A genus definition for *Bacteria* and *Archaea* based on genome relatedness and taxonomic affiliation.

Barco, R.A.[1,5*], Garrity, G.M.[3], Scott, J.J.[4], Amend, J.P.[1,2], Nealson, K.H.[1,2], and Emerson, D.[5]

[1]Department of Earth Sciences, University of Southern California, Los Angeles, CA, USA.
[2]Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA.
[3]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA.
[4]Smithsonian Tropical Research Institute, Panama, Republic of Panama.
[5]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA.

Abstract

Genus assignment is fundamental in the characterization of microbes, yet there is currently no unambiguous way to demarcate genera solely using standard genomic relatedness indices. Here, we propose an approach to demarcate genera that relies on the combined use of the average nucleotide identity, genome alignment fraction, and the distinction between type species and non-type species. More than 750 genomes representing type strains of species from 10 different phyla, and 19 different taxonomic orders/families in Gram-positive/negative, bacterial and archaeal lineages were tested. Overall, all 19 analyzed taxa conserved significant genomic differences between members of a genus and type species of other genera in the same taxonomic family. *Bacillus*, *Flavobacterium, Hydrogenovibrio*, *Lactococcu*s, *Methanosarcina*, *Thiomicrorhabdus*, *Thiomicrospira*, *Shewanella*, and *Vibrio* are discussed in detail. Less than 1% of the type strains analyzed need reclassification, highlighting that the adoption of the 16S rRNA gene as a taxonomic marker has provided consistency to the classification of microorganisms in recent decades. One exception to this is the genus *Bacillus* with 61% of type strains needing reclassification, including the human pathogens *B. cereus* and *B. anthracis*. The results provide a first line of evidence that the combination of genomic indices provides appropriate resolution to effectively demarcate genera within the current taxonomic framework that is based on the 16S rRNA gene. We also identify the emergence of natural breakpoints at the genome level that can further help in the circumscription of genera. Altogether, these results show that a distinct difference between distant relatives and close relatives at the genome level (i.e., genomic coherence) is an emergent property of genera in *Bacteria* and *Archaea*.

Introduction

At the time of writing, 17,300 bacterial and archaeal species and over 3,000 genera with validly published names have been described in the taxonomic literature (Garrity, 2016; Namesforlife Database); however, based on simple phylogenetic definition there are over 200,000 bacterial and archaeal species and 60,000 genera so far detected in the SILVA database (Quast *et al*., 2013; Yarza *et al.*, 2014). Sequence data stored in the Joint Genome Institute (JGI) database, which includes data from other databases, has increased exponentially over the past decade (Markowitz *et al.*, 2015) with >130,000 genomes of

47 bacterial and archaeal isolates, >6,000 metagenome-assembled genomes, and >2,000 single-
48 cell amplified genomes currently available. Because the number of available datasets is
49 constantly increasing (Mukherjee *et al*., 2017), it is becoming easier to access data that
50 represent taxa of interest. Consequently, this added layer of available information could aid in
51 the formal characterization of microorganisms. An essential aspect of this characterization
52 process is the proper assignment of genus and species.
53 Historically, DNA-DNA hybridization (DDH) has been the 'gold standard' for species
54 delineation with a DDH value of ≥70% being recognized as the species boundary between
55 two strains (Tindall *et al.*, 2010 and other references therein). Stackebrandt and Goebel (1994)
56 conducted a correlation analysis between DDH values and 16S rRNA gene sequence
57 identities, and based on this, proposed a boundary of 16S rRNA gene sequence similarity of
58 97% for species delineation. This value, still largely used today for operational-taxonomic-
59 unit (OTU) based analysis of microbial communities, has been updated by Stackebrandt and
60 Ebers (2006) to a value between 98.7-99.0%, based on a greater amount of available sequence
61 data. Subsequently, as genome sequencing has become common, whole genome comparisons
62 became possible leading to the advent of genome relatedness indices such as the average
63 nucleotide identity (ANI), amino acid identity (AAI), and digital DDH. More recently, Kim *et*
64 *al.* (2014) proposed a 16S rRNA gene sequence similarity threshold value of 98.65% for
65 species delineation as it equated ANI values of 95-96%, which in turn have been equated to
66 the classical standard species delineation threshold DDH value of 70% (Goris *et al*., 2007;
67 Richter and Rosselló-Móra, 2009). A method that only relies on protein-coding genes (i.e.
68 neither rRNA nor tRNA are included in analysis) is the Microbial Species Identifier (MiSI),
69 which employs both alignment fractions (AF) and ANI for demarcation of species,
70 recommending threshold AF:ANI values of 0.6:96.5% using complete or nearly complete
71 genomes (Varghese *et al.*, 2015).
72 Despite these advancements in resolving species delineation, practical guidelines that
73 incorporate genomic properties to demarcate genera have been lacking even though genus
74 assignment is key to performing meaningful comparisons regarding the physiology,
75 metabolism and genomic potential of microbes. Methods to demarcate genera have been
76 proposed that are based on either the amino acid identity (AAI) or the percentage of
77 conserved proteins (Konstantinidis and Tiedje, 2007; Qin *et al*., 2014). However, these
78 methods directly rely on the 16S rRNA gene sequence, which is in some cases, insensitive to
79 evolutionary changes in the rest of the genome of a given organism, as revealed by different
80 species sharing >99% identity over the length of this gene. Additionally, the generally-used
81 arbitrary genus threshold of 95% 16S rRNA gene identity has been recently revisited to a
82 lower minimum value of 94.5% with confidence interval of 94.55-95.05 and median sequence
83 identity of 96.4% (Yarza *et al.*, 2014). In borderline cases, interpretation of results may be
84 unclear if there are no alternative ways to confirm genus assignment. This is also the case for
85 microorganisms with multiple, highly-divergent 16S rRNA genes. Here, we propose a novel
86 approach that is based on the application of the MiSI method (Varghese *et al.*, 2015) and
87 provide an objective and reproducible method of delimiting genera. We implement this
88 approach and method by testing a variety of taxonomic groups of *Bacteria* and *Archaea*.
89 Furthermore, supporting evidence is presented to support the reclassification of
90 *Hydrogenovibrio halophilus* and the extensive rearrangement of species within *Bacillus,*
91 including the human pathogens *B. anthracis* and *B. cereus*.
92

Materials and Methods

ANI and AF

ANI and alignment fraction (AF) values were obtained by the Microbial Species Identifier (MiSI) method (https://ani.jgi-psf.org/html/; Varghese *et al.*, 2015), which is also implemented in the JGI-Integrated Microbial Genomes (IMG) system (https://img.jgi.doe.gov/cgi-bin/m/main.cgi) via its Pairwise ANI tool. ANI, as defined by Varghese *et al.* (2015), is calculated for a pair of genomes by averaging the nucleotide identity of orthologous genes identified as bidirectional best hits (BBHs), which are the genes that show ≥70% sequence identity and ≥70% alignment of the shorter gene. AF, as defined by Varghese *et al.* (2015), is calculated as a fraction of the sum of the lengths of BBH genes divided by the sum of the lengths of all genes in a genome.

Strains

The Namesforlife Database (http://www.namesforlife.com; NamesforLife, LLC, East Lansing, MI; Garrity, 2010) was primarily used to retrieve information about type strains and type species associated with different genera. Complementary to this, equivalent strain numbers assigned by different biological resource centers (e.g., DSMZ - German Collection of Microorganism and Cell Cultures) were searched in their respective online catalogues. In addition to the above, primary literature sources were used to confirm some of these designations. All strain designations were cross-referenced in at least two databases.

Genomes and 16S rRNA gene sequences

Genomes were obtained from IMG or the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/). Generally, phyla with taxonomic families/orders that have available genomes of ≥4 type species of genera and ≥4 non-type species in a given genus were considered. Most of the genomes used in this study were sequenced directly by the JGI (47%), National Institute of Technology and Evaluation (https://www.nite.go.jp/; 5%), J. Craig Venter Institute (3%), and Broad Institute (3%). About 38% of the genomes included in this study were sequenced as part of the JGI-Genomic Encyclopedia of *Bacteria* and *Archaea*, a project that focuses on sequencing the genomes of type strains (Wu et al., 2009; Mukherjee et al., 2017). The rest of the genomes were sequenced by >140 different institutions. Sequences of the 16S rRNA gene were obtained from either IMG, NCBI, or EZBioCloud (http://www.ezbiocloud.net/). Alignments were separately performed in the SILVA Incremental Aligner version 1.2.11 (https://www.arb-silva.de/aligner/) and ClustalW via the Geneious platform (version R6; Biomatters, Auckland, New Zealand). Genetic distances were calculated in Geneious. PhyML (Guindon and Gascuel, 2003) was used via the Geneious platform to generate the maximum likelihood phylogenetic tree with the following settings: Hasegawa-Kishino-Yano (HKY85) substitution model, 1000 bootstraps, estimated transition/transversion ratio, estimated proportion of invariable sites, estimated gamma distribution, branch lengths and substitution rate optimized.

Pangenome Analysis

Pangenome analysis of genomes from *Bacillus* and other genera in *Bacillaceae* were processed in anvi'o (version 4)(Eren et al., 2015) following the workflow for microbial pangenomics (merenlab.org/2016/11/08/pangenomics-v2/, last accessed 13 April, 2018;

139 Delmont and Eren, 2018). In brief, we generated contig databases for each genome contig file
140 using the command 'anvi-gen-contigs-database'. Prodigal (Hyatt et al., 2010) was used to
141 identify open reading frames and subsequently each database was populated with HMM
142 profiles by comparison to several single-copy gene collections (Campbell et al., 2011; Rinke
143 et al., 2013) using HMMER (Eddy, 2011). Once contig databases were generated for all
144 genomes, we used 'anvi-gen-genomes-storage' to generate a master genome storage database
145 to use in the pangenome analysis. We used the ncbi-blast option in 'anvi-pan-genome' to
146 calculate gene similarity and MCL (van Dongen and Abreu-Goodger, 2012) for clustering
147 under the following settings: minbit, 0.5; mcl inflation, 2; minimum occurrence, 2. Finally, to
148 organize both gene clusters and genomes, we used Euclidian distance and Ward linkage.
149
150 Statistical Analysis
151 All bidirectional, matched pair values of ANI (ANI1-2; ANI2-1) and AF (AF1-2;
152 AF2-1) were reported as single averaged ANI and AF values in this study. The nonparametric
153 Wilcoxon test was performed separately for each ANI and AF values to determine significant
154 ($P<0.05$) differences between type- and non-type species. A discriminant multivariate analysis
155 was performed using quadratic fitting methods, which resulted in probabilities of group
156 assignment (i.e., type vs. non-type species). All statistical analyses were performed using the
157 statistical software JMP Pro, version 12 (SAS Institute Inc.).
158
159 Rationale
160 The nomenclatural type or "type" is defined by the International Code of
161 Nomenclature of Prokaryotes (ICNP) as "that element of the taxon with which the name is
162 permanently associated" (Parker et al., 2015). A type strain acts as the single reference for a
163 given species. Each validly published species has a type strain. A type species, which is
164 represented by its type strain, acts as the single reference for a given genus. Type strains and
165 type species do not necessarily have to be the most representative of the species or genus,
166 respectively. The rationale used in this study is that sister species in a given genus should be
167 relatively similar to the type species of the cognate genus indicating high similarity at the
168 genome level. Therefore, when a type species of a genus is compared to a non-type, sister
169 species of the same genus, the AF:ANI values should be relatively high. In contrast to this,
170 type species of different sister genera (i.e., in the same taxonomic order or family) should be
171 relatively dissimilar at the genome level. This dissimilarity should be reflected in relatively
172 low AF:ANI values.
173
174 Guidelines
175 In addition to the genome(s) under consideration (e.g., of a cultured strain that needs
176 classification), at least two genomes of type species in the same order or family are needed: a)
177 the primary reference which is the phylogenetically closest relative that is a type species (i.e.,
178 the type species that all other microbes will be compared to), b) all other type species (≥1) in
179 the same order or family. All available genomes of non-type, sister species that belong to the
180 same genus of the primary reference are needed. In order for the results to be meaningful, a
181 species must only be represented by the type strain unless its genome is not available, in
182 which case a non-type strain of the same species that is ≥99 % identical in 16S rRNA gene
183 sequence(s) (i.e., all copies of 16S rRNA gene in the genome) may be used. Accession
184 numbers for these genomes should be reported.

185
186 Interpretation
187 Every data point in the AF:ANI plot is calculated in relation to the primary reference
188 which is defined above as the type strain of the type species of the specific genus to be
189 analyzed; therefore, the plot represents genomic similarity to the reference genome. If the
190 primary reference is compared against an identical genome (e.g., or a direct comparison
191 against itself), an AF value of 1 and ANI value of 100 would result. If the primary reference is
192 compared against relatively dissimilar genomes, the AF and ANI values will reflect relatively
193 lower values. Thus, the plots that are generated effectively illustrate how distant other data
194 points are to the primary reference in terms of AF and ANI values, or in other words, how
195 distant other data points are to the upper-right corner (AF=1 and ANI=100).
196 When all comparisons are made in relation to the primary reference, a type species
197 should always cluster with other type species in the same taxonomic order or family,
198 reflecting relatively low AF and ANI values. A non-type, sister species should always cluster
199 with other non-type, sister species of the same genus, reflecting relatively high AF and ANI
200 values. The distinct clustering of type species and non-type, sister species should be
201 conserved without any overlap. If a non-type, sister species is positioned within the cluster of
202 type species in the same taxonomic order or family, then this indicates that the non-type, sister
203 species could potentially be assigned to a new genus, as it is as genomically different to the
204 primary reference as are other type species in the same taxonomic order or family. If a type
205 species is positioned within the cluster of non-type, sister species, then there is support for
206 potential reclassification of that species within the same genus of the primary reference, given
207 the accumulation of other supporting evidence (e.g., 16S rRNA gene identity and phenotypic
208 properties). It is emphasized that any proposed taxonomic rearrangement or reclassification
209 based on AF:ANI results will necessarily affect nomenclature; therefore, any corresponding
210 changes in nomenclature resulting from such proposal must be in alignment with the ICNP
211 (Parker et al., 2015).
212
213 Results and Discussion
214 Table 1 summarizes the 19 separate taxonomic groups that were investigated in this
215 study along with results of statistical tests for both AF and ANI. A total of 790 genomes
216 including 302 type species and 488 non-type species, representing 705 type strains and 10
217 phyla were used. For each taxonomic group, the combination of AF and ANI genomic indices
218 resulted in distinct clustering of type and non-type species. The means of these clusters were
219 significantly different in all tested cases with AF as a variable. In the majority of the cases
220 (17/19), the means of these clusters were also significantly different with ANI as a variable.
221 Discriminatory multivariate analyses, which give a sense of how separated the clusters are,
222 supported the distinct clustering of type and non-type species in 18 out of 19 cases with 91-
223 100% of the species correctly classified. A few representative cases will be discussed in more
224 detail below.
225 To initially test our approach with Gram-positive bacteria, the order *Lactobacillales*
226 was chosen, since it contains a number of well-characterized genera and species. A continuum
227 of AF:ANI values characterizes a typical result when intergeneric and intrageneric species of
228 the order *Lactobacillales* (e.g., all against all) are compared without distinguishing between
229 type and non-type strains or species (Figure 1A). Because there is no primary taxonomic
230 reference genome to compare to, differentiation between different taxonomic groups is not

231 possible. However, when the type species of *Lactococcus, L. lactis* subsp. *lactis* ATCC
232 19435^T, is used as the primary reference, a clear distinction can be made between type species
233 and non-type species (Figure 1B). The type species of genera in *Lactobacillales* form a
234 distinct cluster toward lower AF:ANI values while the non-type species of *Lactococcus*
235 cluster toward higher AF:ANI values (Table 1; Figure 1C-D). There is no overlap between
236 these two clusters. Discriminatory multivariate analysis supported the distinct clustering of
237 type and non-type species successfully classifying 98% of the samples.
238 To test our approach with Gram-negative microorganisms, type species of genera in
239 the order *Alteromonadales* and non-type *Shewanella* spp. were pairwise-compared to the type
240 species *S. putrefaciens* JCM 20190^T, which is the primary reference (Figure 2A; Table 1). A
241 distinct clustering is seen between the type species and non-type species. Discriminatory
242 multivariate analysis also supported the distinct clustering of type and non-type species
243 successfully classifying 100% of the samples.
244 Additional examples show that this trend is also seen with taxonomic families (Figure
245 2B; Table 1). The *Flavobacteriaceae* was investigated as an example of a diverse family with
246 one of the largest number of validly published genera (currently >160). Despite this, the
247 separate clustering of type species and non-type species was conserved. When the type
248 species of the genus *Flavobacterium, F. aquatile* LMG 4008^T is used as the primary
249 reference, all type species (i.e. with available genomes) in the *Flavobacteriaceae* clustered
250 towards lower AF:ANI values, while the vast majority of the non-type species of
251 *Flavobacterium* clustered toward higher AF:ANI values. Discriminatory multivariate analysis
252 successfully classified 96% of the samples. The five non-type *Flavobacterium* spp. that were
253 outside the AF:ANI genus boundary had 16S rRNA gene sequence identities of <94% to the
254 primary reference. In a separate example, the type species *Vibrio cholerae* ATCC 14035^T was
255 used as the primary reference and compared to non-type species in the genus *Vibrio* and other
256 type species of genera in the family *Vibrionaceae* (Figure 2C; Table 1). A distinct clustering,
257 without overlap, is seen between the type species and non-type species. Discriminatory
258 multivariate analysis also supported the distinct clustering of type and non-type species with
259 >91% of the samples successfully classified, even though the clusters were in much closer
260 proximity to each other than in previous examples.
261 This approach was also tested on *Methanosarcinaceae*, an archaeal taxonomic family
262 in the phylum *Euryarchaeota* (Figure 2D; Table 1). The primary reference was the type
263 species *Methanosarcina barkeri* JCM 10043^T. The results were consistent with previous
264 examples on bacterial taxonomic groups displaying clustering of non-type species separately
265 from type species. Discriminatory multivariate analysis successfully classified 100% of the
266 samples. Similar results were obtained in the family *Haloferacaceae*. Additionally, the family
267 *Thermoproteaceae* in the phylum *Crenarchaeota* was tested with *Pyrobaculum islandicum*
268 DSM 4184^T as a primary reference (Figure S1). A significant distinction between type and
269 non-type species was seen with AF (P=0.0200) but not with ANI (P=0.713). Despite this,
270 there was no overlap of the type/non-type clusters and discriminatory multivariate analysis
271 successfully classified 100% of the samples.
272 In a similar manner, genus assignment in the recently rearranged *Hydrogenovibrio-*
273 *Thioalkalimicrobium-Thiomicrospira* cluster was tested as an example of a bacterial
274 taxonomic group with historical taxonomic issues (Takai *et al*., 2004; Tourova *et al*., 2006;
275 Figure S2). Boden *et al.* (2017) provided a detailed evaluation of the taxonomy of this cluster
276 which belongs to the family *Piscirickettsiaceae* and proposed to place many of the species

277 into different genera. We tested these newly proposed assignments using the approach
278 described above and the MiSI method. When testing genus assignment to *Hydrogenovibrio*,
279 the type species *H. marinus* DSM 11271^T was used as a primary reference (Figure 3). The
280 single data point that crossed the boundaries set by the AF:ANI values between type species
281 of genera in the family *Piscirickettsiaceae* and the non-type species of the genus
282 *Hydrogenovibrio* belonged to *Hydrogenovibrio halophilus* DSM 15072^T. Aside from this data
283 point, there was no overlap of AF:ANI values between the type- and non-type species,
284 indicating support for the assignment of *H. crunogenus*, *H. thermophilus*, and *H. kuenenii* to
285 the genus *Hydrogenovibrio*. Similarly, the recently proposed new genus *Thiomicrorhabdus*
286 (Boden *et al*., 2017) was also tested (Figure S3). The results indicate that the new
287 classification is supported by AF and ANI values as the type species of genera in the family
288 *Piscirickettsiaceae* and *Thiomicrorhabdus* spp. remain well separated. The reclassification of
289 *Thioalkalimicrobium* spp. as *Thiomicrospira* spp. (Boden *et al*., 2017) is also supported by
290 ANI and AF values (Figure S4), as the *Thioalkalimicrobium* spp. (now non-type
291 *Thiomicrospira* spp.) cluster away from the type species of genera in the family
292 *Piscirickettsiaceae*.
293 Our approach has identified a potential misclassification of *H. halophilus* (basonym
294 *Thiomicrospira halophila*) because it clusters with the type species of genera in the family
295 *Piscirickettsiaceae,* other than *Hydrogenovibrio*. Below we show additional evidence that
296 could support a potential reclassification of *H. halophilus*. A closer look at the multiple 16S
297 rRNA gene sequences of *H. halophilus* revealed that they have 94.0-94.2% pairwise-identities
298 to the type species, *H. marinus* DSM 11271^T. Our results differ from those of Boden *et al.*,
299 who reported 95.7 % sequence identity (2017) between *H. halophilus* and *H. marinus*. The
300 difference in identity appears to be due to differences in the 16S rRNA gene sequence
301 between the near-full-length version of Sorokin *et al*. (2006; 1420 bp; accession number
302 DQ390450) and the full-length versions of this gene (1439 bp each; IMG ID 2518265101 and
303 2518265550) originating from a draft genome (Scott *et al.,* 2018) and used in this study. The
304 phylogenetic placement of this organism also supports the idea that *H. halophilus* is distinctly
305 positioned in its own clade as it branches away from the cluster of other *Hydrogenovibrio* spp.
306 (Figure 3B). One differentiating characteristic of *H. halophilus* is the DNA G+C content of
307 56.6%, much higher than the ca. 44% of the other newly classified *Hydrogenovibrio* spp.
308 Another major differentiating aspect is the higher NaCl optimum/maximum of 1.5/3.5 M for
309 *H. halophilus* versus the lower 0.2-0.5/0.6-1.2 M for other *Hydrogenovibrio* spp. Therefore,
310 based on AF and ANI results, molecular, phylogenetic, and physiological evidence, *H.*
311 *halophilus* could potentially be assigned to a new genus.
312 The *Bacillaceae* was also investigated as an example of taxonomic family that is
313 medically relevant (Figure 4A). Generally, there were significant differences between non-
314 type *Bacillus* species and type species of genera in the family *Bacillaceae* when *B. subtilis*
315 6051^T (Kabisch et al., 2013) was used as a primary reference (Table 1; Figure S5). However,
316 a large number of non-type *Bacillus* species clustered with type species (50% misclassified by
317 multivariate, discriminatory method), which strongly suggests that the genus *Bacillus* is in
318 need of an extensive and thorough rearrangement. Of note, it is highlighted that human
319 pathogens *B. anthracis* and *B. cereus* cluster with type species of genera in *Bacillaceae*,
320 indicating that they are at least as genomically different to *B. subtilis* as are other type species
321 in this family, warranting a taxonomic rearrangement. As shown above with *H. halophilus*, a
322 similar case could be made for the taxonomic rearrangement of each one of these species

323 (e.g., <94.5% identity in 16S rRNA gene sequence to the type species *B. subtilis* 6051^T).
324 Compared to our analyses of other taxonomic groups, these results are atypical in the sense
325 that the non-type species cluster overlaps considerably with the type species cluster. We
326 highlight the fact that a large number of these specimens were designated non-type species of
327 *Bacillus* prior to the advent of the universal use of 16S rRNA gene as a taxonomic marker in
328 the 1990s (i.e., >20% of all type strains of non-type *Bacillus* with available genomes in our
329 dataset). For example, *B. anthracis* was first described in 1872 (Cohn). When the primary
330 reference is changed to *Anoxybacillus pushchinoensis* K1^T, a type species of a genus
331 established in 2000 (Pikuta *et al*.) using the taxonomic framework of the 16S rRNA gene, the
332 clustering of type species of genera in *Bacillaceae* and non-type species of *Anoxybacillus* is
333 distinct and without overlap (Table 1).
334 The AF:ANI plots reveal different rates of differentiation at the DNA level. In the few
335 cases where there are sufficient samples to analyze, and values are evenly distributed, the
336 AF:ANI plots reveal a polynomial shape approximating a quartic function, with R^2 values
337 >0.97. A second derivative of this function results in a quadratic function that can be used to
338 determine inflection points with precision. Estimates based on several examples in this study
339 indicate two inflection points. The first inflection point at 71-74% ANI, generally
340 approximates the genus demarcation line in all cases herein presented based on taxonomic
341 affiliation. However, the distribution of non-type, sister species in a given genus generally
342 includes values that fall below this inflection point at lower AF:ANI values, suggesting a
343 potential underestimation of genera diversity within a taxonomic order or family. Estimation
344 of this genus inflection point via Logistic and Gompertz functions generally agree with the
345 estimations by quartic function and their use is recommended for accuracy. The second
346 inflection point at 89-90% ANI, is consistently 5-7% lower than current ANI species
347 thresholds (95-96.5% ANI), and is only seen with the quartic function. Because the second
348 inflection point potentially deals with species delineation, detailed exploration (e.g., using
349 subspecies type strains) and discussion of this topic are outside of the scope of this
350 manuscript. However, it is noted that ANI values of 89-90% largely correspond to 16S rRNA
351 gene identity values of ≥98.65 (Figure 3 in Kim et al., 2014).
352 The genus inflection point serves as practical guide for maximum AF:ANI values of
353 genus boundary and for the identification of a genus transition zone. Ideally, and assuming
354 abundance of type strains genomes and consequently data points, the estimated genus
355 inflection point should closely match the genus boundary. As an example, three different
356 functions were used to estimate the genus inflection point in *Bacillus* (Figure 4B). The
357 obtained values give an estimation of this inflection point at AF=0.39-0.45 and ANI=72.71-
358 73.80. An analysis of the pangenome of *Bacillus* using the most conservative estimated genus
359 inflection point as a guide shows that only non-type *Bacillus* spp. above the genus inflection
360 point cluster with the type species *B. subtilis* (the primary reference) while *B. anthracis* and *B.*
361 *cereus* cluster with type species of other genera in *Bacillaceae* (Figure 5), corroborating, as
362 pointed out above, that these two pathogens warrant reclassification and renaming. However,
363 doing so would require considerable care so as not to raise objections over safety issues and a
364 push for conservation of the current names over newly proposed names (Parker *et al.*, 2015).
365 These results were further corroborated in phylogenomic analyses of *Bacillaceae* taking into
366 account either 138 single copy genes or only ribosomal genes (Figures S13-S14).
367 It is important to note that the genus inflection point is *estimated* based on nonlinear
368 regression and only serves as a guide for future taxonomic designation by providing

369 maximum AF:ANI values. This is particularly important in the case of a new taxonomic
370 designation with a genome displaying AF:ANI values that are right above the effective genus
371 boundary, in a transition zone; is it a novel type species or a novel non-type species? Having
372 an estimated inflection point helps in making this decision. Thus far, this could be properly
373 calculated only for a few taxonomic groups due to lack of genomes of type strains and/or
374 actual isolates. *B. pumilus*, *B. safensis*, *B. invictae*, *B. altitudinis*, *B. zhangzhouensis*, *B.*
375 *xiamenensis*, and *B. australimaris* are borderline species positioned within the limits of the
376 most and least conservative estimated genus inflection points. Their 16S rRNA gene identities
377 in relation to the primary reference range from 96.8% to 97.4%, suggesting that the genus
378 inflection point of *Bacillus* approximates a value that is much higher than the current genus
379 threshold based on 16S rRNA gene identity (94.5%).
380 The approach we have used to demarcate genera is complementary and does not
381 replace the conventional polyphasic method of designating a genus, which includes thorough
382 analyses of the full-length 16S rRNA gene, phylogeny, physiology and metabolism, among
383 other aspects. The AF:ANI boundary values for genus demarcation in a given order or family
384 can be refined as more type strains and type species are genomically sequenced and formally
385 described. This approach emphasizes the importance of type strains and type species in the
386 evaluation of taxonomy with genomic indices (e.g., ANI or AF), as these specimens function
387 as reference points for their respective taxonomic groups. In this respect, our approach is in
388 line with recommendations regarding taxonomy, including relevant comparisons to type
389 strains and type species for the characterization of novel *Bacteria* and *Archaea* (Tindall *et al.*,
390 2010). It is reiterated that any rearrangement or reclassification of taxa should be in alignment
391 with the ICNP (Parker *et al.*, 2015). In particular, a potential rearrangement of type species
392 must be carefully analyzed as a genus can only contain one type species, and rules and
393 principles of nomenclature must be followed in order to properly do that in accordance with
394 the ICNP.
395 Uncultivated *Bacteria* and *Archaea* represented by metagenomic assembled genomes
396 (MAGs) and/or single cell amplified genomes (SAGs) could potentially be used in the
397 implementation of this approach; however, it is noted that type strains and type species should
398 be used as references for such analyses, if genomes are available (see *Guidelines*). The
399 designation of type strains and type species is governed by the ICNP and require a thorough
400 characterization of the microorganism and the deposit of the type strain in at least two
401 different culture collections. Naturally, the proper characterization, and therefore the type
402 designation is, in most cases, impossible for uncultivated *Bacteria* and *Archaea* given the
403 current guidelines set by the ICNP. This could change if genomes are subsequently
404 considered type material, as recently proposed (Whitman, 2015) and discussed
405 (Konstantinidis *et al*., 2017). In lieu of type strains and type species, proper and consistent
406 genomic references must be established in order to analyze taxonomic groups without
407 cultivated representatives. Caution is advised in the interpretation of data resulting from
408 MAGs since they can represent composite genomes of different strains. In theory, SAGs
409 could serve as good taxonomic frames of reference for uncultivated genera; however, genome
410 incompleteness in SAGs is an issue that would need to be addressed.
411 The MiSI method explicitly excludes tRNA and rRNA genes in order to avoid
412 inflation of AF or ANI values. However, the current taxonomic framework is largely based on
413 the use of the 16S rRNA gene for classification. Therefore, the approach herein investigated is
414 not completely independent of the 16S rRNA gene. The fact that in all the groups tested, with

415 the exception of the *Bacillaceae* (note: although significant differences between type and non-
416 type species were still detected for this group), there was strong concordance of type species
417 of genera being delineated from non-type species of the primary reference provides evidence
418 to the success of the use of the 16S rRNA gene within the taxonomic framework. The 16S
419 rRNA gene has served and is continuing to serve the scientific community well, especially in
420 maintaining consistency in classifications. However, our estimation of inflection points
421 indicates that the 16S RNA gene identity minimum threshold value of 94.5% could
422 underestimate genera diversity in some taxa [e.g. *Bacillaceae (Bacillus)*] and overestimate it
423 in others [e.g. *Enterobacterales (Xenorhabdus)*], especially if multiple non-identical 16S
424 rRNA genes are present in the genomes. It is clear that different taxonomic groups are
425 characterized by different AF:ANI threshold values for genus demarcation. However, even
426 within established taxonomic groups, the effective AF:ANI threshold values will still be
427 "moving targets", dependent on the assignment and re-assignment of new type species. Thus,
428 a single, universal, stationary threshold for genus delineation will not be able to sensitively
429 resolve genus assignments for all taxa.
430 The reader is made aware that variations in bidirectional, matched pair AF values
431 (AF1-2; AF2-1) exist. These variations are usually small, but could be relevant in the
432 interpretation of data, especially in borderline cases. Bidirectional, matched pair ANI values
433 (ANI1-2; ANI2-1) differ by <0.1 % in the majority of the cases, with a few cases differing by
434 <0.3%, which would make the error bars invisible to the naked eye. Figures with
435 corresponding standard deviation values for AF are included in a supplemental file (Figure
436 S6-S12).
437 The use of ANI has been previously stated as not applicable for the demarcation of
438 genera (Qin *et al*, 2014; Kim *et al*., 2014). Our results indicate that the MiSI method, and
439 therefore AF and ANI, can be used to visualize natural breakpoints that can be used to
440 circumscribe genera, if the guidelines presented in this study are followed. Adaptation of this
441 method to demarcate higher taxonomic ranks, including phyla, has not been tested as it is
442 beyond the scope of this study. Nonetheless, such an approach would be warranted. Finally,
443 we highlight that the results demonstrate a conserved genomic coherence at the genus level
444 for numerous different taxa, shedding light on a fundamental emergent property of *Bacteria*
445 and *Archaea*, a property that has proven largely elusive at the species level.
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460

Table 1. Summary of taxonomic groups and genomes included in this study. Gray fill-color indicates if either the taxonomic family or order was investigated. For a detailed list including accession numbers, see Table S1.

									Wilcoxon Test (T vs. NT)		Discriminatory Test
Domain	**Phylum**	**Class**	**Order***	**Family***	**Genus**	**Primary reference**	**# of Type Species of Genera§**	**# of Non-Type Species**	**AF (P-value)**	**ANI (P-value)**	**% Correctly Classified**
Archaea	*Crenarchaeota*	*Thermoprotei*	*Thermoproteales*	*Thermoproteaceae*	*Pyrobaculum*	*P. islandicum* DSM 4184^T	5	5	0.0200	0.7130	100
Archaea	*Euryarchaeota*	*Halobacteria*	*Haloferacales*	*Haloferacaceae*	*Haloferax*	*H. volcanii* DS2^T	9	10	0.0004	0.0004	100
Archaea	*Euryarchaeota*	*Methanococci*	*Methanosarcinales*	*Methanosarcinaceae*	*Methanosarcina*	*M. barkeri* JCM 10043^T	7	9	0.0016	0.0018	100
Bacteria	*Actinobacteria*	*Actinomycetes*	*Actinomycetales*	*Microbacteriaceae*	*Curtobacterium*	*C. citreum* NS330#	23	4	0.0020	0.0020	100
Bacteria	*Bacteroidetes*	*Flavobacteriia*	*Flavobacteriales*	*Flavobacteriaceae*	*Flavobacterium*	*F. aquatile* LMG 4008^T	70	83	<0.0001	<0.0001	96
Bacteria	*Bacteroidetes*	*Flavobacteriia*	*Flavobacteriales*	*Flavobacteriaceae*	*Maribacter*	*M. sedimenticola* DSM 19840^T	70	12	<0.0001	<0.0001	100
Bacteria	*Deinococcus-Thermus*	*Deinococci*	*Thermales*	*Thermaceae*	*Thermus*	*T. aquaticus* YT-1^T	4	12	0.0113	0.0115	100
Bacteria	*Firmicutes*	*Bacilli*	*Bacillales*	*Bacillaceae*	*Anoxybacillus*	*A. pushchinoensis* K1^T	27	8	<0.0001	<0.0001	97
Bacteria	*Firmicutes*	*Bacilli*	*Bacillales*	*Bacillaceae*	*Bacillus*	*B. subtilis* 6051-HGW^T	"	110	<0.0001	<0.0001	50
Bacteria	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Streptococcaceae*	*Lactococcus*	*L. lactis* ATCC 19435^T	38	6	0.0001	0.0001	97
Bacteria	*Fusobacteria*	*Fusobacteriia*	*Fusobacteriales*	*Leptotrichiaceae*	*Streptobacillus*	*S. moniliformis* DSM 12112^T	9	3	0.0184	0.0189	100
Bacteria	*Fusobacteria*	*Fusobacteriia*	*Fusobacteriales*	*Fusobacteriaceae*	*Fusobacterium*	*F. nucleatum* ATCC 25586^T	"	9	0.0081	0.0081	100
Bacteria	*Planctomycetes*	*Planctomycetacia*	*Planctomycetales*	*Planctomycetaceae*	*Rhodopirellula*	*R. baltica* SH 1^T	13	2	0.0358	0.0354	100
Bacteria	*Proteobacteria*	*Gammaproteobacteria*	*Alteromonadales*	*Shewanellaceae*	*Shewanella*	*S. putrefaciens* JCM 20190^T	20	23	<0.0001	<0.0001	100
Bacteria	*Proteobacteria*	*Deltaproteobacteria*	*Desulfuromonadales*	*Geobacteraceae*	*Geobacter*	*G. metallireducens* GS-15^T	6	12	0.0019	0.0019	100
Bacteria	*Proteobacteria*	*Gammaproteobacteria*	*Enterobacterales*	*Morganellaceae*	*Xenorhabdus*	*X. nematophila* ATCC 19061^T	46	10	<0.0001	<0.0001	100
Bacteria	*Proteobacteria*	*Gammaproteobacteria*	*Vibrionales*	*Vibrionaceae*	*Vibrio*	*V. cholerae* ATCC 14035^T	6	65	0.0005	0.0005	91
Bacteria	*Proteobacteria*	*Gammaproteobacteria*	*Thiotrichales*	*Piscirickettsiaceae*	*Hydrogenovibrio*	*H. marinus* DSM 11271^T	11	4	0.013	0.0089	100
Bacteria	*Spirochaetes*	*Spirochaetia*	*Spirochaetales*	*Spirochaetaceae*	*Leptospira*	*L. interrogans* ATCC 43642^T	6	19	0.0008	0.1768	100

* = shading indicates the taxonomic resolution at which the group was analyzed.

§ = including the primary reference.

= *>99% identity to 16S rRNA gene sequence of C. citreum* DSM 20528^T, which does not have a sequenced genome.

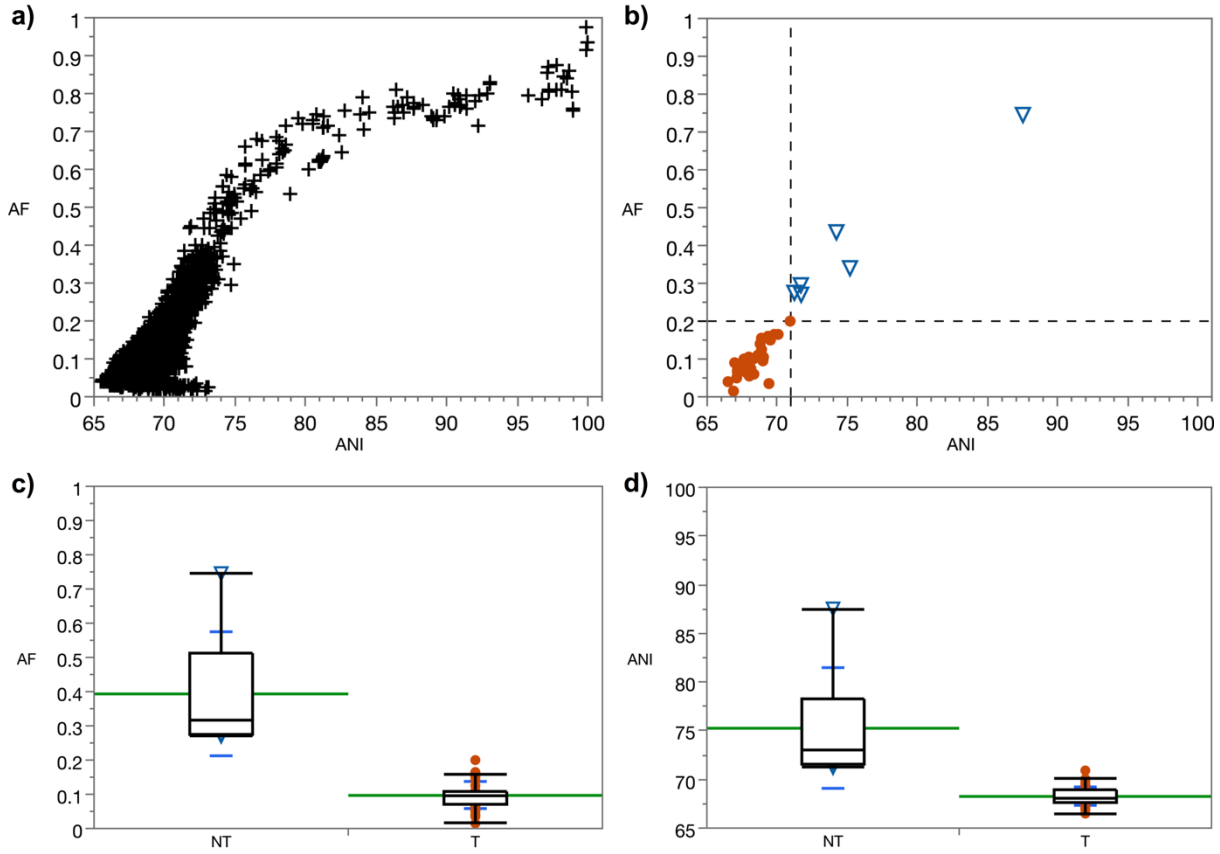


Figure 1. AF:ANI pairwise genome comparisons in the order *Lactobacillales*. A. Type and non-type strains/species of 15 different genera in the order *Lactobacillales* were pairwise-compared (n=6090 comparisons). B. Type species (n=37) of genera within the order *Lactobacillales* (circles) and non-type species (n=6) of the genus *Lactococcus* (in order *Lactobacillales*; triangles) were pairwise-compared *only* to the primary reference, type species *Lactococcus lactis* subsp. *lactis* ATCC 19435^{T}. The bottom-left quadrant demarcates the boundary between type and non-type species. Boxplot diagrams of AF (C) and ANI (D) values as related to non-type species (NT) of the genus *Lactococcus* and type species (T) of genera within the order *Lactobacillales*. Means are shown in green. Standard deviations are shown in blue.

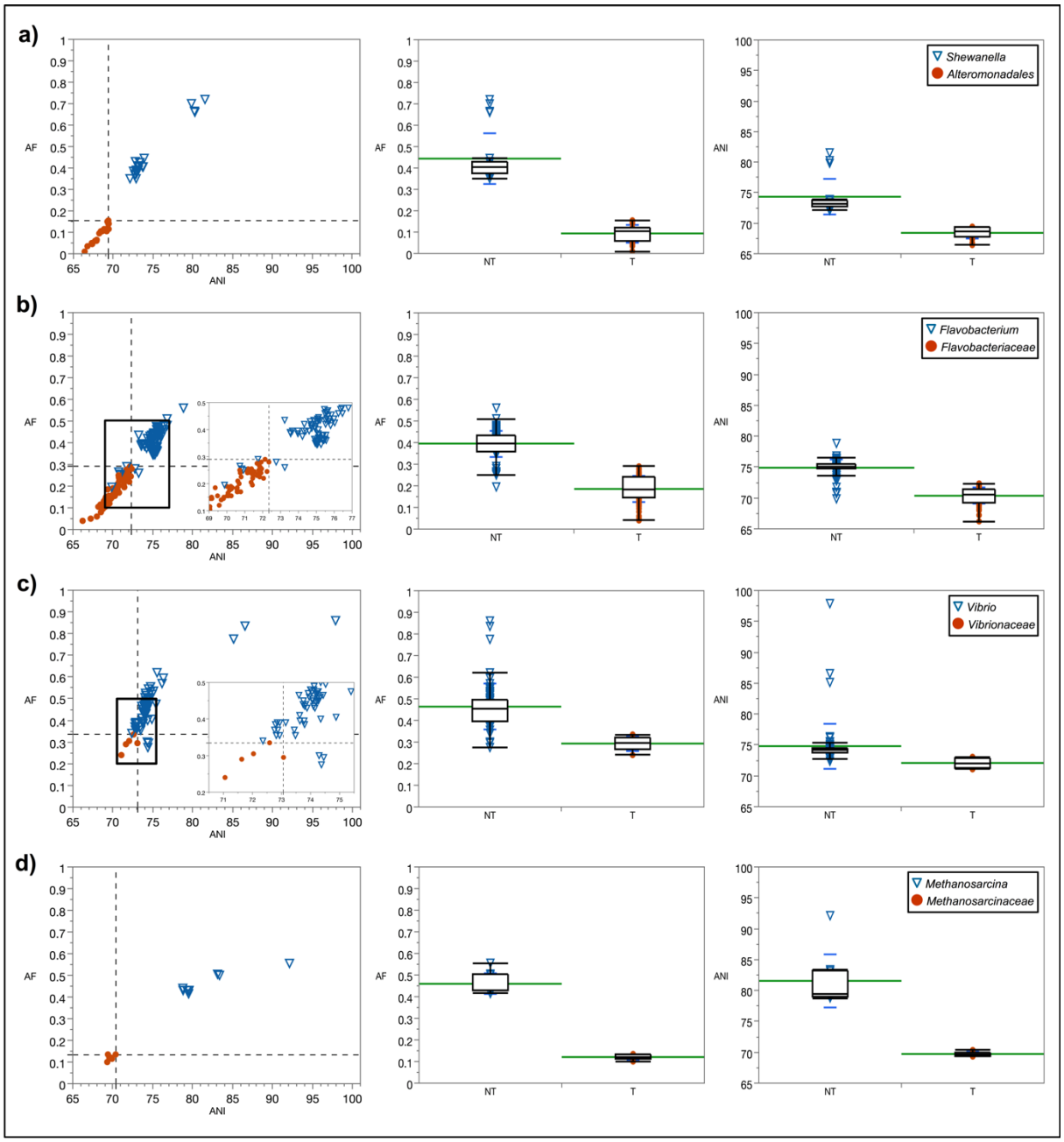


Figure 2. Pairwise genome comparisons to the following primary references: a) type species of the genus *Shewanella*: *S. putrefaciens* JCM 20190^{T}. In circles: type species (n=19) of genera within the order *Alteromonadales*. In triangles: non-type species (n=23) of the genus *Shewanella*; b) type species of the genus *Flavobacterium*: *F. aquatile* LMG 4008^{T}. In circles: type species (n=69) of genera within the family *Flavobacteriaceae*. In triangles: non-type species (n=83) of the genus *Flavobacterium*. Inset: zoomed-in boxed area; c) type species of the genus *Vibrio*: *V. cholerae* ATCC 14035^{T}. Circles: type species (n=5) of genera within the family *Vibrionaceae*. Triangles: non-type *Vibrio* spp. (n=65). Inset: zoomed-in boxed area; d) type species of the genus *Methanosarcina*: *M. barkeri* JCM 10043^{T}. Circles: type species (n=6) of genera within the family *Methanosarcinaceae*. Triangles: non-type *Methanosarcina*

spp. (n=9). In all cases, the bottom-left quadrant demarcates the boundary between type and non-type species. Boxplots of AF (center panel) and ANI (right panel) values as related to non-type (NT) species of genus analyzed and type species (T) of genera within the order or family analyzed. Means are shown in green. Standard deviations are shown in blue.

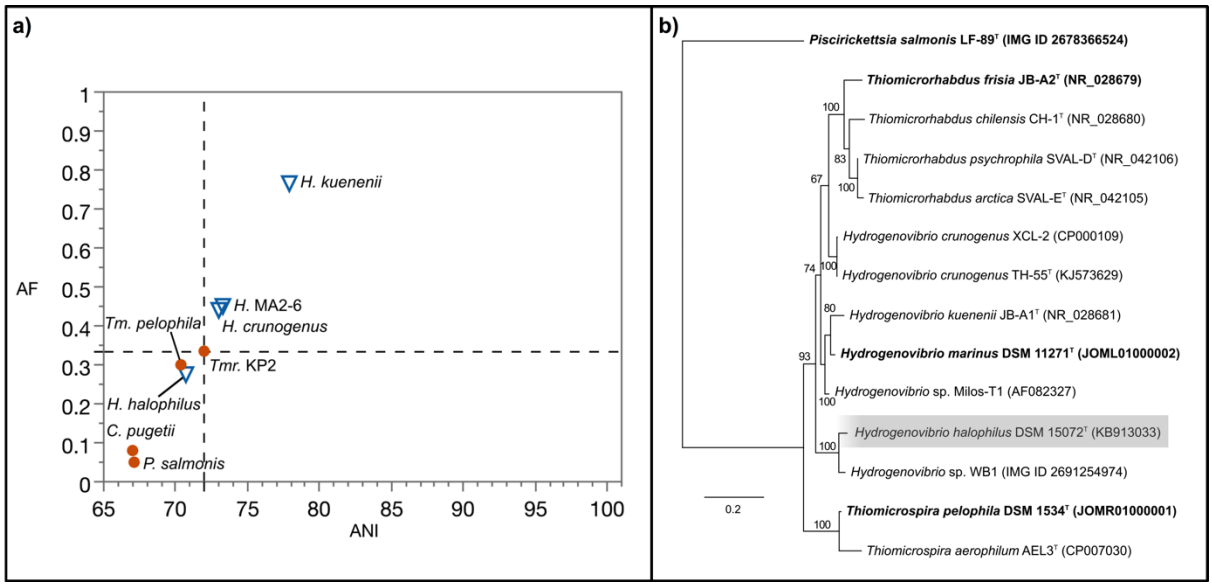


Figure 3. Analysis of *Hydrogenovibrio*, as rearranged by Boden *et al.* (2007). A. Pairwise genome comparisons to the primary reference *Hydrogenovibrio marinus* DSM 11271^T. Non-type species within the genus *Hydrogenovibrio* are shown in triangles. The type strain *H. thermophilus* I78^T does not have a sequenced genome; therefore, *H.* sp. MA2-6 (>99% pairwise identity in 16S rRNA gene sequence) was used instead. Similarly, *H. crunogenus* XCL-2 (>99% pairwise identity in 16S rRNA gene sequence) is used instead of the type strain *H. crunogenus* TH-55^T, as the latter does not have a sequenced genome. Type species of genera within the family *Piscirickettsiaceae* are shown in circles. The bottom-left quadrant demarcates the boundary between type and non-type species. B. Maximum-likelihood phylogenetic tree based on an alignment of 16S rRNA genes indicating the phylogenetic positioning of *H. halophilus* DSM 15072^T. Type species of genera are shown in bold. The scale bar indicates 20% sequence divergence.

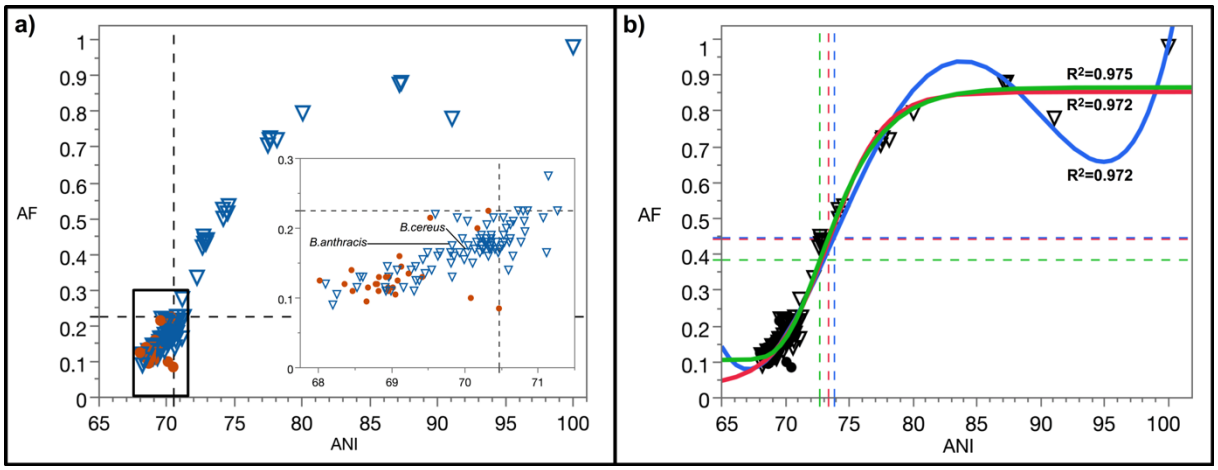
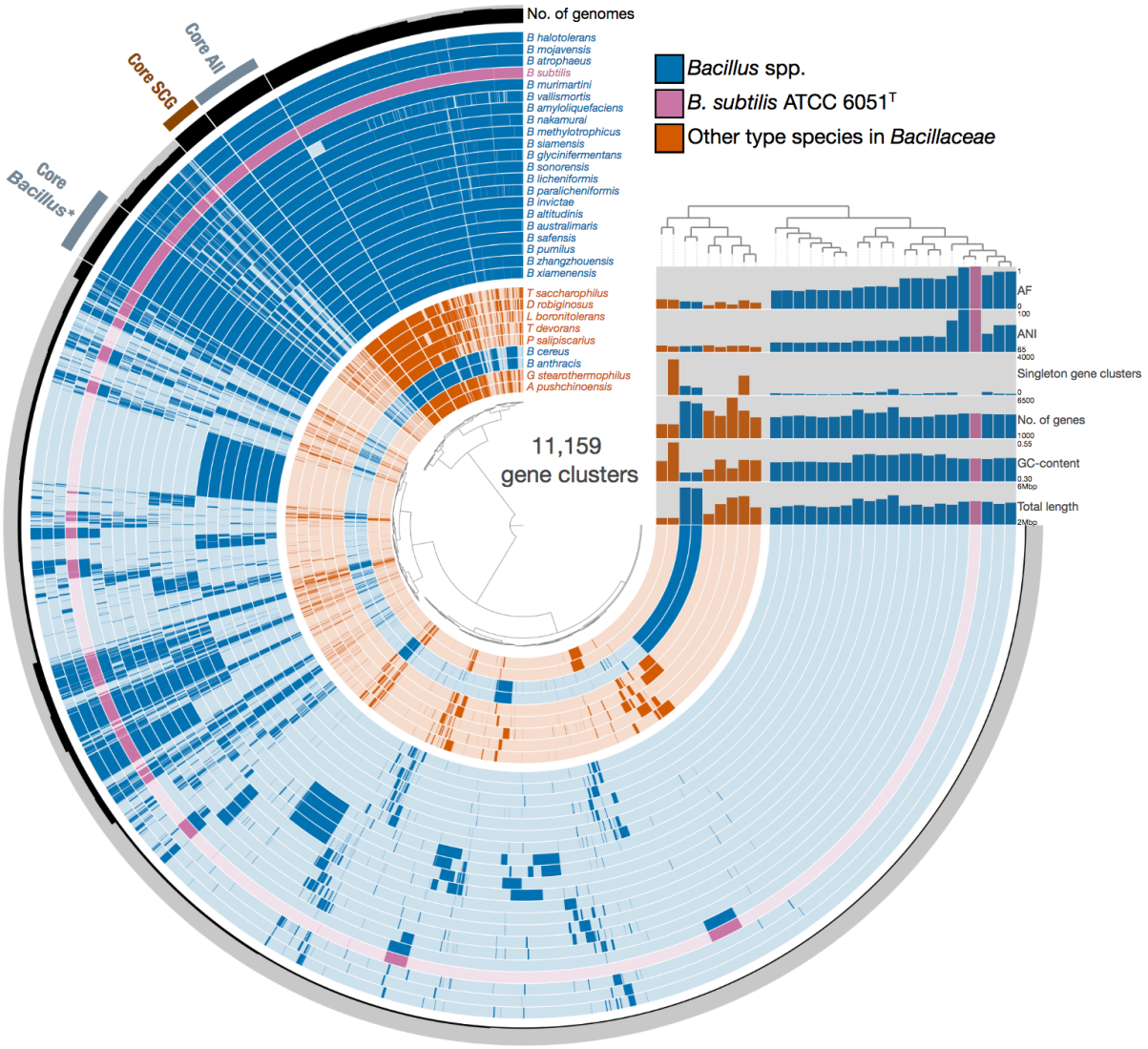


Figure 4. Pairwise genome comparisons to the type species of the genus *Bacillus*: *B. subtilis* ATCC 6051^{T}. Circles: type species (n=26) of genera within the family *Bacillaceae*. Triangles: non-type *Bacillus* spp. (n=110). a) The bottom-left quadrant (black dashed-line) demarcates the boundary between type and non-type species; inset: zoomed-in boxed area. b) Estimated inflection points (colored dashed-lines) for *Bacillus* based on quartic (blue), logistic 4P (red), and Gompertz 4P (green) functions. The quartic function is forced to go through the AF:ANI value of 1:100.

556



557
558
559 Figure 5. Pangenome of *Bacillus* (in blue). Type species of genera in the family *Bacillaceae*
560 (i.e., other than *B. subtilis* ATCC 6051^{T}) that are relatively close to the genus demarcation
561 AF:ANI threshold (see Figure 4A) are shown in orange. For this pangenomic view, only
562 *Bacillus* spp. with AF:ANI values above an estimated genus inflection point (i.e.,
563 >0.39:72.71; Figure 4B) are shown, with exception of *B. anthracis* (0.18:69.8) and *B. cereus*
564 (0.17:70.0). ANI and AF values were obtained via pairwise genome comparisons to the
565 primary reference, *B. subtilis* ATCC 6051^{T} (in purple). Hierarchical clustering was performed
566 on the presence-absence of gene clusters using Euclidean distance and Ward linkage. * =
567 *Bacillus* spp. above the estimated genus inflection point; SCG = single copy genes.
568
569 Acknowledgements
570
571 We would like to thank Anna Ratner at JGI for updating the interface of the Pairwise ANI
572 tool on the IMG platform. We would also like to thank Trudy Wassenaar and Nikos Kyrpides

for comments and suggestions on this manuscript. RAB was supported by the National Science Foundation postdoctoral research fellowship in Biology (award #1523639). JJS was supported by the Gordon and Betty Moore Foundation through a grant to the Smithsonian Tropical Research Institute (GBMF5603).

References:

Boden, R., Scott, K.M., Williams, J., Russel, S., Antonen, K., Rae, A.W., and Hutt, L.P. (2017) An evaluation of *Thiomicrospira*, *Hydrogenovibrio* and *Thioalkalimicrobium*: reclassification of four species of *Thiomicrospira* to each *Thiomicrorhabdus* gen. nov. and *Hydrogenovibrio*, and reclassification of all four species of *Thioalkalimicrobium* to *Thiomicrospira*. *Int. J. Syst. Evol. Microbiol*. 67:1140-1151. doi:10.1099/ijsem.0.001855

Campbell, B.J., Yu, L., Heidelberg, J.F., and Kirchman, D.L. (2011) Activity of abundant and rare bacteria in a coastal ocean. *Proc. Natl. Acad. Sci. USA*. 108(31):12776-12781. doi:10.1073/pnas.1101405108

Cohn, F. (1872) Untersuchungen über Bakterien. *Beitr. Biol. Pflanz*. 1:127-224.

Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., Delmont, T.O. (2015) Anvi'o: an advanced analysis and visualization platform for 'omics' data. *PeerJ*. 3:e1319.

Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Comp. Biol*. 7(10):e1002195. doi:10.1371/journal.pcbi.1002195.

Delmont, T.O. and Eren, A.M. (2018) Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ*. 6:e4320.

Garrity, G.M. (2010) NamesforLife: BrowserTool takes expertise out of the database and puts it right in the browser. *Microbiology Today*. 2(2):9.

Garrity, G.M. (2016) A new genomics-driven taxonomy of Bacteria and Archaea: are we there yet? *J. Clin. Microbiol*. 54:1956-1963. doi:10.1128/JCM.00200-16

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P. and Tiedje, J.M. (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol*. 57:81-91. doi:10.1099/ijs.0.64483-0

Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*. 52(5):696-704.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 11:119. doi:10.1186/1471-2105-11-119

619
620 Kabisch, J., Thürmer, A., Hübel, T., Popper, L., Daniel, R., Schweder, T. (2013) Journal of
621 Biotechnology. 163(2):91-04. doi:10.1016/j.jbiotec.2012.06.034
622
623 Kim, M., Oh, H-S., Park, S-C., and Chun, J. (2014) Towards a taxonomic coherence between
624 average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation
625 of prokaryotes. *Int. J. Syst. Evol. Microbiol*. 64:346-351. doi:10.1099/ijs.0.059774-0
626
627 Konstantinidis, K.T. and Tiedje, J.M. (2005) Towards a genome-based taxonomy for
628 prokaryotes. *J. Bacteriol.* 187(18):6258-6264. doi:10.1128/JB.187.18.6258-6264.2005
629
630 Konstantinidis, K.T. and Tiedje, J.M. (2007) Prokaryotic taxonomy and phylogeny in the
631 genomic era: advancements and challenges ahead. *Current Opinion in Microbiology*. 10:504-
632 509.
633
634 Konstantinidis, K.T., Rosselló-Móra, R., and Amann, R. (2017) Uncultivated microbes in
635 need of their own taxonomy. *ISME J.* 11:2399-2406.
636
637 Markowitz, V.M., Chen, I.M., Chu, K., Pati, A., Ivanova, N.N., and Kyrpides, N.C. (2015)
638 Ten years of maintaining and expanding a microbial genome and metagenome analysis
639 system. *Trends in Microbiology*. 23(11):730-741. doi:10.1016/j.tim.2015.07.012
640
641 Mukherjee, S., Seshadri, R., Varghese, N.J., Eloe-Fadrosh, E.A., Meier-Kolthoff, J.P., Goker,
642 M., Coates, R.C., Hadjithomas, M., Pavlopoulos, G.A., Paez-Espino, D., Yoshikuni, Y.,
643 Visel, A., Whitman, W.B., Garrity, G.M., Eisen, J.A., Hugenholtz, P., Pati, A., Ivanova, N.N.,
644 Woyke, T., Klenk, H.-P., and Kyrpides, N.C. (2017) 1,003 reference genomes of bacterial and
645 archaeal isolates expand coverage of the tree of life. *Nature Biotechnology*. 35:676-683.
646 doi:10.1038/nbt.3886
647
648 Parte, A.C. (2014) LPSN — list of prokaryotic names with standing nomenclature. *Nucleic*
649 *Acids Research*. 42:D613-D616. doi:10.1093/nar/gkt1111
650
651 Parker, C.T., Tindall, B.J., Garrity, G.M. (2015) International Code of Nomenclature of
652 Prokaryotes. *Int. J. Syst. Evol. Microbiol*. doi:10.1099/ijsem.0.000778
653
654 Pikuta, E., Lysenko, A., Chuvilskaya, N., Mendrock, U., Hippe, H., Suzina, N., Nikitin, D.,
655 Osipov, G., Laurinavichius, K. (2000) *Anoxybacillus pushchinensis* gen. nov., sp. nov., a
656 novel anaerobic, alkaliphilic, moderately thermophilic bacterium from manure, and
657 description of *Anoxybacillus flavithermus* comb. nov. *Int. J. Syst. Evol. Microbiol.* 50:2109-
658 2117. doi:10.1099/00207713-50-6-2109
659
660 Pruesse, E., Peplies, J. and Glöckner, F.O. (2012) SINA: accurate high-throughput multiple
661 sequence alignment of ribosomal RNA genes. *Bioinformatics.* 28:1823-1829.
662

Qin, Q., Xie, B., Zhang, X., Chen, X., Zhou, B., Zhou, J., Oren, A., Zhang, Y. (2014) A proposed genus boundary for the Prokaryotes based on genomic insights. *Journal of Bacteriology*. 196(12): 2210-2215. doi:10.1128/JB.01688-14

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 41:D590-D596.

Richter, M. and Rosselló-Móra, R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. USA.* 106(45):19126-19131. doi:10.1073/pnas.0906412106

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., Dodsworth, J.A., Hedlund, B.P., Tsiamis, G., Sievert, S.M., Liu, W.-T., Eisen, J.A., Hallam, S.J., Kyrpides, N.C., Stepanauskas, R., Rubin, E.M., Hugenholtz, P., and Woyke, T. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature*. 499:431-437. doi:10.1038/nature12352

Scott, K.M., Williams, J., Porter, C.M.B., Russel, S., Harmer, T.L., Paul, J.H., Antonen, K.M., Bridges, M.K., Camper, G.J., Campla, C.K., Casella, L.G., Chase, E., Conrad, J.W., Cruz, M.C., Dunlap, D.S., Duran, L., Fahsbender, E.M., Goldsmith, D.B., Keeley, R.F., Kondoff, M.R., Kussy, B.I., Lane, M.K., Lawler, S., Leigh, B.A., Lewis, C., Lostal, L.M., Marking, D., Mancera, P.A., McClenthan, E.C., McIntyre, E.A., Mine, J.A., Modi, S., Moore, B.D., Morgan, W.A., Nelson, K.M., Nguyen, K.N., Ogburn, N., Parrino, D.G., Pedapudi, A.D., Pelham, R.P., Preece, A.M., Rampersad, E.A., Richardson, J.C., Rodgers, C.M., Schaffer, B.L., Sheridan, N.E., Solone, M.R., Staley, Z.R., Tabuchi, M., Waide, R.J., Wanjugi, P.W., Young, S., Clum, A., Daum, C., Huntemann, M., Ivanova, N., Kyrpides, N., Mikhailova, N., Palaniappan, K., Pillay, M., Reddy, T.B.K., Shapiro, N., Stamatis, D., Varghese, N., Woyke, T., Boden, R., Freyermuth, S.K., Kerfeld, C.A. (2018) Genomes of ubiquitous marine and hypersaline *Hydrogenovibrio*, *Thiomicrorhabdus* and *Thiomicrospira* spp. encode a diversity of mechanisms to sustain chemolithoautotrophy in heterogeneous environments. *Environ. Microbiol*. doi:10.1111/1462-2920.14090

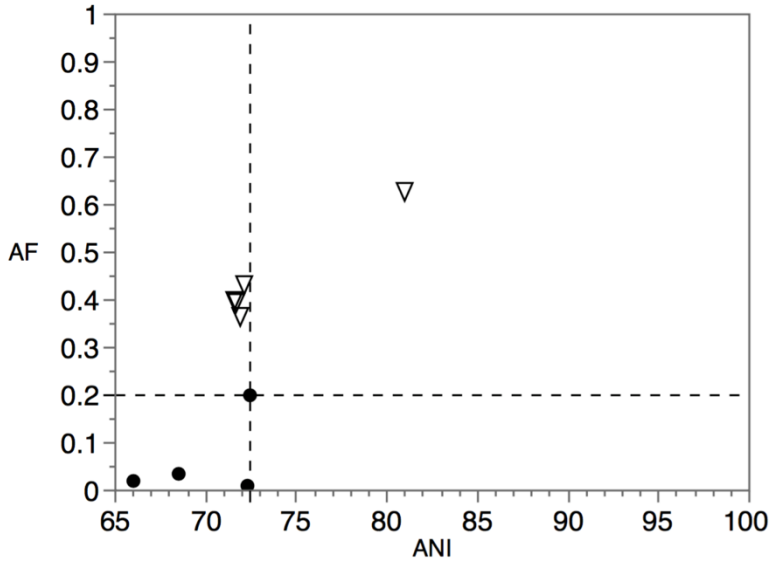
Stackebrandt, E. and Goebel, B.M. (1994) Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in Bacteriology. *Int. J. Syst. Bacteriol*. 44(4):846-849.

Stackebrandt, E. and Ebers, J. (2006) Taxonomic parameters revisited: tarnished gold standards. *Microbiol. Today*. 33:152-155.

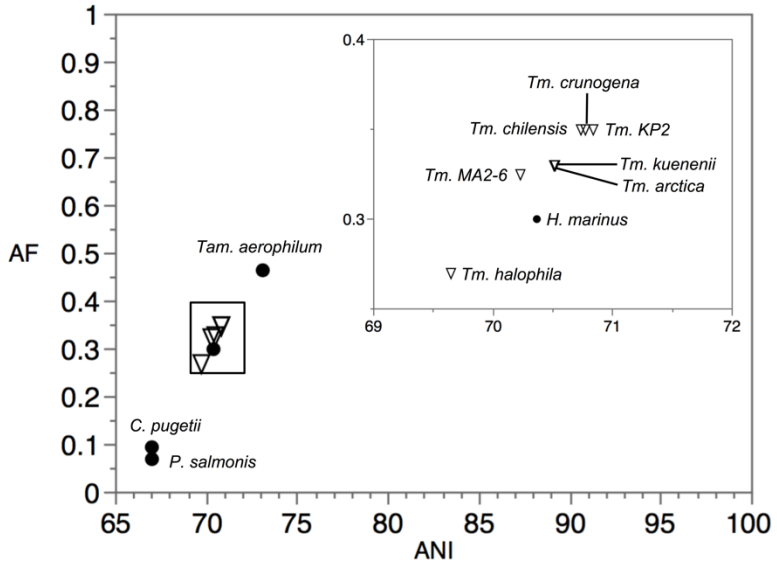
Sorokin, D.Y., Tourova, T.P., Kolganova, T.V., Spiridonova, E.M., Berg, I.A., and Muyzer, G. (2006) *Thiomicrospira halophila* sp. nov., a moderately halophilic, obligately chemolithoautotrophic, sulfur-oxidizing bacterium from hypersaline lakes. *Int. J. Syst. Evol. Microbiol*. 56:2375-2380. doi:10.1099/ijs.0.64445-0

708 Takai, K., Hirayama, H., Nakagawa, T., Suzuki, Y., Nealson, K.H., and Horikoshi, K. (2004)
709 *Thiomicrospira thermophila* sp. nov., a novel microaerobi, thermotolerant, sulfur-oxidizing
710 chemolithomixotroph isolated from a deep-sea hydrothermal fumarole in the TOTO caldera,
711 Mariana Arc, Western Pacific. *Int. J. Syst. Evol. Microbiol.* 54:2325-2333.
712 doi:10.1099/ijs.0.63284-0
713
714 Tindall, B.J., Rosselló-Móra, R., Busse, H.-J., Ludwig, W. and Kämpfer, P. (2010) Notes on
715 the characterization of prokaryote strains for taxonomic purposes. *Int. J. Syst. Evol.*
716 *Microbiol.* 60:249-266.
717
718 Tourova, T.P., Spiridonova, E.M., Berg, I.A., Kuznetsov, B.B., and Sorokin, D.Y. (2006)
719 Occurrence, phylogeny and evolution of ribulose-1,5-bisphosphate carboxylase/oxygenase
720 genes in obligately chemolithoautotrophic sulfur-oxidizing bacteria of the genera
721 *Thiomicrospira* and *Thioalkalimicrobium*. *Microbiology*. 152:2159-2169.
722 doi:10.1099/mic.0.28699-0
723
724 van Dongen, S. and Abreu-Goodger, C. (2012) Using MCL to extract clusters from networks.
725 *Methods Mol. Biol.* 804:281-295. doi:10.1007/978-1-61779-361-5_15
726
727 Varghese, N.J., Mukherjee, S., Ivanova, N., Konstantinidis, K.T., Mavrommatis, K.,
728 Kyrpides, N.C., and Pati, A. (2015) Microbial species delineation using whole genome
729 sequences. *Nucleic Acids Research.* 43(14): 6761-6771. doi:10.1093/nar/gkv657
730
731 Verslyppe, B., De Smet, W., De Baets, B., De Vos, P., and Dawyndt, P. (2014) StrainInfo
732 introduces electronic passports for microorganisms. *Syst. Appl. Microbiol.* 37(1):42-50.
733 doi:10.1016/j.syapm.2013.11.002
734
735 Whitman, W.B. (2015) Genome sequences as the type material for taxonomic descriptions of
736 prokaryotes. *Syst. Appl. Microbiol*. 38:217-222.
737
738 Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N.N., Kunin, V.,
739 Goodwin, L., Wu, M., Tindall, B.J., Hooper, S.D., Pati, A., Lykidis, A., Spring, S., Anderson,
740 I.J., D'haeseleer, P., Zemla, A., Singer, M., Lapidus, A., Nolan, M., Copeland, A., Han, C.,
741 Chen, F., Cheng, J.-F., Lucas, S., Kerfeld, C., Lang, E., Gronow, S., Chain, P., Bruce, D.,
742 Rubin, E.M., Kyrpides, N.C., Klenk, H.-P., and Eisen, J.A. (2009) A phylogeny-driven
743 genomic encyclopaedia of Bacteria and Archaea. *Nature*. 462(7276):1056-1060.
744 doi:10.1038/nature08656
745
746 Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K., Whitman,
747 W.B., Euzéby, J., Amann, R., and Móra, R.R. (2014). Uniting the classification of cultured
748 and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews*
749 *Microbiology*. 12: 635-645. doi:10.1038/nrmicro.3330
750
751 Yoon, S. H., Ha, S. M., Kwon, S., Lim, J., Kim, Y., Seo, H. and Chun, J. (2017). Introducing
752 EzBioCloud: A taxonomically united database of 16S rRNA and whole genome
753 assemblies. *Int. J. Syst. Evol. Microbiol*. 67(5):1613-1617. doi:10.1099/ijsem.0.001755.

754 =========
755 Supplemental Section
756



757
758
759 Figure S1. Pairwise genome comparisons to the type species of the genus *Pyrobaculum*: *P.*
760 *islandicum* DSM 4184^{T}. Circles: type species (n=4) of genera within the family
761 *Thermoproteaceae*. Triangles: non-type *Pyrobaculum* spp. (n=5).
762



763
764
765 Figure S2. Pairwise genome comparisons to the type species *Thiomicrospira pelophia* DSM
766 1534^{T} using the old classification scheme (*i.e.*, prior to rearrangement by Boden *et al*., 2017)
767 show clear taxonomic issues and no distinct clustering between type and non-type species.
768 Non-type species within the genus *Thiomicrospira* are shown in triangles. Type species of
769 genera within the family *Piscirickettsiaceae* are shown in circles. Inset: zoomed-in boxed
770 area.

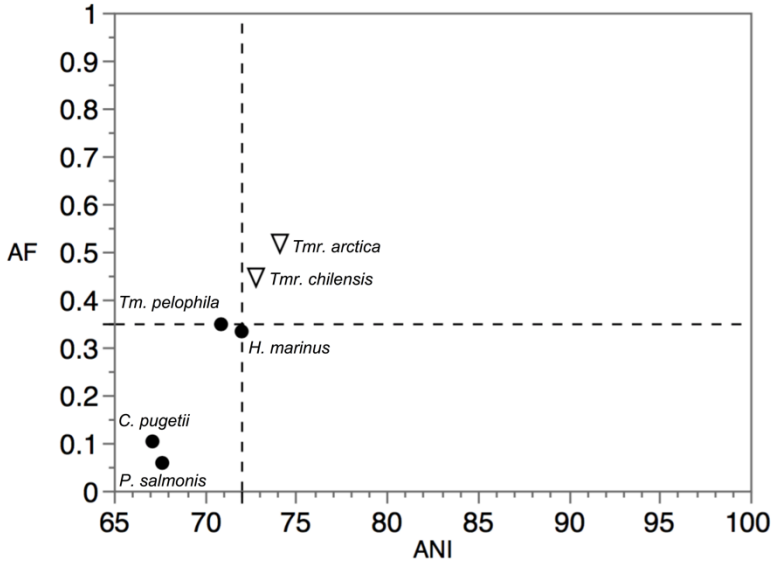


Figure S3. Pairwise genome comparisons to *Thiomicrorhabdus* sp. KP2 (> 99% pairwise identity to the 16S rRNA gene sequence of the type species *Tmr. frisia* JB-A2^T, which does not have a sequenced genome) using the new re-classification scheme by Boden *et al.* (2017). Non-type species within the genus *Thiomicrorhabdus* are shown in triangles. Type species of genera within the family *Piscirickettsiaceae* are shown in circles. The bottom-left quadrant demarcates the boundary between type and non-type species.

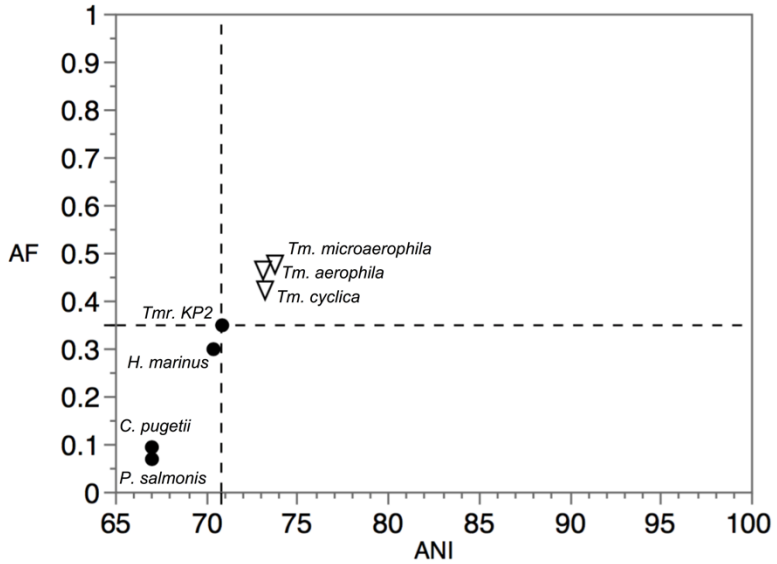


Figure S4. Pairwise genome comparisons to *Thiomicrospira pelophila* DSM 1534^T using the new re-classification scheme by Boden *et al*. (2017). Non-type species within the genus *Thiomicrospira* are shown in triangles. Type species of genera within the family

Piscirickettsiaceae are shown in circles, with the exception of *Thiomicrorhabdus frisia* JB-A2^T which does not have a sequenced genome; *Thiomicrorhabdus* sp. KP2 (> 99% pairwise identity to the 16S rRNA gene sequence) was used instead. The bottom-left quadrant demarcates the boundaries between type and non-type species.

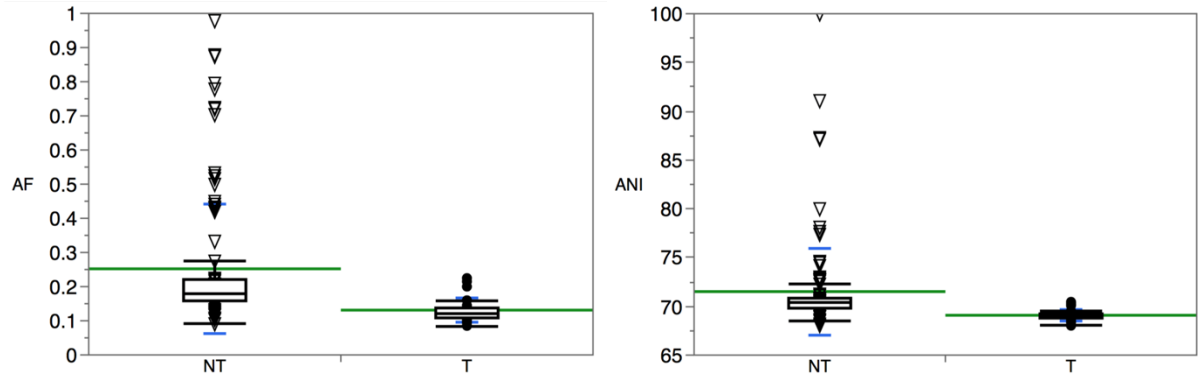


Figure S5. Boxplot diagrams of AF (left panel) and ANI (right panel) values as related to non-type species (NT; n=120) of the genus *Bacillus* and type species (T, n=26) of genera within the family *Bacillaceae*. Means are shown in green. Standard deviations are shown in blue.

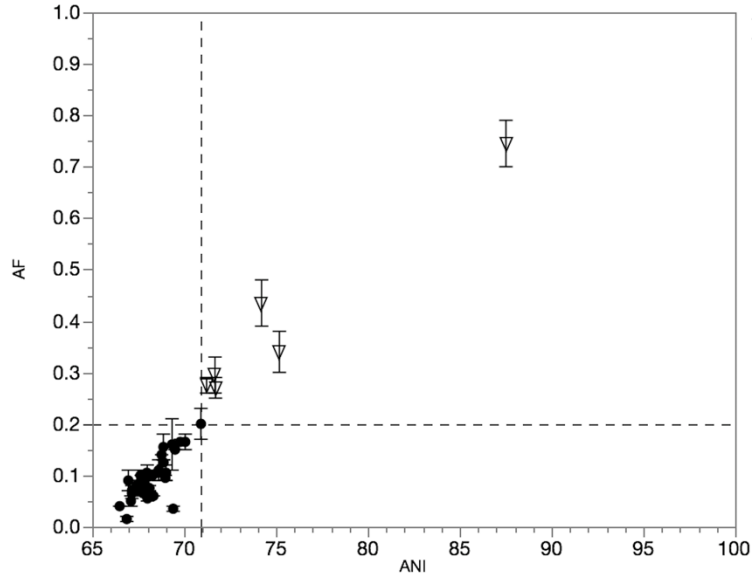


Figure S6. AF:ANI pairwise genome comparisons in the order *Lactobacillales* (Figure 1-right). Type species (n=37) of genera within the order *Lactobacillales* are shown in circles. Non-type species (n=6) of the genus *Lactococcus* are shown in triangles. The primary reference is the type species *Lactococcus lactis* subsp. *lactis* ATCC 19435^T. The bottom-left quadrant demarcates the boundary between type and non-type species. Error bars indicate one standard deviation from the mean.

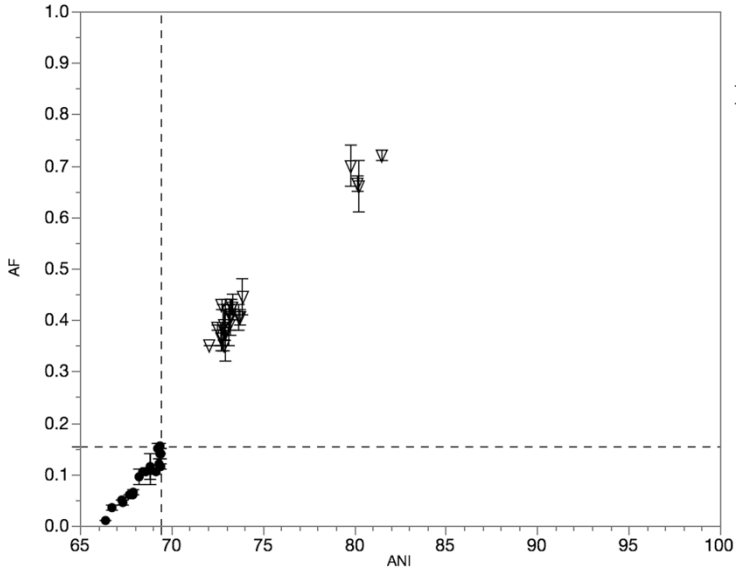


Figure S7. Pairwise genome comparisons to the type species of the genus *Shewanella*: *S. putrefaciens* JCM 20190^{T} (Figure 2A). In circles: type species (n=19) of genera within the order *Alteromonadales*. In triangles: non-type species (n=23) of the genus *Shewanella*. The bottom-left quadrant demarcates the boundary between type and non-type species. Error bars indicate one standard deviation from the mean.

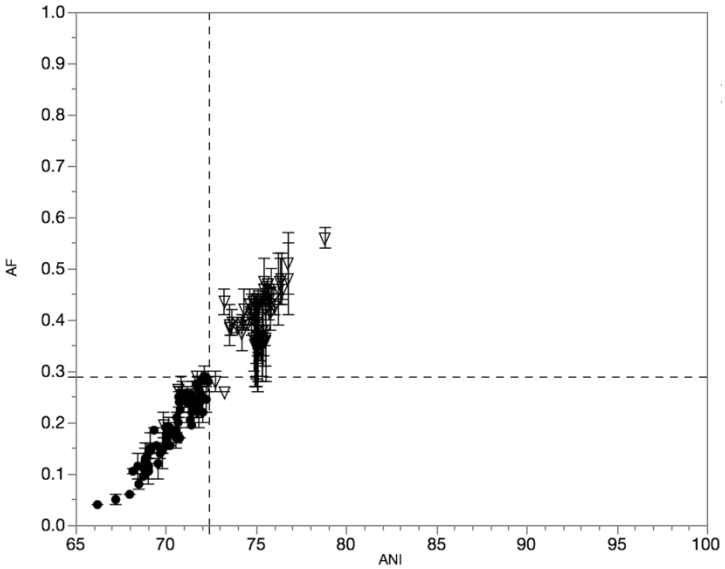


Figure S8. Pairwise genome comparisons to the type species of the genus *Flavobacterium*: *F. aquatile* LMG 4008^{T} (Figure 2B). In circles: type species (n=69) of genera within the family *Flavobacteriaceae*. In triangles: non-type species (n=83) of the genus *Flavobacterium*. The bottom-left quadrant demarcates the boundary between type and non-type species. Error bars indicate one standard deviation from the mean.

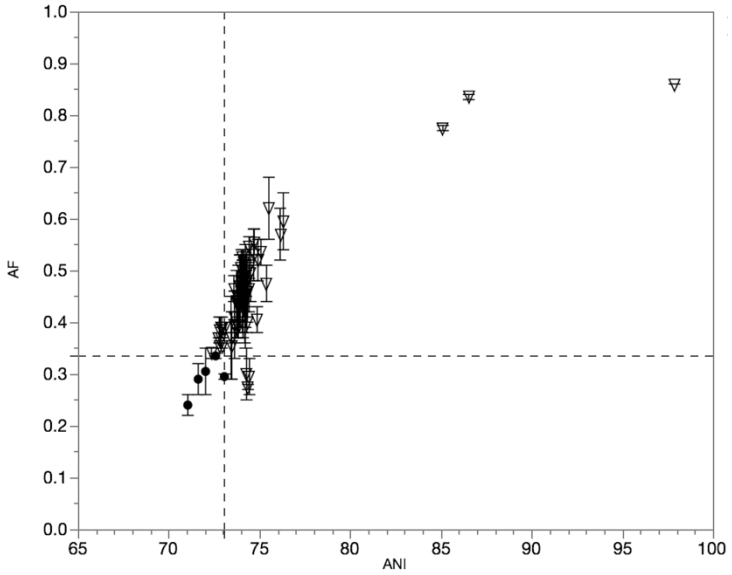


Figure S9. Pairwise genome comparisons to the type species of the genus *Vibrio*: *V. cholerae* ATCC 14035^T (Figure 2C). Circles: type species (n=5) of genera within the family *Vibrionaceae*. Triangles: non-type *Vibrio* spp. (n=65). The bottom-left quadrant demarcates the boundary between type and non-type species. Error bars indicate one standard deviation from the mean.

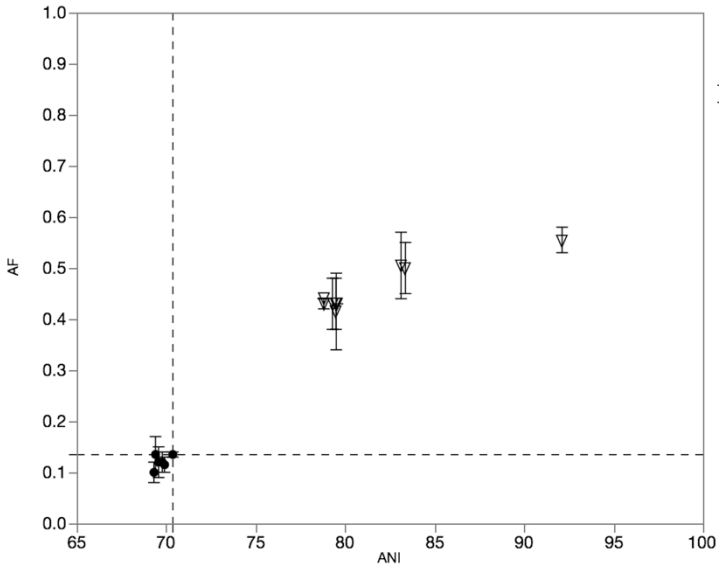
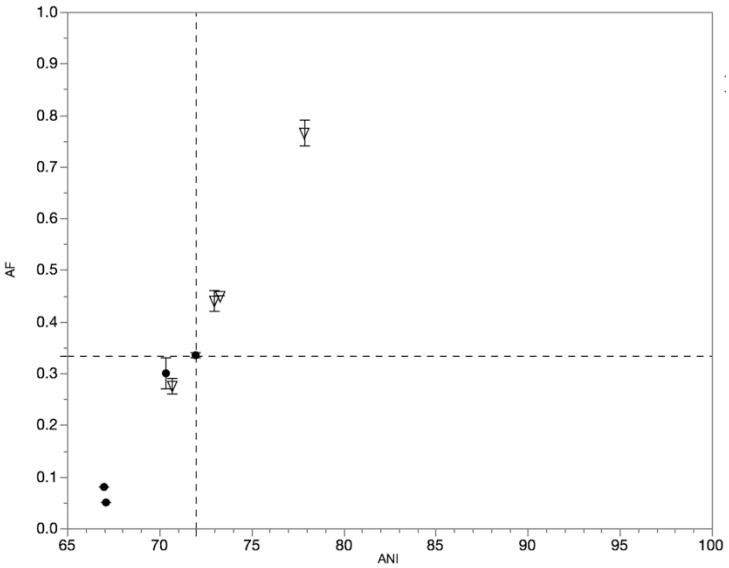
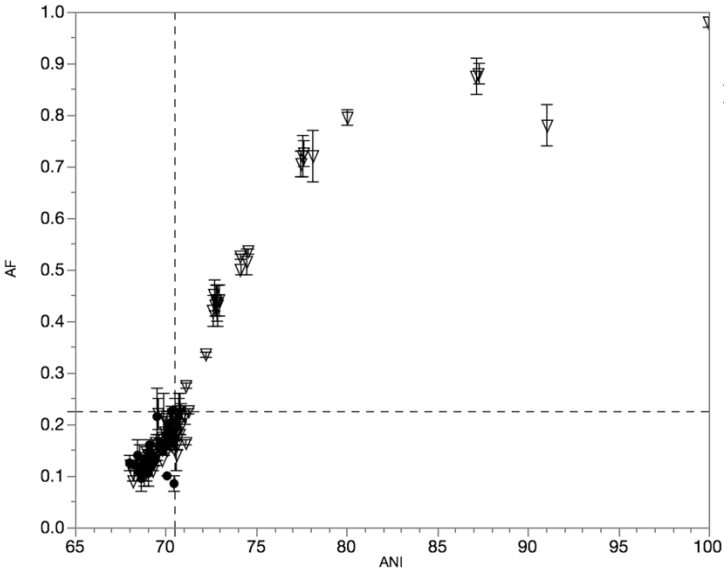


Figure S10. Pairwise genome comparisons to the type species of the genus *Methanosarcina*: *M. barkeri* JCM 10043^T (Figure 2D). Circles: type species (n=6) of genera within the family *Methanosarcinaceae*. Triangles: non-type *Methanosarcina* spp. (n=9). Error bars indicate one standard deviation from the mean.

845
846



847
848 Figure S11. Pairwise genome comparisons to *Hydrogenovibrio marinus* DSM 11271^{T} using
849 the new re-classification scheme by Boden *et al.* (2017)(Figure 3A). Non-type species within
850 the genus *Hydrogenovibrio* are shown in triangles. Type species of genera within the family
851 *Piscirickettsiaceae* are shown in circles. The bottom-left quadrant demarcates the boundary
852 between type and non-type species. Error bars indicate one standard deviation from the mean.
853
854



856
857 Figure S12. Pairwise genome comparisons to the type species of the genus *Bacillus*: *B.*
858 *subtilis* ATCC 6051^{T}. Circles: type species (n=26) of genera within the family *Bacillaceae*.
859 Triangles: non-type *Bacillus* spp. (n=110). The bottom-left quadrant demarcates the boundary
860 between type and non-type species. Error bars indicate one standard deviation from the mean.
861

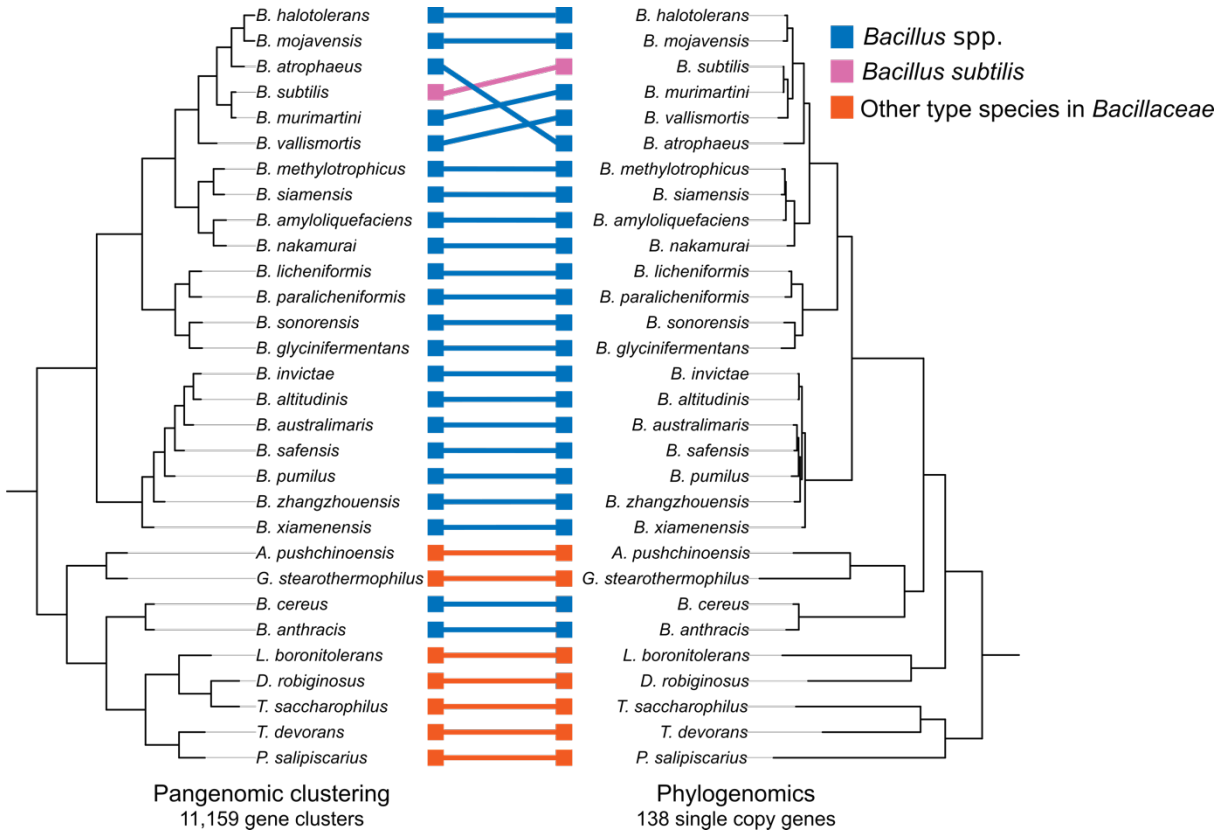
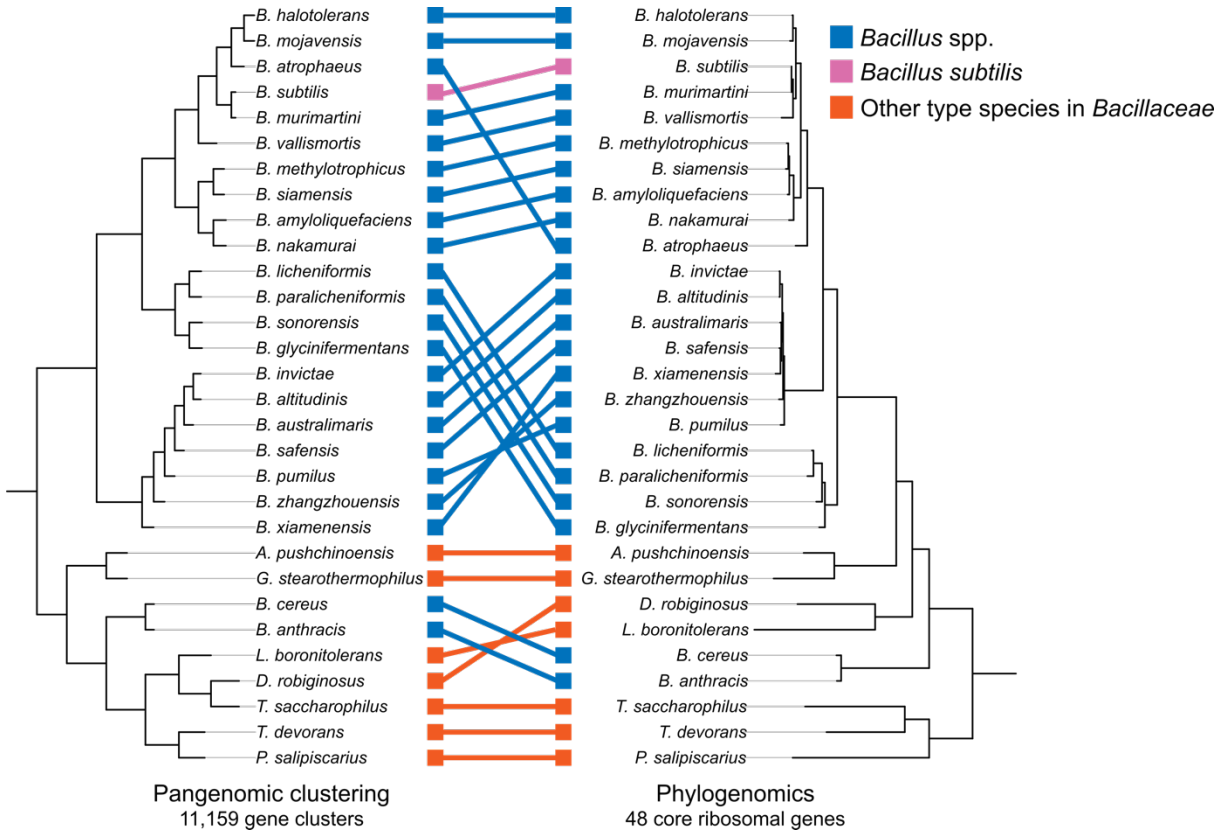


Figure S13. Genome clustering based on either pangenomic or phylogenomic analysis. All 138 single copy genes were included in the phylogenomic analysis.

887



888
889
890 Figure S14. Genome clustering based on either pangenomic or phylogenomic analysis. Only
891 the ribosomal genes were included in the phylogenomic analysis.