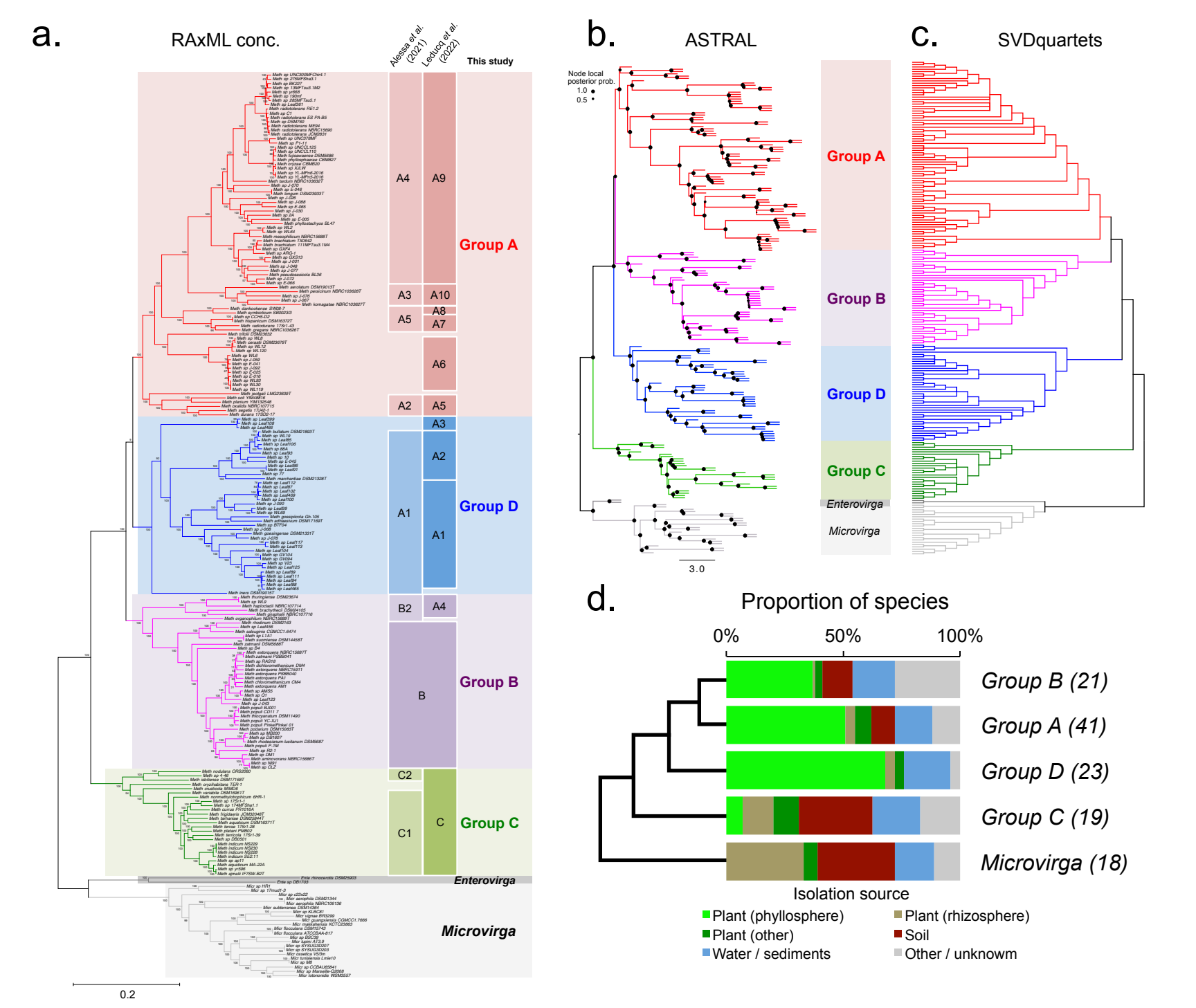
Shapiro BJ, Leducq J-B, Mallet J. 2016. What is speciation? PLoS Genet. 12:e1005860.

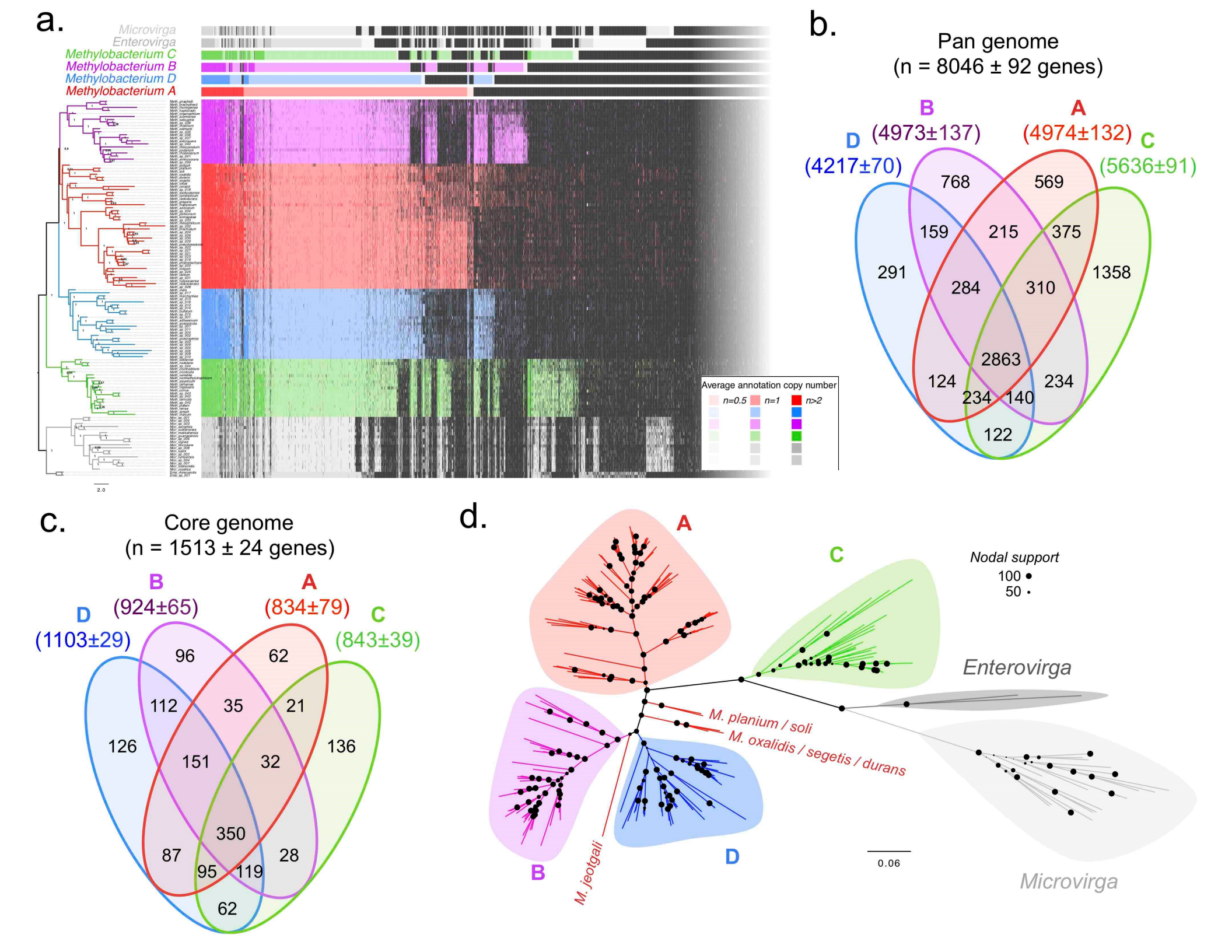
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313. doi: 10.1093/bioinformatics/btu033.

Sy A et al. 2001. Methylotrophic *Methylobacterium* Bacteria Nodulate and Fix Nitrogen in Symbiosis with Legumes. J Bacteriol. 183:214–220. doi: 10.1128/JB.183.1.214-220.2001.

Tani A, Ogura Y, Hayashi T, Kimbara K. 2015. Complete Genome Sequence of *Methylobacterium aquaticum* Strain 22A, Isolated from *Racomitrium japonicum* Moss.

1125 Genome Announc. 3. doi: 10.1128/genomeA.00266-15.
1126 Ten LN et al. 2020. *Methylobacterium segetis* sp. nov., a novel member of the family
1127 *Methylobacteriaceae* isolated from soil on Jeju Island. Arch Microbiol. 202:747–754.
1128 doi: 10.1007/s00203-019-01784-z.
1129 Wang S, Meade A, Lam H-M, Luo H. 2020. Evolutionary Timeline and Genomic Plasticity
1130 Underlying the Lifestyle Diversity in *Rhizobiales* Allard, SM, editor. mSystems.
1131 5:e00438-20, /msystems/5/4/msys.00438-20.atom. doi: 10.1128/mSystems.00438-20.
1132 Wilgenbusch JC, Swofford D. 2003. Inferring Evolutionary Trees with PAUP*. Current
1133 Protocols in Bioinformatics. 00:6.4.1-6.4.28. doi: 10.1002/0471250953.bi0604s00.
1134 Zervas A, Zeng Y, Madsen AM, Hansen LH. 2019. Genomics of Aerobic Photoheterotrophs
1135 in Wheat Phyllosphere Reveals Divergent Evolutionary Patterns of Photosynthetic
1136 Genes in *Methylobacterium spp*. Genome Biology and Evolution. 11:2895–2908. doi:
1137 10.1093/gbe/evz204.
1138 Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree
1139 reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:153. doi:
1140 10.1186/s12859-018-2129-y.
1141


a.
RAxML conc.
Alessa et al. (2021)
Leducq et al. (2022)
This study
A4
A9
Group A
A3
A10
A8
A5
A7
A6
A2
A5
A3
A2
A1
A1
Group D
B2
A4
B
Group B
C2
C1
C
Group C
Enterovirga
Microvirga
0.2
b.
ASTRAL
Node local posterior prob.
1.0
0.5
Group A
Group B
Group D
Group C
Enterovirga
Microvirga
3.0
c.
SVDquartets
d.
Proportion of species
0%
50%
100%
Group B (21)
Group A (41)
Group D (23)
Group C (19)
Microvirga (18)
Isolation source
Plant (phyllosphere)
Plant (rhizosphere)
Plant (other)
Soil
Water / sediments
Other / unknown


a.
Microvirga
Enterovirga
Methylobacterium C
Methylobacterium B
Methylobacterium D
Methylobacterium A
Average annotation copy number
n=0.5
n=1
n>2
2.0
b.
Pan genome
(n = 8046 ± 92 genes)
B
(4973±137)
A
(4974±132)
D
(4217±70)
C
(5636±91)
768
569
159
215
375
291
284
310
1358
2863
124
234
234
140
122
c.
Core genome
(n = 1513 ± 24 genes)
B
(924±65)
A
(834±79)
D
(1103±29)
C
(843±39)
96
62
112
35
21
126
151
32
136
350
87
28
95
119
62
d.
A
C
Nodal support
100
50
Enterovirga
M. planium / soli
M. oxalidis / segetis / durans
B
M. jeotgali
D
0.06
Microvirga

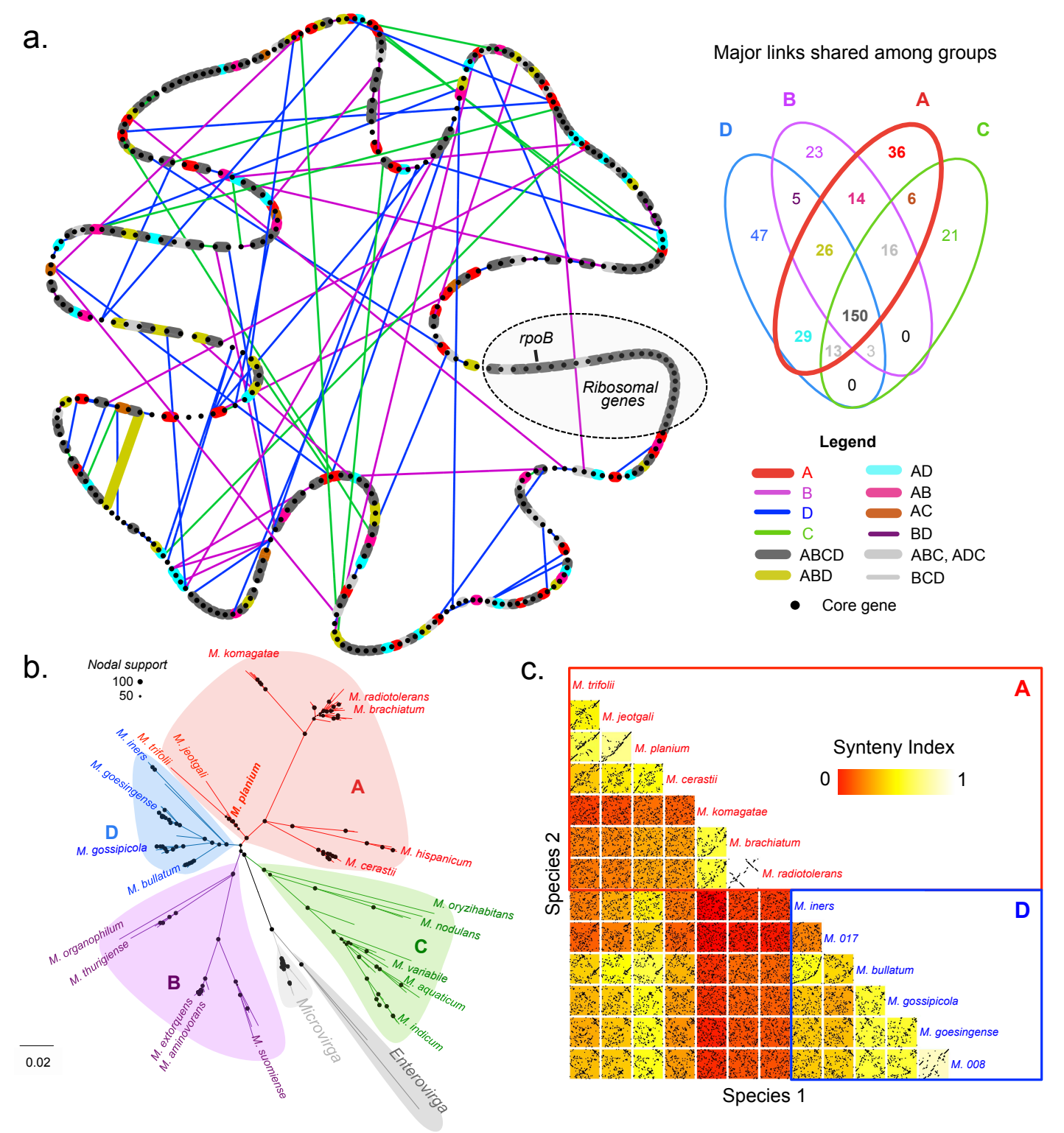

a.
rpoB
Ribosomal genes
Major links shared among groups
B
A
D
C
23
36
5
14
6
47
21
26
16
150
29
0
13
3
0
Legend
A
B
D
C
ABCD
ABD
AD
AB
AC
BD
ABC, ADC
BCD
Core gene
b.
Nodal support
100
50
M. komagatae
M. radiotolerans
M. brachiatum
M. iners
M. trifolii
M. jeotgali
M. planium
M. goesingense
D
M. gossipicola
M. bullatum
A
M. hispanicum
M. cerastii
M. oryzihabitans
M. nodulans
C
M. variabile
M. aquaticum
M. indicum
M. organophilum
M. thurigiense
B
M. extorquens
M. aminovorans
M. suomiense
Microviriga
Enteroviriga
0.02
c.
M. trifolii
M. jeotgali
M. planium
M. cerastii
M. komagatae
M. brachiatum
M. radiotolerans
A
Synteny Index
0
1
M. iners
M. 017
M. bullatum
M. gossipicola
M. goesingense
M. 008
D
Species 2
Species 1

Figure S1 : Proportion of genes per genome present in 1 (blue), 2 (green), 3 (orange), 4 (red) or 5 (purple) copies per genome in function of genome assembly quality (given by the number of scaffolds per genome, log scale).

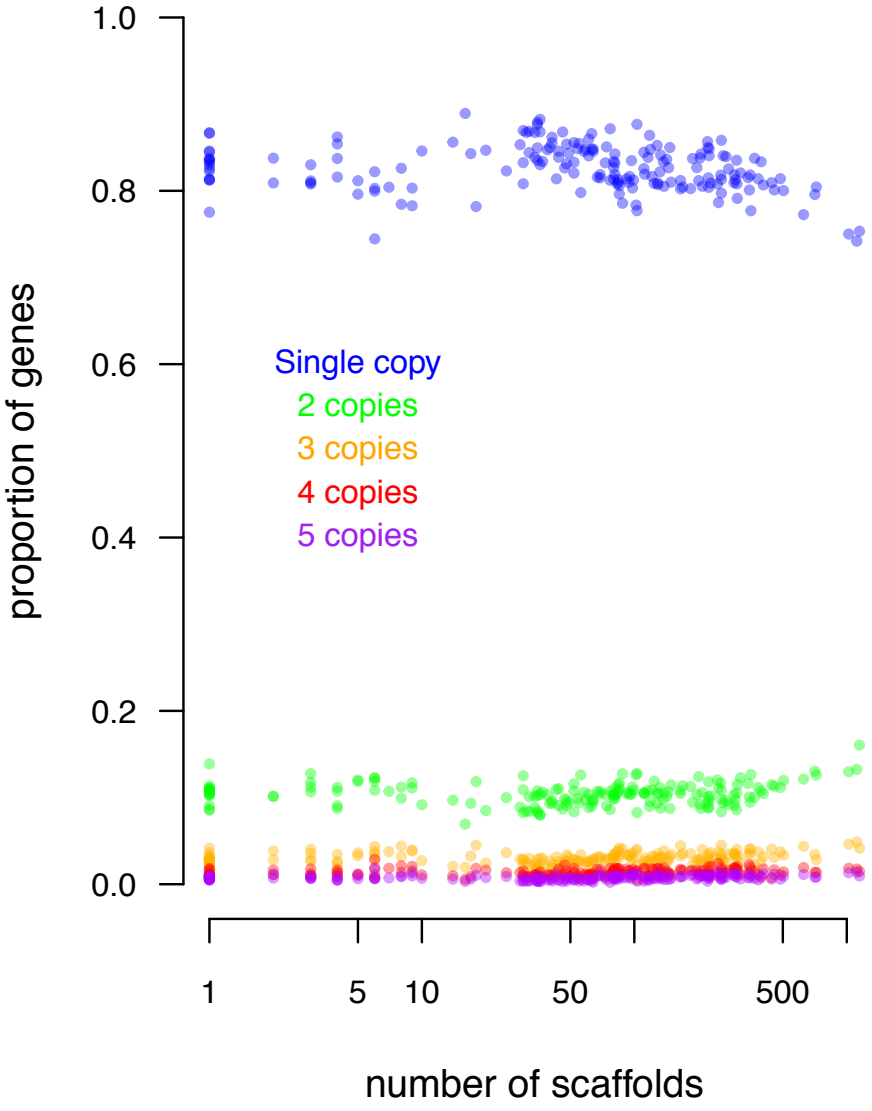


Figure S2: Identification of true core genes among 893 candidate core genes present in a single copy in at least 90% of 184 *Methylobacteriaceae* genomes. Average gene size normalized (divided) by the average nucleotide sequence size observed in complete genomes (defined as genomes with *N50* > 3×10^6 Mb) was plotted against the number of copies observed per genome. Each dot represents one copy in one genome. Lines represent the expected copy number for each normalized size/observed copy number combination. 398 genes for which at least one genome had more than one copy with normalized size >0.75 were considered as true duplicates and removed from the analysis (red). For the 495 remaining candidate core genes, single-copy genes with normalized size >1.3 and gene copies with normalized size <0.7 (regardless copy number) were considered as missing data (blue). Of the remaining genes, 384 genes with a single copy in at least 180 genomes were considered as true core genes.

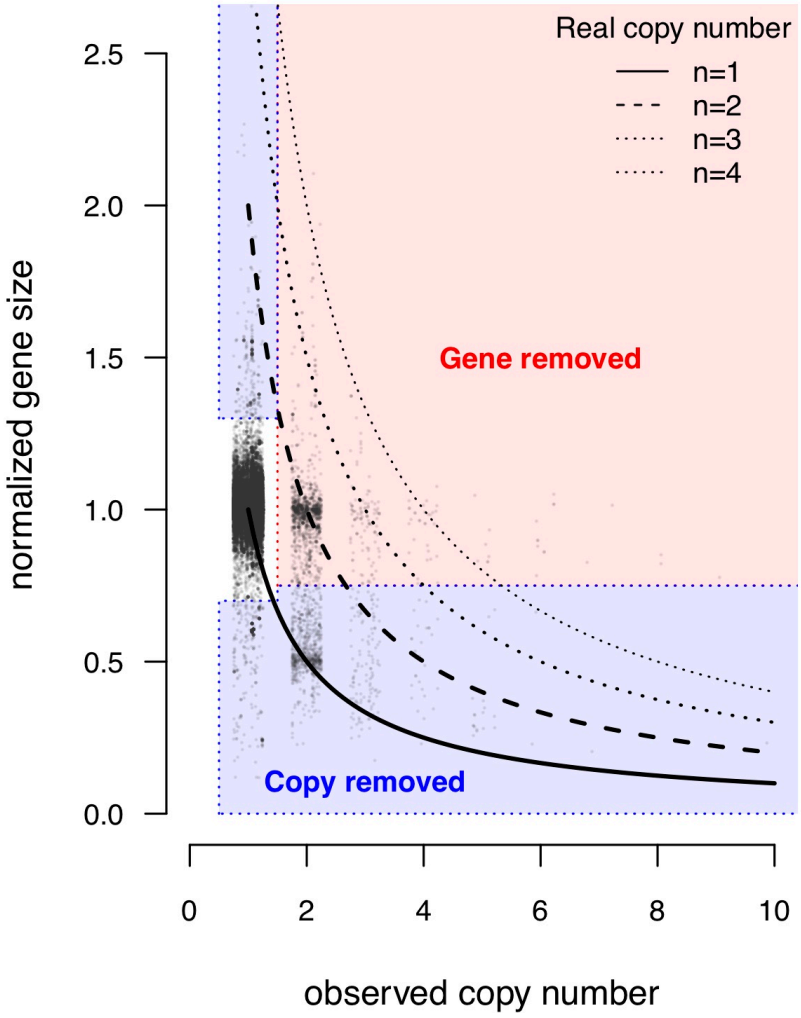


Figure S3: Detailed lineage trees of *Methylobacteriaceae* reconstructed from 213 genomes using ASTRAL and SVDquartet. a) ASTRAL tree inferred from 384 core gene ML trees. Each gene ML tree was inferred in RAxML assuming a GTRgamma model (1,000 replicated trees; nodes with less than 10% of support collapsed) and combined in ASTRAL-III. Branch lengths are in coalescent units. Nodal support values represent local posterior probability. b) SVD quartet tree inferred from the concatenated alignments of 384 core gene nucleotide sequences. Nodal support values indicate the proportion of quartets supporting each node. Trees were rooted on *Microvirga* and *Enterovirga*. Branches and genomes names (Genus, species, strain) were colored according to assignation to *Methylobacterium* groups (A: red; B: purple; C: green; D: blue) and outgroups (*Microvirga*: grey; *Enterovirga*: dark grey)

Figure S4: Normalized RF distance distribution between the RAxML majority consensus rule lineage tree and the 512 replicate trees (grey; see Figure 1a) and normalized RF distances between lineage trees (points; legend on top right).

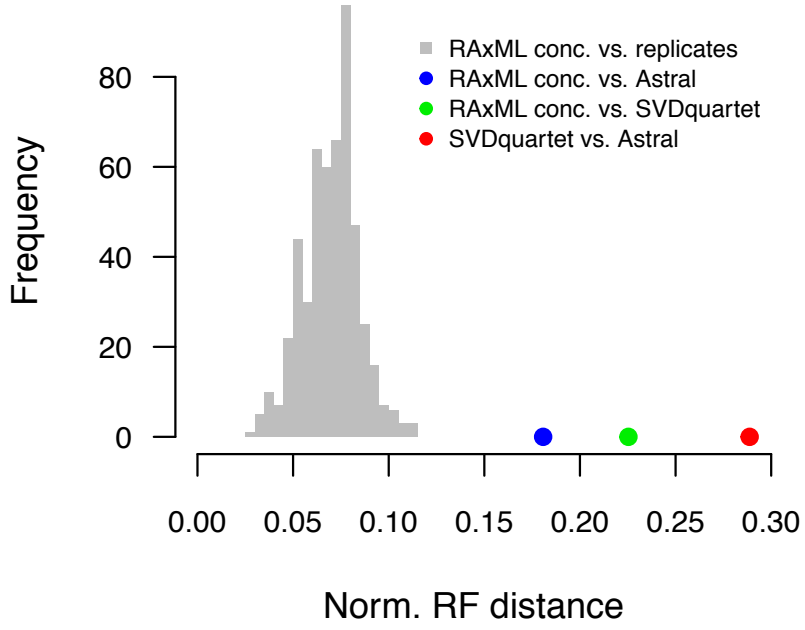


Figure S5: Detailed species trees of *Methylobacteriaceae* inferred from 213 genomes assuming **124 *Methylobacterium* species**. a) ASTRAL tree inferred from 384 core gene ML trees. Each gene ML tree was inferred in RAxML assuming a GTRgamma model (1,000 replicated trees; nodes with less than 10% of support collapsed) and combined in ASTRAL-III. Branch lengths are in coalescent units. Nodal support values represent local posterior probability. b) SVD quartet tree inferred from the concatenated alignments of 384 core gene nucleotide sequences. Nodal support values indicate the proportion of quartets supporting each node. Trees were rooted on *Microvirga* and *Enterovirga*. Branches were colored according to assignation to *Methylobacterium* groups (A: red; B: purple; C: green; D: blue) and outgroups (*Microvirga*: grey; *Enterovirga*: dark grey)

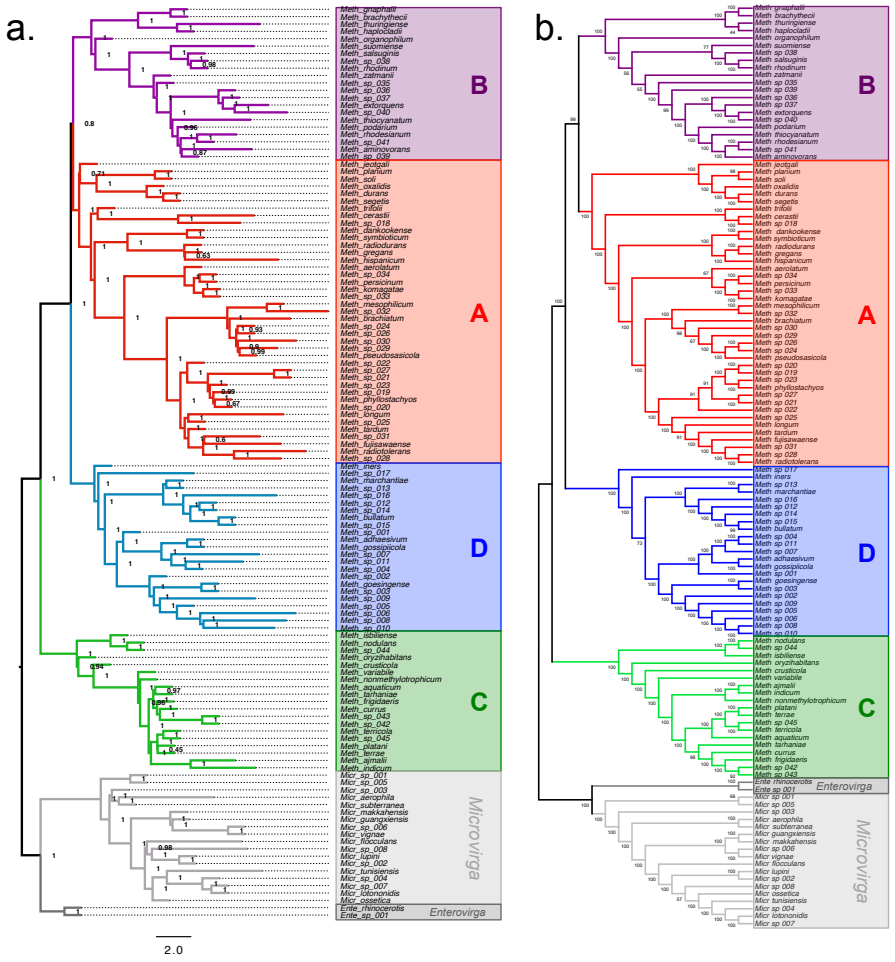


Figure S6: Comparison of genome characteristics among 4 *Methylobacterium* groups. Genome size, number of unique gene annotations per genome (unknown proteins, repeat elements and mobile elements excluded), average gene copy number per genome, and GC content (in coding genome) are compared to each other. Each point represents values for a genome, colored according to assignment to *Methylobacterium* groups (A: red; B: purple; C: green; D: blue). Ellipses indicate 50% of standard deviation centered on average values for each group.

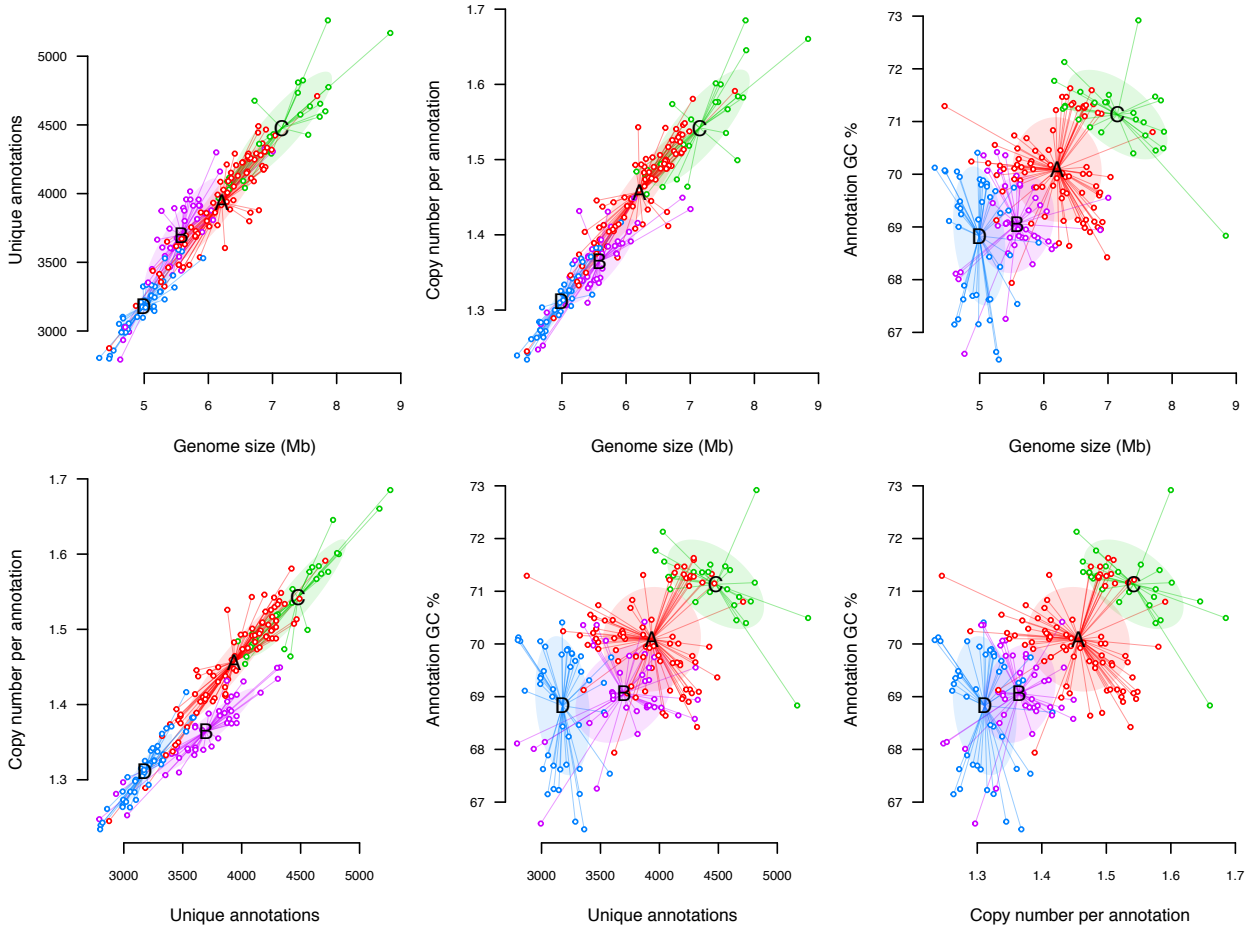


Figure S7: Pan (a) and core (b) genome size estimations in four *Methylobacterium* groups and *Microvirga*. Genome sizes per group (number of genes per groups; y-axis) were estimated for every number of species assumed in the range 1-n (n = maximum number of species per group; x-axis). For each group and number of species, average (lines) and standard deviation (frames) over 100 random resampling of n species per group were estimated. Dotted lines indicate the value for which pan genome and core genome size where estimated (n=15 species per group; Figure 2).

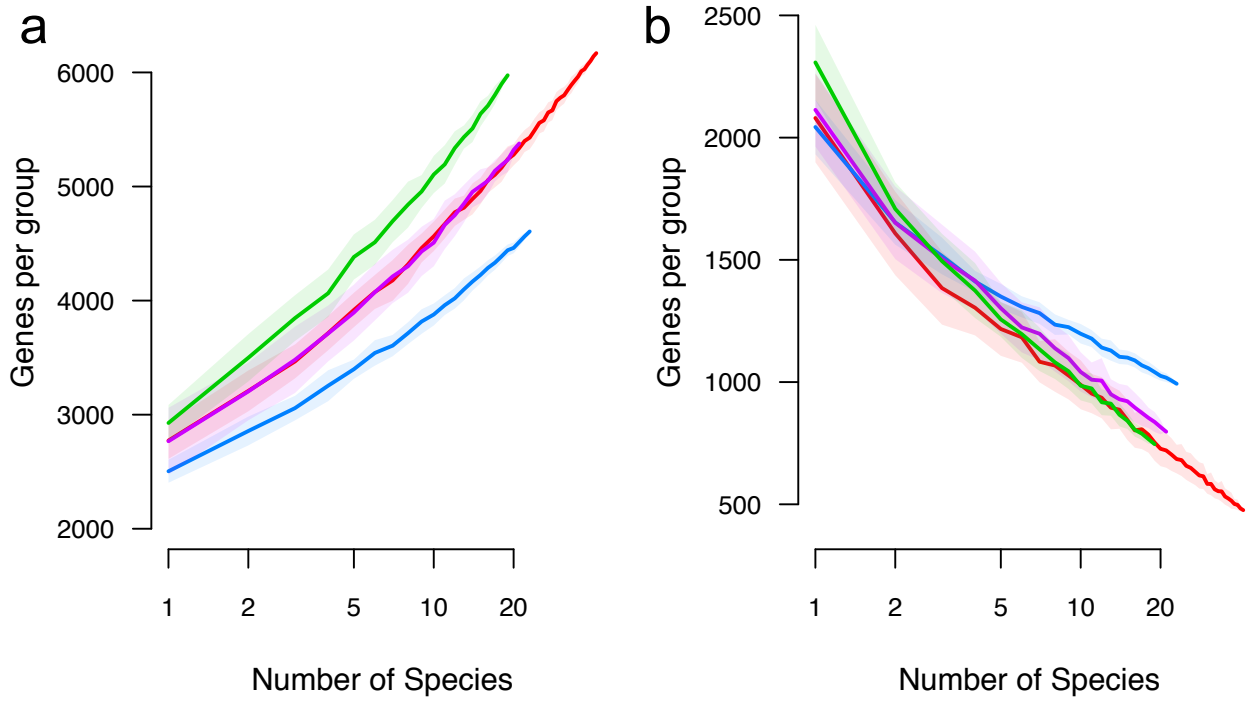


Figure S8: a) ASTRAL species tree. b) Heatmap of dissimilarity in gene content (*BC* index; blue scale; above diagonal) and similarity in core genome architecture (Synteny index; orange scale; below diagonal) among 104 *Methylobacterium* and 20 outgroups species, matching the species tree.

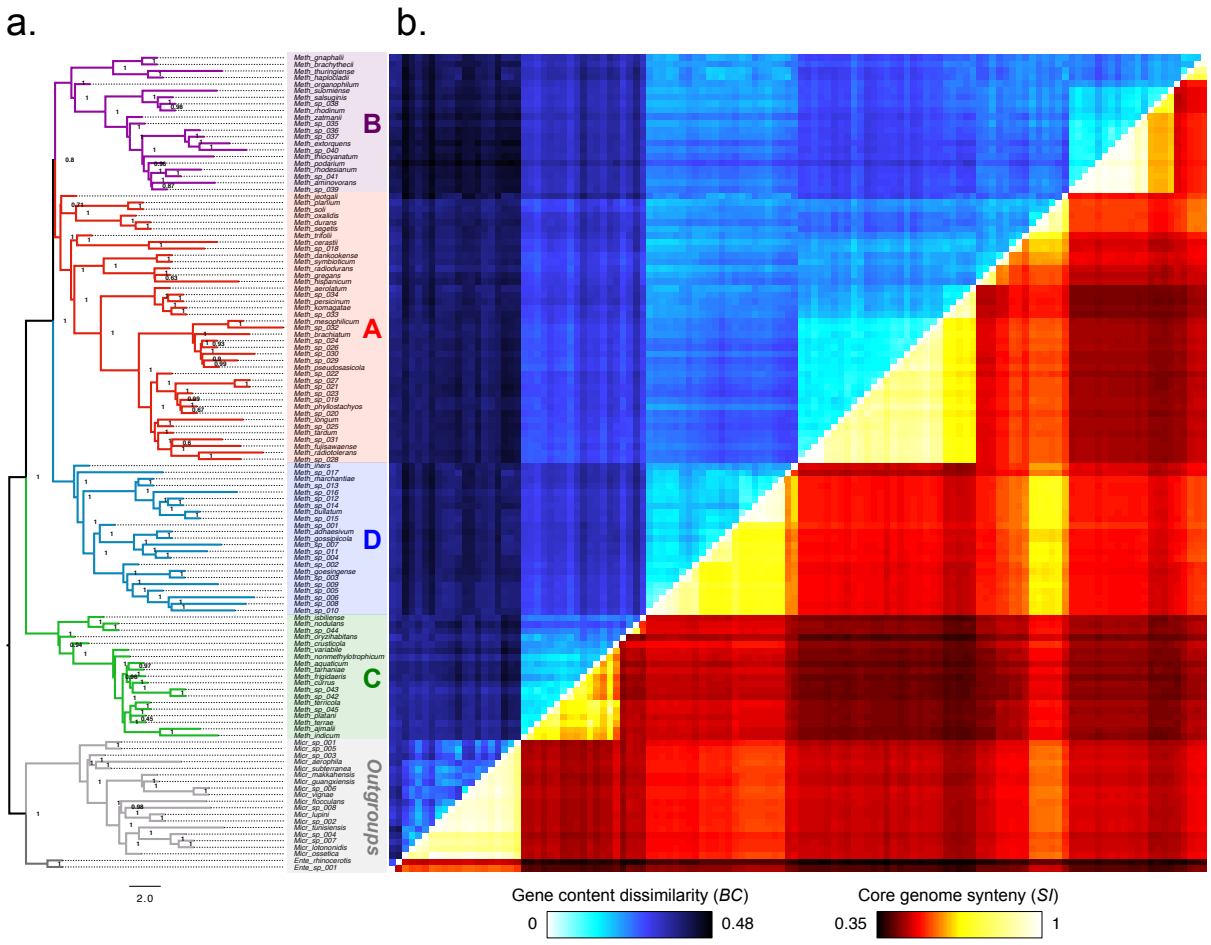


Figure S9: Detailed *Methylobacteriaceae* lineage trees inferred from gene content (a) and core genome synteny (b). Each ML tree was inferred in RAxML assuming a BINCAT model (1,000 replicated trees). Nodal support values indicate the proportion of replicate tree supporting each node. Trees were rooted on *Microvirga* and *Enterovirga*. Branches were colored according to assignation to *Methylobacterium* groups (A: red; B: purple; C: green; D: blue) and outgroups (*Microvirga*: grey; *Enterovirga*: dark grey)

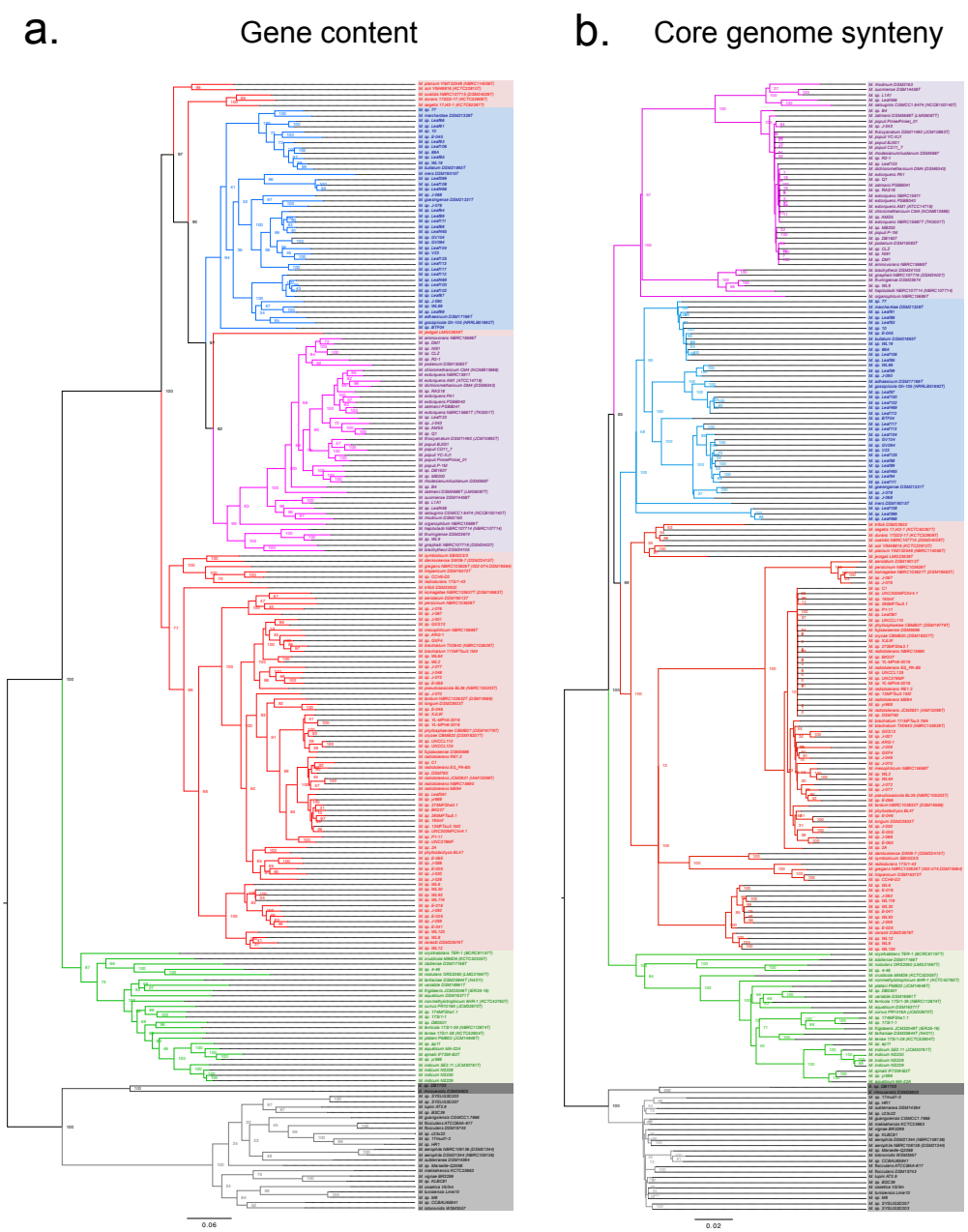


Figure S10: Comparison of core genome architecture among 124 *Methylobacteriaceae* species (rows; ordered according to the ASTRAL species tree, left) using 384 links (pairs of contiguous core genes) observed in the *M. planium* genome as a reference (links are ordered according to the reference genome). For each species, *M. planium* links are colored according to their group (B: purple; A: red; D: blue; C: green; outgroups: grey) when observed. Average *SI* values per link per *Methylobacterium* group are indicated in the top diagram. Blue points indicate links that are more conserved in group D than group A. Average *SI* values with *M. planium* are indicated per species on the right diagram.