An ANI gap within bacterial species that advances the definitions of intra-species units.

Luis M. Rodriguez-R[1*], Roth E. Conrad[2,*], Tomeu Viver[3], Dorian J. Feistel[2], Blake G. Lindner[2], , Fanus Venter[4], Luis Orellana[5], Rudolf Amann[5], Ramon Rossello-Mora[3], and Konstantinos T. Konstantinidis[2]

[1]Department of Microbiology, and Digital Science Center (DiSC), University of Innsbruck, Innsbruck, Austria.

[2]School of Civil and Environmental Engineering, and School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA.

[3]Marine Microbiology Group, Department of Animal and Microbial Biodiversity, Mediterranean Institutes for Advanced Studies (IMEDEA, CSIC-UIB), Esporles, Spain

[4]Department of Biochemistry, Genetics and Microbiology, and Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa.

[4]Department of Molecular Ecology, Max Planck Institute for Marine Microbiology, Bremen, Germany.

Correspondence should be addressed to K.T.K. (email: kostas.konstantinidis@gatech.edu)

*equal first authors.

Keywords: ANI, strain definition, micro-diversity, epidemiology, clonal complex

Abstract

Large-scale surveys of prokaryotic communities (metagenomes) as well as isolate genomes have revealed that their diversity is predominantly organized in sequence-discrete units that may be equated to species. Specifically, genomes of the same species commonly show genome-aggregate average nucleotide identity (ANI) >95% among themselves and ANI <90% to members of other species, while genomes showing ANI 90-95% are comparatively rare. However, it remains unclear if such "discontinuities" or gaps in ANI values can be observed within species and thus, used to advance and standardize intra-species units such as strains and sequence types. By analyzing 18,123 complete isolate genomes from 330 bacterial species with at least ten genome representatives each as well as available long-read metagenomes, we show that such a discontinuity exists between 99.2-99.8% (mean 99.5%) ANI. The 99.5% ANI threshold is largely consistent with how sequence types have been defined in previous epidemiological studies but provides clusters with ~20% higher accuracy in terms of evolutionary and gene-content relatedness of the grouped genomes, while strains should be consequently defined at higher ANI values (>99.99% proposed). Collectively, our results should facilitate future micro-diversity studies across clinical or environmental settings because they provide a more natural definition of intra-species units of diversity.

Introduction

Discrete, or somewhat discrete (1), ecological, functional or evolutionary units within bacterial species have been recognized for some time. These units have been designated with various terms such as subspecies, ecotypes, clonal complexes, serotypes and strains, among several others [reviewed in (2)]. However, the application of these units has commonly been inconsistent between different taxa and studies, e.g., different marker genes and standards for each marker are used, creating challenges in communication about diversity. If diversity within species is indeed organized in discrete units and these units show similar intra-unit relatedness levels, that is the units are consistent across taxa, this information could be used to standardize unit definition and recognition. These challenges are represented well by the use of clonal complex (CC) and sequence type (ST), two terms that are commonly employed to catalogue

30 intra-species diversity. These terms have been successfully used, especially in medical
31 microbiology and epidemiological studies, to identify an outbreak caused by a specific
32 pathogenic organism (or pure isolate) or groups (complexes) of highly related organisms. A ST is
33 typically defined as a collection of genomes with no nucleotide sequence diversity (zero single-
34 nucleotide polymorphisms) in 6-7 selected genetic loci (3). These loci are typically distributed
35 across the genome to avoid co-selection evolutionary events, and represent fragments (PCR
36 amplicons) of genes shared by most members of a species; that is, core genes. While this
37 definition is pragmatic and operational, it has its own inherent limitations. Most notably,
38 different gene markers are often used for different taxa, and core genes tend to be more
39 conserved than the genome average. Thus, it remains somewhat speculative how similar (or
40 not) organisms of the same ST may be in the rest of their genome, and this may also depend, at
41 least partly, on the exact loci used in the analysis since different genes often evolve under
42 varied selective constrains. Several recent efforts have also employed all core genes in the
43 genome, providing a different (higher) resolution level compared to earlier efforts with 6-7
44 genes, which, nonetheless, makes isolate-typing results based on different sets of genes not
45 directly comparable (4, 5). On the other hand, CCs are defined as closely related STs
46 (complexes) based on phylogenetic analysis of the corresponding sequences but there are no
47 established standards on how closely related STs should be to be grouped together under the
48 same CC (1). Rather, CCs are usually defined by the clustering patterns (e.g., monophyly) of STs,
49 and after overlaying epidemiological data, that is, in an empirical way. If the intra-species
50 diversity was organized in consistent, discrete units across different taxa, this would have
51 provided for more natural and precise intra-species unit definition(s) compared to the existing
52 practice and thus, improved communication about intra-species diversity. While it has been
53 recently recognized that prokaryotic organisms may form such discrete units at the species
54 level (6), it remains unclear whether or not such consistent units across different taxa also exist
55 within species.
56
57 Specifically, culture-independent (metagenomic) studies of natural microbial
58 populations during the past decade revealed that bacteria and archaea predominantly form

59 sequence-discrete populations with intra-population genomic sequence relatedness typically
60 ranging from ~95% to ~100% ANI depending on the population considered. For example,
61 younger populations since the last population diversity sweep event show lower levels of intra-
62 population diversity. In contrast, ANI values between distinct populations are typically lower
63 than 90% (7). Intermediate identity genotypes, for example, sharing 85–95% ANI, when
64 present, are generally ecologically differentiated and scarcer in abundance, and thus should
65 probably be considered distinct species (6, 8, 9) rather than representing cultivation or other
66 sampling biases (10). Such sequence-discrete populations have been recovered from many
67 different habitats, including marine, freshwater, soils, human gut, and biofilms, and are usually
68 persistent over time and space [e.g., (11-15)] indicating that they are not ephemeral but long-
69 lived entities. Further, these sequence-discrete populations commonly harbor substantial intra-
70 population gene content diversity (i.e., they are rarely clonal) (11, 14). Therefore, these
71 populations appear to be "species-like" and may constitute important units of microbial
72 communities. Moreover, the 95% ANI threshold appears to be largely consistent with how
73 isolate genomes have been classified into (named) species in the last couple decades; that is,
74 ~97% of named species include only organisms with genomes sharing >95% ANI (16). In
75 summary, it appears that a natural gap in ANI values can be used to define prokaryotic species
76 and has been largely consistent with how species are recognized (16). In this context,
77 “discontinuity” or “gap” refers to the dearth of genome pairs showing 85–95% ANI relative to
78 counts of pairs showing ANI >95% or <85%. Whether or not a similar ANI gap exists within
79 species has not been evaluated yet.
80
81 A related term that is also commonly used to catalogue intra-species diversity is strain.
82 The “strain” represents a fundamental unit of microbial diversity that is commonly used across
83 medical or environment studies to designate the smallest distinguishable unit, presumably a
84 unit within STs. Unlike the stringent and precise definition of ST and CC, the definition of a
85 strain is more relaxed and often appears context dependent. Specifically, the concept of a strain
86 is primarily based on the notion of pure cultures (17). The Bacteriological Code defines strain as
87 the group “*of the descendants of a single isolation in pure culture*” (18). Accordingly, a strain is

88 expected to represent a genome or a collection of genomes that have no single-nucleotide or
89 gene content differences or, if such differences exist, they are expected to not encode for
90 important phenotypic differences (19). Unfortunately, this definition of a strain – while being
91 commonly used across microbiological fields – remains problematic, and often leads to
92 confusion in communication. Notably, it is not always clear when two distinct genomes or cells
93 should be considered the same or separate strains since cell ancestry information is often
94 missing, such as in environmental (culture-independent) surveys. Additionally, the isolation of
95 an organism (wild-type) in the laboratory is frequently accompanied by phenotypic changes due
96 to adaptation to laboratory conditions; yet, the wild-type and the lab-adapted cells are typically
97 considered the same strain (17), even though their observed phenotypes may not fully overlap
98 in some cases. The existence of phenotypic differences among members of the same strain
99 could be confusing because high phenotypic similarity is expected at this (the strain) level. To
100 circumvent some of these limitations, we have recently proposed a definition threshold for
101 strains at 99.99% ANI on the basis of high gene-content conservation (>99.0% of total genes,
102 typically) and thus, phenotypic relatedness among 138 isolate genomes of *Salinibacter ruber*
103 recovered from two saltern sites in Spain (20). However, testing this definition with a larger
104 collection of genomes and how it precisely relates to STs and CCs have not been performed yet.
105

106 **Results and Discussion**

107 **An ANI gap within species around 99.2-99.8%**

108 In the process of assessing cultivation biases as a possible explanation for the ANI-based
109 sequence discrete populations previously (6), we observed another discontinuity (or gap) in ANI
110 values that may be used to more reliably and systematically define the units within a species.
111 Specifically, the analysis of 18,123 complete genomes from 330 species available in NCBI's
112 Assembly database with at least ten genome representatives per species revealed a clear
113 bimodal distribution in the ANI values within named species or 95% ANI-defined groups of
114 genomes (genomospecies). That is, there is a scarcity of genome pairs showing 99.2-99.8% ANI
115 (average around 99.5% ANI) in contrast to genome pairs showing ANI >99.8% or <99.2%.
116 Specifically, among the 18,123 complete genomes in our dataset, there are 4,280,133 genome

117 pairs showing ANI >96%, which would translate to about 107,000 pairs per every 0.1 percent
118 unit of ANI if there was no bimodal distribution and the ANI values among these genome pairs
119 were evenly distributed between 96% and 100% ANI. Our analysis revealed only 235,527
120 genome pairs between 99.2% and 99.8% ANI, which is three-fold fewer data points than
121 expected by chance alone in a uniform ANI value distribution (642,000 pairs expected). No
122 other ANI range within 96-100% had such a strong bias based on our dataset. That is, a
123 pronounced gap in ANI values is observed among very closely related members of a species
124 around 99.2-99.8% ANI (Fig. 1). Importantly, this ANI gap appears to be consistent across
125 phylogenetically diverse species from a dozen of distinct bacterial phyla evaluated, including
126 gram-negative and gram-positive. About half of the 330 species evaluated appear to be
127 environmental species (as opposed to pathogen or human/animal host-associated) or species
128 of biotechnological interest, revealing no major bias based on the (presumed) ecology of the
129 organisms evaluated (Fig. S1). Further, the ANI gap does not seem to be driven by a couple or a
130 few species based on a sub-sampling of all species to the same number of genomes (n=10) (Fig.
131 S2). Instead, it represents a nearly universal property of the 330 species evaluated (see also Fig.
132 S3 for specific species examples). Therefore, it appears that another important level of genomic
133 differentiation may exist within species.
134
135 It is unlikely that this 99.2-99.8% intra-species ANI gap is due to cultivation or
136 classification biases due the reasons mentioned previously, such as that cultivation media
137 usually do not distinguish between members of the same or closely related species (i.e., similar
138 organisms) (6), and that random subsampling provided similar patterns (Fig. S2). It is also highly
139 likely that the intra-species ANI gap is even more pronounced in nature because very closely
140 related genomes (e.g., showing ANI >99.8% to each other) are often selected against for
141 genome sequencing (and thus, are likely underrepresented in our collection) based on pre-
142 screening using fingerprinting techniques (e.g., RAPD, MLST) in order to avoid sequencing of
143 redundant genomes. Further, we were not able to identify clear exceptions to this 99.2-99.8%
144 intra-species ANI gap when examining individual species with enough sequenced
145 representatives, although such exceptions likely exist. For instance, several species in our

146 collection did not have enough highly related genome representatives (showing ANI >99% to
147 each other) to assess the critical area of ANI value distribution (i.e., the 99-100% range), and
148 this could be due to the pre-screening biases mentioned above or reflect their actual natural
149 diversity patterns (see also Fig. 4, Panel D for an example of an exception within *E. coli*).
150 Further, for a few species (n < 10) such as *Listeria monocytogenes* and *Bordetella*
151 *bronchiseptica*, the intra-species ANI gap appears to exist but is shifted compared to the 99.2-
152 99.8% ANI that characterized most well-sampled species (Fig. S3; all 330 species are available
153 on the GitHub at https://github.com/rotheconrad/bacterial_strain_definition). Therefore, for
154 future studies, we suggest evaluating the ANI value distribution for the species of interest, and
155 if the data indicate so, to adjust the ANI threshold to match the gap in the observed ANI value
156 distribution. The 99.2-99.8% ANI range should represent the gap for most species based on the
157 dataset evaluated here.
158
159 **Support from a high-throughput cultivation environmental study and long-read**
160 **metagenomes**
161 Our team has recently described a collection of high-draft, isolate genomes of
162 *Salinibacter ruber* (n=162), chosen for sequencing at random from a larger collection of isolates
163 (n=257) recovered from two solar saltern sites on the Mallorca and Fuerteventura Islands
164 (Spain) (9). Solar salterns are human-controlled environments used for salt production. They
165 are operated in repeated cycles of feeding with natural saltwater, followed by water
166 evaporation due to ambient sunlight, and finally, salt precipitation. Salterns from different parts
167 of the world have been reported to harbour similar microbial communities, generally consisting
168 of two major lineages, i.e., the archaeal *Halobacteria* class and the bacterial family of
169 *Salinibacteraceae*, class *Rhodothermia* (21). Within each lineage, the species richness is
170 relatively high (22-24). *S. ruber* often makes up 5-20% of the total microbial community of
171 saltern sites, i.e., it is an abundant population *in situ*, and the growth media used in our
172 previous study have been tested and found to not bias the diversity of *S. ruber* isolates that can
173 be recovered in culture (9, 21). Notably, our analysis of the 162 genomes revealed that only
174 0.35% of the total 13,122 ANI comparisons fell between 99.6% and 99.8% vs. 1.24% expected if

175 the total ANI values higher than 99% were distributed uniformly at random between 99% and
176 100% (~4-fold reduction in datapoints; Fig. 2). Therefore, the intra-species ANI values between
177 members of the same *S. ruber* species from a single sampling site and year (e.g., Mallorca
178 Island) or two sites separated by about 2,000 km (Mallorca vs. Fuerteventura Islands) revealed
179 a remarkably similar ANI gap to that observed with the heterogenous genome collections
180 available in NCBI.
181
182 We also examined recently available long-read metagenomes from a variety of habitats
183 (25, 26) to offer a culture-independent assessment of the intra-species diversity patterns. It
184 should be noted that short-read metagenomes are not ideal for our purposes because the
185 shorter sequence fragments show a larger dispersion of identity values around the (genome-)
186 average identity value and thus, could mask (obscure) the ANI gap we observed in the analysis
187 of both NCBI complete and *S. ruber* draft isolate genomes described above. Moreover, direct
188 read mapping of short reads can only offer identity resolutions of 1 divided by the read length
189 (e.g., 1% resolution for 100 bp, or 0.4% resolution for 250 bp), making it impossible to resolve a
190 gap so close to 100% identity. Nonetheless, previous short-read metagenomic surveys that
191 recovered and compared metagenome-assembled genomes (MAGs) of the same species from
192 different samples have revealed a scarcity of MAGs that are related around 99.0-99.5% ANI
193 (15), although the resolution – and thus, the sharpness of this ANI gap – is not as great as that
194 presented here based on isolates above. The latter could be due, at least in part, to the
195 assembly process frequently merging sequences related at 97-98% nucleotide identity or higher
196 into a consensus sequence.
197
198 To avoid the possibility of the assembly obfuscating sequence similarities in the data, we
199 did not assemble the long-read sequences but made direct comparisons against each other or
200 against a reference genome or MAG recovered from the same sample, using only reads longer
201 than 10 Kbp and from both Oxford Nanopore and PacBio datasets. We also examined,
202 independently, the diversity patterns based on full sequences of the ribosomal polymerase
203 subunit B (*rpoB*) gene recovered by a subset of the long-reads to offer higher resolution, since

204 *rpoB* has been shown to represent a reliable marker that reflects well the whole-genome
205 identity (5). The analysis of *rpoB* metagenomic sequences also circumvented the effect of
206 varied degrees of sequence conservation between different genes on the signature; i.e., that
207 some genes are more or less conserved than the genome average, which could introduce noise
208 with respect to gaps in nucleotide diversity patterns. Collectively, the results showed that the
209 intraspecies ANI gap is present in most populations that were abundant enough to be
210 adequately sampled by the metagenomic library in the human gut (Fig. 3) as well as soil and
211 ocean habitats (Fig. S4). In a few cases, especially of oceanic populations, the ANI gap was not
212 observed, mostly because either the populations were too clonal to assess (e.g., no reads
213 showed nucleotide identities <99%; majority of the cases observed) or harbored extensive
214 intra-population that was – more or less – evenly distributed between 96% and 100% ANI (Fig.
215 S4, B and C). For several of the latter cases, however, the ANI gap became (more) obvious when
216 the analysis was restricted to the *rpoB* gene (Fig. S4D), revealing that the effect of varied
217 degrees of sequence conservation between different genes on (blurring) the signature was
218 significant even based on long reads. Collectively, these results showed that while exceptions to
219 the pattern exist, the predominant picture is that the signature is present in natural
220 populations, providing further support for the intra-species ANI gap revealed by the analysis of
221 the isolate genomes described above.
222
223 **Gene content diversity within the >99.8% ANI clusters**
224 Another notable observation from the data from all species comparisons is that shared
225 gene content generally decreases as ANI distance (or genomic divergence) increases within the
226 95% ANI clusters, but the decrease is biphasic. That is, shared gene content decreases quickly
227 among genome pairs sharing 99.8-100% ANI but then, the decrease is less dramatic in genome
228 pairs sharing between 96.0%-99.8% ANI. In other words, genome pairs sharing between 99.8%-
229 100% ANI (i.e., one fourth of a unit of ANI) may differ in their total gene content by up to 10%
230 (average values of genome pairs showing ~99.8% ANI) and more divergent genomes of the
231 same species (i.e., showing 96.0%<ANI<99.8%) may differ by up to 20% (average values of
232 genome pairs showing ~96.0% ANI), adding another ~10% of gene content differences for 3.75

233 additional units of ANI (vs. 0.25 unit in the 99.8-100% range). Further, the genes that differed
234 between genomes sharing >99.8% ANI are more enriched in hypothetical and mobile (e.g.,
235 prophage and transposases) functions compared to the functions that differ between more
236 divergent genomes by ~10% of the total genes in the genome ($p<0.001$, z-test; Fig. S5),
237 consistent with what was reported earlier for intra-species gene-content diversity (27).
238 Collectively, these results show that genome pairs showing ANI >99.8% are also expected to be
239 much more similar in gene-content compared to more divergent genomes of the same species.
240 Nonetheless, it is important to highlight that even very closely related genomes (showing ANI
241 >99.8%) often show substantial gene-content differences, up to about 10% of the total genes
242 based on our evaluation, albeit most of these differences are likely ephemeral and
243 metabolically/ecologically not-important genes (e.g., Fig. 5S). Hence, members of the same
244 99.8%-ANI-based intra-species unit should be expected to be overall more similar in shared
245 functional gene content and thus, phenotype relative to comparison between such intra-
246 species units, but not necessarily identical.
247
248 **Comparison to Sequence Types (or STs)**
249 We also assessed how consistent the 99.5% ANI threshold (as the midpoint of the 99.2-
250 99.8% range that corresponds to the gap) is with the assignment of genomes to other intra-
251 species units. We found that the 99.5% ANI threshold is most similar to the Sequence Type (ST),
252 the latter defined as identical sequences for 6-7 genetic loci, among the units evaluated (data
253 not shown). We primarily report on the analysis of the *E. coli* species below because it is a good
254 representative of the ANI patterns observed within other species in our dataset, the large
255 number of closed *E. coli* genomes available (n=1218), and the availability of a robust Multi-
256 Locus Sequence Typing/Analysis (MLST/MLSA) scheme (28) that has been used for at least two
257 decades to provide below-species resolution and identify outbreaks of *E. coli* pathogens. Under
258 the *E. coli* MLST scheme, genomes are assigned to the same ST based on identical sequences in
259 seven *E. coli* core genes (namely, *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA*) (28). Our
260 evaluation showed that the 99.5% ANI threshold is largely consistent with how genomes are
261 assigned to STs; that is, ~80% of ST assignments, for the four most abundant STs (n=615

262 genomes), were supported by the 99.5% ANI threshold (Fig. 4). In other words, only about 20%
263 of the genomes that were assigned to the same ST showed <99.5% ANI among themselves (high
264 recall). The existence of a clear gap around 99.5% ANI in the latter cases indicated that these
265 STs could be split in two (or more) STs for more homogenous STs in terms of overall genomic
266 relatedness (e.g., ST-10 and ST131, Fig. 4). The higher resolution provided by ANI in these cases
267 is due, at least in part, to the fact that the core genes used in current MLST schemes tend to be
268 more highly conserved, at the sequence level, than the genome average (represented by ANI).
269 These results are also consistent with our previous conclusion that the 95% ANI threshold for
270 species demarcation should be adjusted upwards if the ANI is based on a few universal or core
271 genes, as opposed to the whole-genome (16). Further, ~3.5% of the genomes with ANI >99.5%
272 were assigned to different ST, revealing even higher precision (Table S1; Suppl. Fig. S6). Similar
273 results to those reported here for *E. coli* were observed for several of the 14 most-sampled
274 species in our dataset with available MLST schemas, albeit recall was slightly better than that
275 observed for the *E. coli* dataset (average across all ST comparisons: 83.1%, stdev 0.31) while
276 precision was worse (average 91.6%, stdev 0.19).
277 The high precision (and recall) of the 99.5% ANI threshold compared to ST assignments
278 is due, at least in part, to the fact that hundreds, if not thousands (e.g., at least 3,000 for the *E.*
279 *coli* pairs), of genes are used in each ANI calculation at this level of high relatedness. Thus, the
280 ANI threshold is robust against horizontal gene transfer or other evolutionary events that could
281 affect sequence identity of one or a few loci, a known limitation of traditional MLST
282 approaches. Importantly, the 99.5% ANI threshold, or a different ANI threshold to better
283 capture intra-species diversity patterns as recommended above, is a property that emerges
284 from the data themselves as opposed to a manmade threshold (such as "identical sequences in
285 seven loci" used in MLST applications) and thus, it should capture the natural diversity patterns
286 better. Consistent with this interpretation, our preliminary results from applying the 99.5% ANI
287 threshold to a collection of *E. coli* isolate genomes collected over a period of 18 months in
288 Northern coastal Ecuador as part of the EcoZUR study (for "*E. coli* en Zonas Urbanas y Rurales")
289 (29) shows that the 99.5%-ANI-defined STs map well to local outbreaks of pathogenic *E. coli*
290 (Feistel et al., in preparation). Therefore, using the 99.5% ANI to define new or refine existing

291 STs could provide data-informed groups that encompass genomically more homogenous
292 organisms compared to the existing practice in some cases. Another important advantage of
293 the 99.5% ANI is that it can be automatically implemented and thus, does not require manual
294 curation, which is the case when establishing new ST numbers when novel (meaning, not seen
295 previously) sequences become available (3).
296
297 **ANI-based definitions for strain, sequence type (ST) and clonal complex (CC)**
298 Our evaluation shows that the 99.5% ANI threshold, or the 99.8% if the upper end of the
299 observed ANI gap is used instead, is directly comparable to how STs have been defined (e.g.,
300 Fig. 4) and thus, could be used to complement or even substitute in the fugure the requirement
301 for 6-7 identical loci that has the shortcomings explained above. We propose the midpoint
302 (99.5% ANI) as opposed to the upper value (99.8% ANI) of the gap as a more conservative
303 threshold and in order to account for the variation observed among different species with
304 respect to the ANI value range that corresponded to their own gap (e.g., Fig. S3). We also
305 suggest that the term genomovar could be used to refer to these 99.5%-ANI intra-species units,
306 instead of STs, in case that ST should maintain its original conception of 6-7 identical loci for
307 historic, pragmatic or other reasons. The term genomovar was originally used to name distinct
308 genomic groups within species that cannot be distinguished phenotypically from each other
309 and therefore, cannot be classified as distinct species based on the standard taxonomic
310 practices (30). Hence, genomovar may capture conceptually well the 99.5% ANI groups, as we
311 also proposed recently elsewhere (20).
312
313 We think that CCs should continue to be defined based on phylogenetic analysis of STs;
314 in fact, we see the 99.5% ANI threshold proposed here as a complementary, rather than
315 competitive, approach to the phylogenetics because it provides convenient means to define
316 and/or refine STs, which can then be used in phylogenetic analysis to define CCs, assess
317 evolutionary relationships between STs, etc. Related to this, it is important to note that the
318 computation of ANI is, on average, two orders of magnitude faster compared to the
319 phylogenetic placement of a genome using all core genes (16). Therefore, the ANI-based

320 approach and thresholds proposed here should provide highly efficient and practical means to
321 classify (type) large collections of genomes into STs at the whole-genome level and
322 subsequently perform phylogenetic (or other) analysis of representatives of the resulting STs.
323
324 Finally, we did not observe another ANI gap within the 99.5% ANI clusters and thus,
325 recommend the use of the term strain only for nearly-identical genomes. We recently proposed
326 to define a strain as a collection of genomes sharing ANI >99.99% based on the extent of gene-
327 content differences among the *S. ruber* genomes originating from the same saltern site (20).
328 This threshold ensures high gene content similarly, e.g., typically, >99.0% gene content is
329 shared at this ANI level based on the *S. ruber* genomes (20) and the data presented here (Fig. 1
330 and Fig. 2). Further, our previous study showed that this ANI level encompasses well the typical
331 sequencing and assembly noise observed when splitting the raw reads from a *S. ruber* isolate
332 genome project in two halves and assembling the two subsets independently for comparisons
333 of the resulting genome sequences (20). It should be noted, however, that 99.99% ANI
334 threshold represents only practical and convenient means for defining strains, and it could be
335 neglected or adjusted should key phenotypic differences distinguishing organisms sharing ANI
336 >99.99% are known/found such as antibiotic resistance or catabolic genes carried by plasmids.
337
338 **What are the underlying mechanisms for the 99.5% ANI gap?**
339 The mechanism(s) that underly the 99.5% ANI gap (or the earlier 95% ANI gap for the
340 species level) remain essentially speculative and should be the subject of future research in
341 order to further advance the mechanistic understanding of the microbial diversity patterns
342 observed in nature. Most notable is the idea that members of a population cohere together via
343 means of unbiased (random) genetic exchange which is more frequent within vs. between
344 populations or CCs (i.e., *the biological* or *sexual species concept*) (31). A competing hypothesis is
345 that several members of the species are functionally differentiated from each other either due
346 to specialization for different growth conditions or different affinities for the same energy
347 substrate and thus, selection over time for these functions purge diversity (i.e., *the ecological*
348 *species concept*) (32-34). It is intriguing to note that the ecological explanation is also consistent

349 with the notion that STs or different strains of the same species are somewhat ecologically
350 and/or functionally distinguishable from each other. Notably, given an estimated mutation rate
351 of ~$4x10^{-10}$ per nucleotide per generation (35) and between 100 to 300 generations per year
352 (36), it would take two distinct *E. coli* lineages or STs at least twenty thousand years since their
353 last common ancestor to accumulate 0.5% difference (i.e., fixed mutations) in their core genes
354 or 99.5% ANI. Therefore, there is enough time, at least theoretically, for the ecological purging
355 of diversity to take place at around the 99.5% ANI level and thus, account for the ANI patterns
356 observed herein. Intriguingly, it has been shown that the explicit inclusion of extinction events
357 in a neutral model of evolution can also result in punctuated distributions of genetic
358 differentiation, opening up a third possibility of historical contingency from stochastic events
359 (37). However, we note that while stochasticity can explain bimodal (or multimodal) distance
360 distributions, a scarcity of ANI values in the exact same range (i.e., around 99.5% ANI) would be
361 unlikely to repeatedly emerge by chance alone across many different species with distinct
362 lifestyles and evolutionary tempo, as opposed to this range varying between species. In any
363 case, the data available in support of one of these (or another) hypotheses remain sparse
364 and/or anecdotal to date, to the best of our knowledge, and the analysis presented in this study
365 did not aim to advance this issue further but rather to present a highly intriguing observation of
366 patterns of diversity that could have major practical consequences in defining strains, STs, and
367 species more broadly.
368
369 **Conclusions**
370 Regardless of what the underlying mechanisms are for the 99.5% ANI gap, the results
371 presented here show that the patterns of natural diversity among thousands of sequenced
372 genomes are consistent with a 99.5% ANI threshold that can be used to identify STs more
373 reliably and precisely compared to the current practice. Regarding the use of this threshold to
374 define (or refine) STs vs. strains, we believe that the threshold is highly appropriate, as well as it
375 matches well the intended meaning and use of STs, and thus its application to ST definition is
376 straightforward, although we recommend the alternative term of genomovar to avoid
377 confusion with historical ST definitions. Notably, recent large-scale surveys of the human gut

378 microbiome (38) have also revealed distinct intra-species units closely matching the 99.5% ANI
379 threshold and genomovar definition proposed here, even though these units were often called
380 strains previously (39). However, for the strain level, genomes showing ~0.5% or ~0.2%
381 difference in ANI (99.5% or 99.8% ANI, respectively) often show substantial sequence and gene
382 content divergence (e.g., Fig. 1) and thus, phenotypic differences. Therefore, multiple (distinct)
383 strains should be expected to be grouped together within the same 99.5% ANI cluster in such
384 cases, and strain, in general, represents a more fine-grained level of resolution than the 99.5%
385 ANI level. The results presented here based on all well-sampled bacterial species further
386 reinforced our recent proposal based on the *S. ruber* isolate genomes from the same site to use
387 99.99% ANI as the threshold to define strains (20). Collectively, we expect that the findings
388 reported here will advance the molecular toolbox for accurately delineating and following the
389 important units of diversity withing prokaryotic species and thus, would greatly facilitate future
390 epidemiological and micro-diversity studies.
391
392 **Methods**
393 Step by step methods, including how average trendlines were fit to the data, custom
394 Python code, NCBI Assembly accession numbers for selected genomes, and plots for each
395 selected species are available from:
396 https://github.com/rotheconrad/bacterial_strain_definition. Briefly, all genome sequences
397 were obtained from NCBI's RefSeq Assembly database on April 20th, 2022 and were labeled as
398 "complete" and "latest". ANI values and the shared genome fractions were directly obtained
399 from the output of FastANI version 1.32, "One to Many" mode with default settings (16).
400 Results were concatenated to create within species all vs. all output. Self matches were
401 removed, and genome pairs were filtered by minimum ANI values according to the axes of each
402 figure. Selected individual species plots (i.e., all vs. all output) of shared genome fraction vs. ANI
403 are shown in the Supplementary Material. *E. coli* genomes were assigned to sequence types
404 (ST) using the using the command-line tool mlst (https://github.com/tseemann/mlst) version
405 2.19.0 (28) with default settings.
406

Acknowledgments

This work has been supported by the US National Science Foundation (Award No 1759831 and 2129823) to KTK. LO and RA acknowledge funding by the Max Planck Society.

Code and data availability

All code and data details are available from https://github.com/rotheconrad/bacterial_strain_definition.

Competing interests

The authors declare no competing interests.

Author contact list

Luis M. Rodriguez-R: lmrodriguezr@gmail.com

Roth E. Conrad: rotheconrad@gatech.edu

Dorian J. Feistel: dfeistel3@gatech.edu

Tomeu Viver: tviver@imedea.uib-csic.es

Fanus Venter: fanus.venter@fabi.up.ac.za

Luis Orellana: lorellanar@gmail.com

Rudolph Amann: ramann@mpi-bremen.de

Ramon Rossello-Mora: ramon@imedea.uib-csic.es

Konstantinos T. Konstantinidis: kostas.konstantinidis@gatech.edu

FIGURES

Figure 1. ANI vs. shared gene content for the 17,283 complete genomes used in this study. Each datapoint represents a comparison between a pair of genomes. FastANI (16) was used to generate ANI values between the genomes of a pair (x-axis) and their shared genome fraction (y-axis). The shared genome fraction was calculated by dividing the number of bidirectional fragment mappings over the total query fragments determined by FastANI. Only a single set of values is reported per pair, the one that used the longer genome as the reference (and the reverse comparison was omitted). Note that only datapoints representing genome pairs sharing ANI >95% are shown (n=4346079 datapoints), and that panel B is a zoomed-in version of panel A (n=2676181 datapoints). The main scatter plot is shaded by the density of points using the Datashader package in Python with Matplotlib. The trendline was calculated using linearGAM from pyGAM and includes the 95% confidence interval. The marginal plots outside the two axes show histograms for the density of datapoints of each axis. Note the low-density region in the ANI value distribution around 99.2-99.8% (Panel A), which becomes more obvious when zooming in to the 98-100% ANI range (Panel B).

Figure 1, panel A.

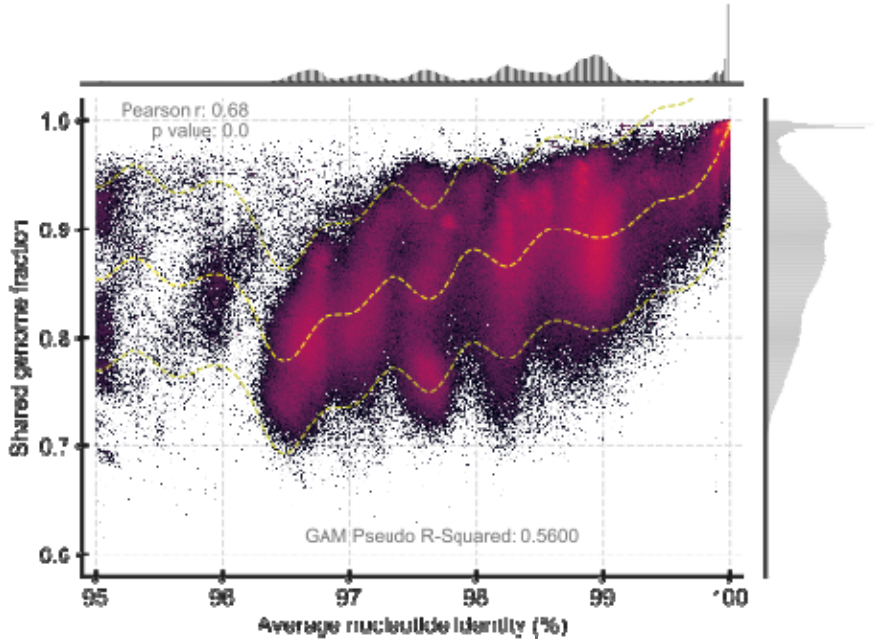


Figure 1, panel B.

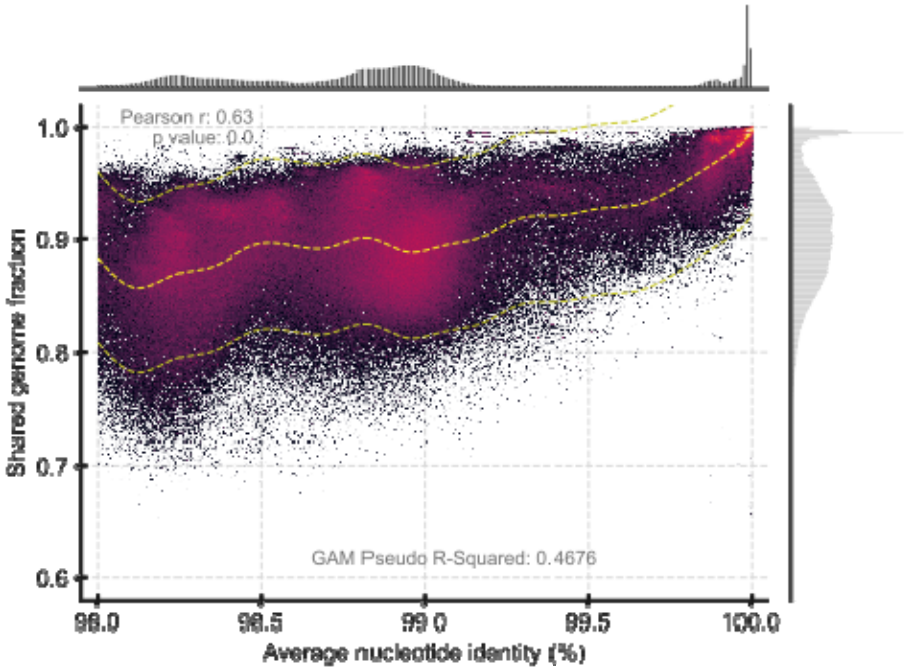


Figure 2. ANI vs. shared gene content for *Salinibacter ruber* isolate genomes recovered for the same site. Fig. 2 is identical to Fig. 1B except that the underlying data represent 162 *S. ruber* isolate genomes (n= 3847 datapoints with >98% ANI) recovered from two saltern sites on the Mallorca and Fuerteventura Islands (Spain) as previously described (9). Note the lack of genome pairs showing ANI values between 99.2% and 99.8% and the sharp increase in gene-content differences in pairs of genomes showing ANI <99.8%, which echoed the results obtained with NCBI genomes (Fig. 1). Similar results were obtained when the genomes from each site were analyzed independently (data not shown).

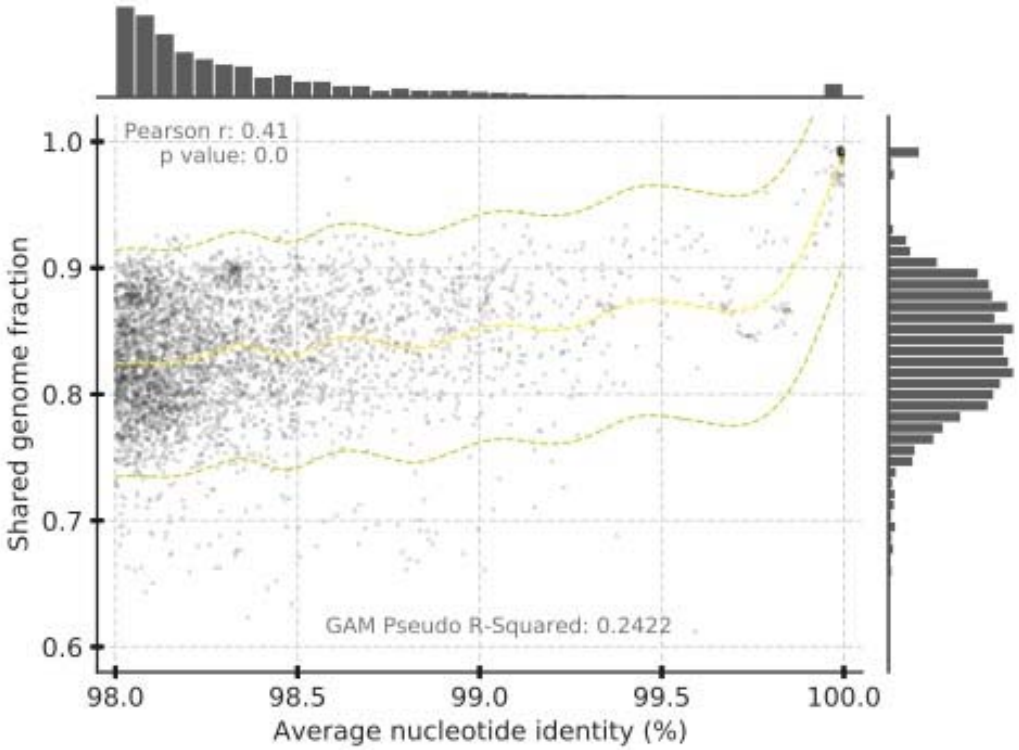


Figure 3. The intra-species ANI gap is present in natural populations recovered by long read metagenomic sequencing of human fecal samples. The underlying data are Oxford Nanopore long read metagenomes of human fecal samples (109 samples, about 136Gbp, in total, after read quality trimming) described previously (26). The graphs show the nucleotide (nt) identity distribution of individual long reads against the MAGs recovered by the same study. All reads were quality-trimmed with Filtlong (v0.2.1) for minimum read lengths of 1kb and average quality scores of >90% across 1kb sliding windows. Reads were mapped non-competitively with minimap2 (v2.24) allowing for spliced alignments (-x splice). Only alignments sharing at least 20% sequence overlap with a MAG were reported. Similar distributions were obtained when reads were simply mapped all versus all (data not shown). Top histograms show all reads sharing nt identities above 90%; bottom histograms show the subset of these reads that share nt identities above 98%. Panel A summarizes the mapping of all long reads against all MAGs (n=4,497). Panel B shows mapping of all long reads against a single MAG taxonomically identified as *Phocaeicola vulgatus*. Note the sparsity of reads mapping between 99.4-99.9% nt identity for panel A and 98.8-99.3% for panel B.

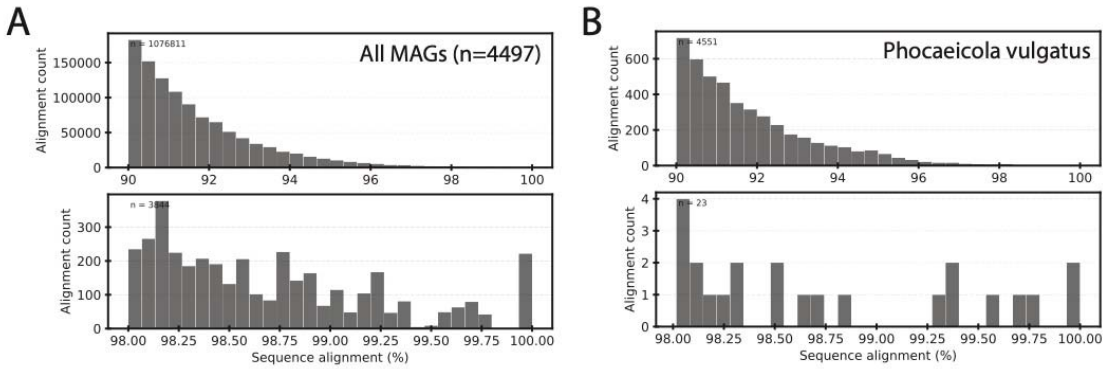
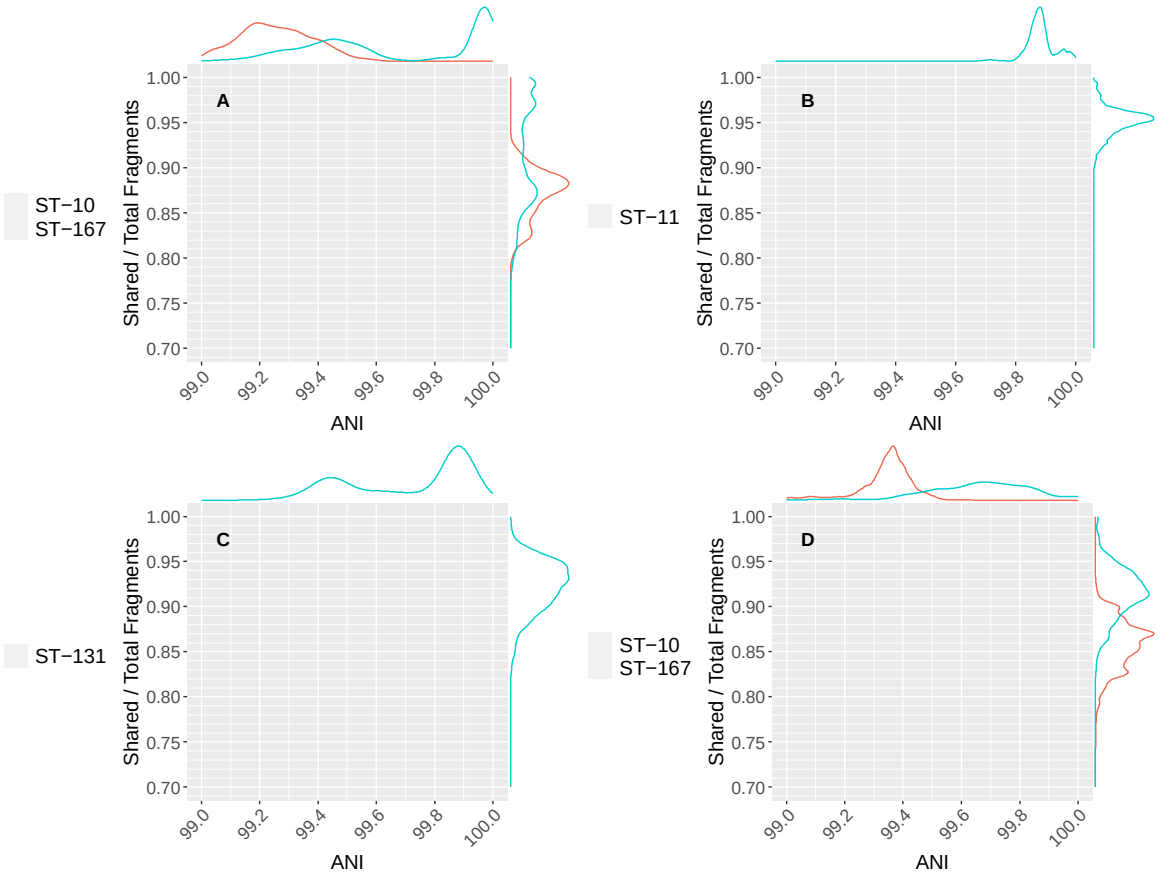


Figure 4. Comparison of the 99.5% ANI threshold to available Sequence Types (STs) of *E. coli*. All *E. coli* genomes were assigned to a ST using the tool mlst version 2.19.0 (28). The four panels show the four most abundant STs based on the number of genomes assigned to them (see Table S1 and Fig. S6 for underlying data and F1 statistic, respectively). Each datapoint is a comparison between two genomes, similar to those shown in Fig. 1; the marginal plots show the kernel density estimate of datapoints for each axis. Datapoints are cyan if both genomes in the pair were assigned to the same, reference ST; red datapoints represent pairs for which one of the genomes in the pair is assigned to a closely related, yet distinct ST than the reference ST. Note that for ST-11, recall and precision of 99.5% ANI vs. ST designation is perfect because there are no genomes, and thus STs, that are closely related to ST-11, which is also consistent with a pronounced ANI gap at 99.5% for ST-11, and the substantial overlap in terms of ANI values between the closely related ST-10 and ST167 (low recall). ST-131 and ST-10 appear to harbor too much genomic diversity and could be split in more than one ST based on the 99.5% ANI criterion (and the bimodal ANI value distribution around 99.7% ANI). ST-167 represents a possible deviation from the main pattern, meaning no clear ANI gap is obvious for this ST around 99.5% but a gap may be obvious at lower ANI values, around 99.0%-99.2%. A couple more similar examples to ST-167, out of about 330 cases examined, were noted (Fig. S3 and https://github.com/rotheconrad/bacterial_strain_definition), indicating that exceptions to the main pattern are generally infrequent.

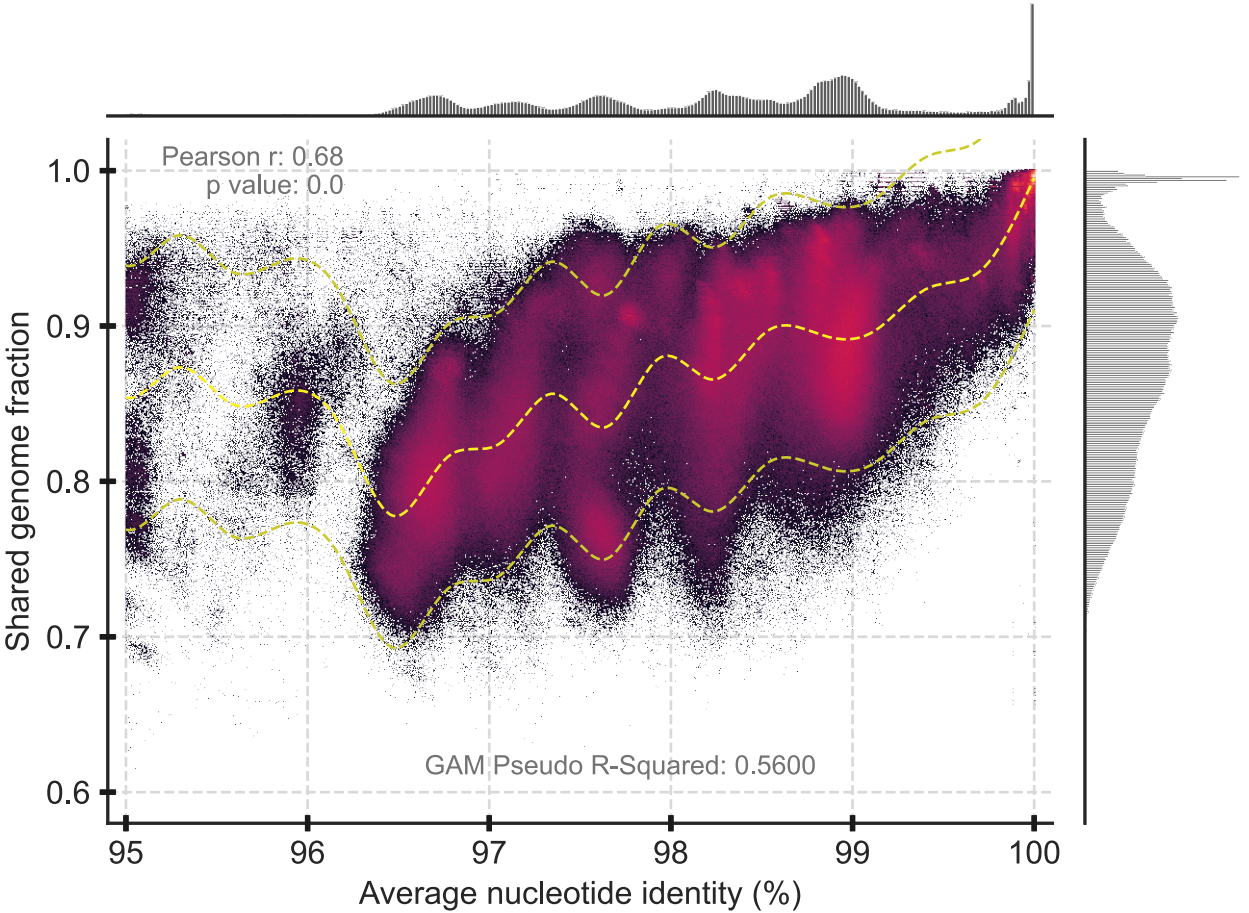

A
B
C
D
ST−10
ST−167
ST−11
ST−131
ST−10
ST−167
Shared / Total Fragments
ANI
1.00
0.95
0.90
0.85
0.80
0.75
0.70
99.0
99.2
99.4
99.6
99.8
100.0

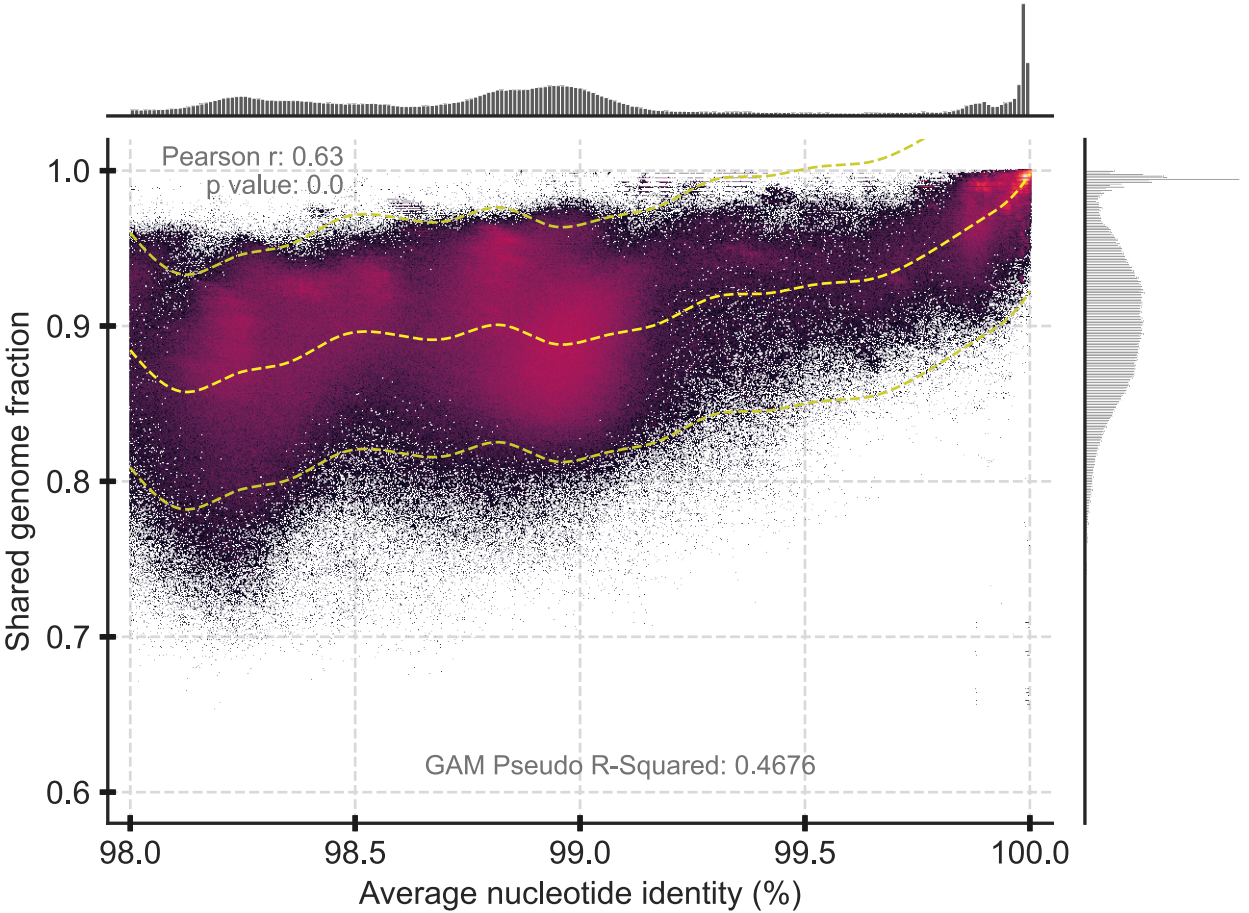
References cited.

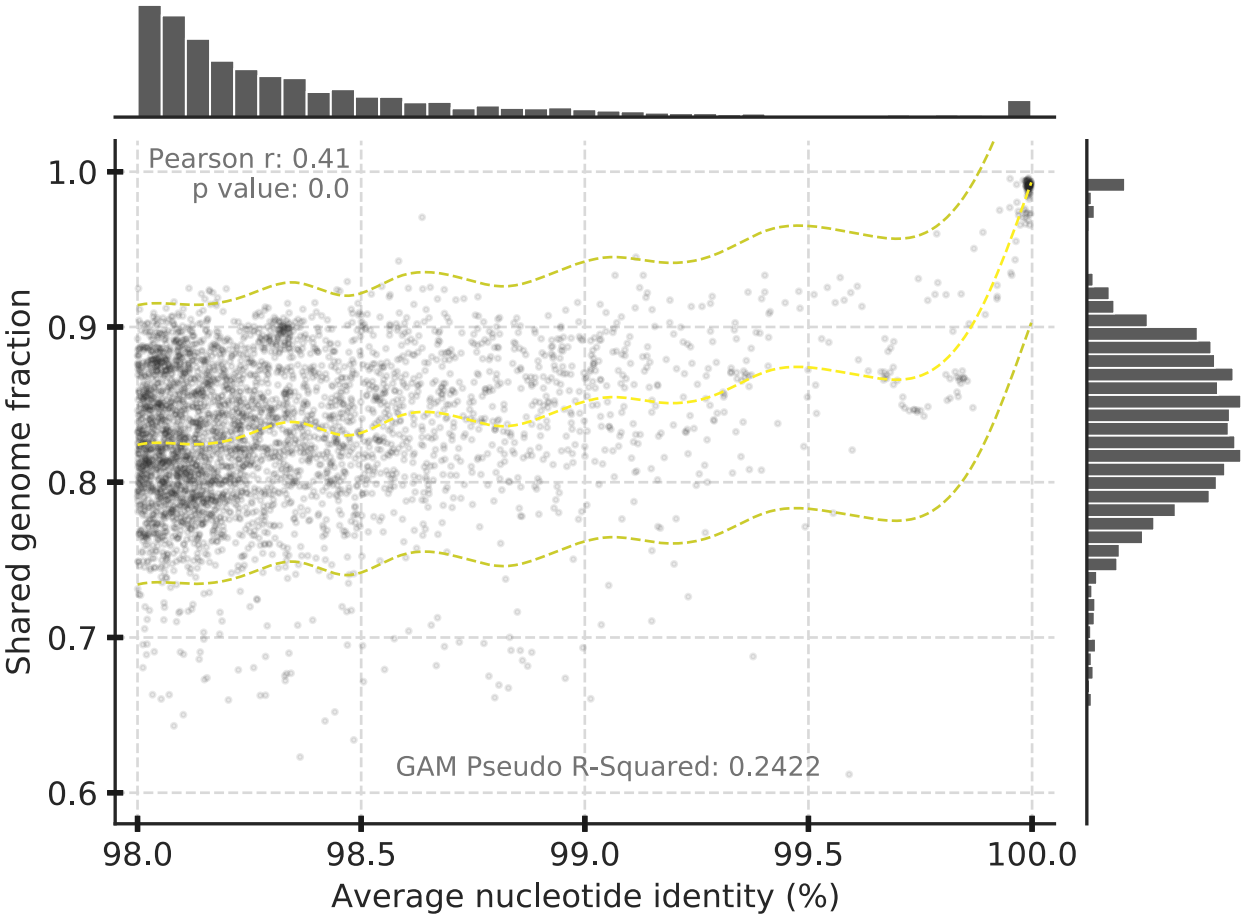
1. W. P. Hanage, C. Fraser, B. G. Spratt, Fuzzy species among recombinogenic bacteria. *BMC Biol* **3**, 6 (2005).
2. R. Rossello-Mora, R. Amann, Past and future species definitions for Bacteria and Archaea. *Systematic and applied microbiology* **38**, 209-216 (2015).
3. M. C. Maiden *et al.*, Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95**, 3140-3145 (1998).
4. D. A. Baltrus, K. Dougherty, S. M. Beckstrom-Sternberg, J. S. Beckstrom-Sternberg, J. T. Foster, Incongruence between multi-locus sequence analysis (MLSA) and whole-genome-based phylogenies: Pseudomonas syringae pathovar pisi as a cautionary tale. *Mol Plant Pathol* **15**, 461-465 (2014).
5. K. T. Konstantinidis, A. Ramette, J. M. Tiedje, Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. *Appl Environ Microbiol* **72**, 7286-7293 (2006).
6. R. L. Rodriguez, C. Jain, R. E. Conrad, S. Aluru, K. T. Konstantinidis, Reply to: "Re-evaluating the evidence for a universal genetic boundary among microbial species". *Nat Commun* **12**, 4060 (2021).
7. A. Caro-Quintero, K. T. Konstantinidis, Bacterial species may exist, metagenomics reveal. *Environ Microbiol* **14**, 347-355 (2012).
8. T. Viver *et al.*, Distinct ecotypes within a natural haloarchaeal population enable adaptation to changing environmental conditions without causing population sweeps. *ISME J* 10.1038/s41396-020-00842-5 (2020).
9. R. E. Conrad *et al.*, Toward quantifying the adaptive role of bacterial pangenomes during environmental perturbations. *ISME J* **16**, 1222-1234 (2022).
10. C. S. Murray, Y. Gao, M. Wu, Re-evaluating the evidence for a universal genetic boundary among microbial species. *Nat Commun* **12**, 4059 (2021).
11. K. T. Konstantinidis, E. F. DeLong, Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J* **2**, 1052-1065 (2008).
12. M. L. Bendall *et al.*, Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J* **10**, 1589-1601 (2016).
13. E. R. Johnston *et al.*, Metagenomics Reveals Pervasive Bacterial Populations and Reduced Community Diversity across the Alaska Tundra Ecosystem. *Front Microbiol* **7**, 579 (2016).
14. A. Meziti *et al.*, Quantifying the changes in genetic diversity within sequence-discrete bacterial populations across a spatial and temporal riverine gradient. *ISME J* **13**, 767-779 (2019).
15. M. R. Olm *et al.*, Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems* **5** (2020).
16. C. Jain, R. L. Rodriguez, A. M. Phillippy, K. T. Konstantinidis, S. Aluru, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114 (2018).

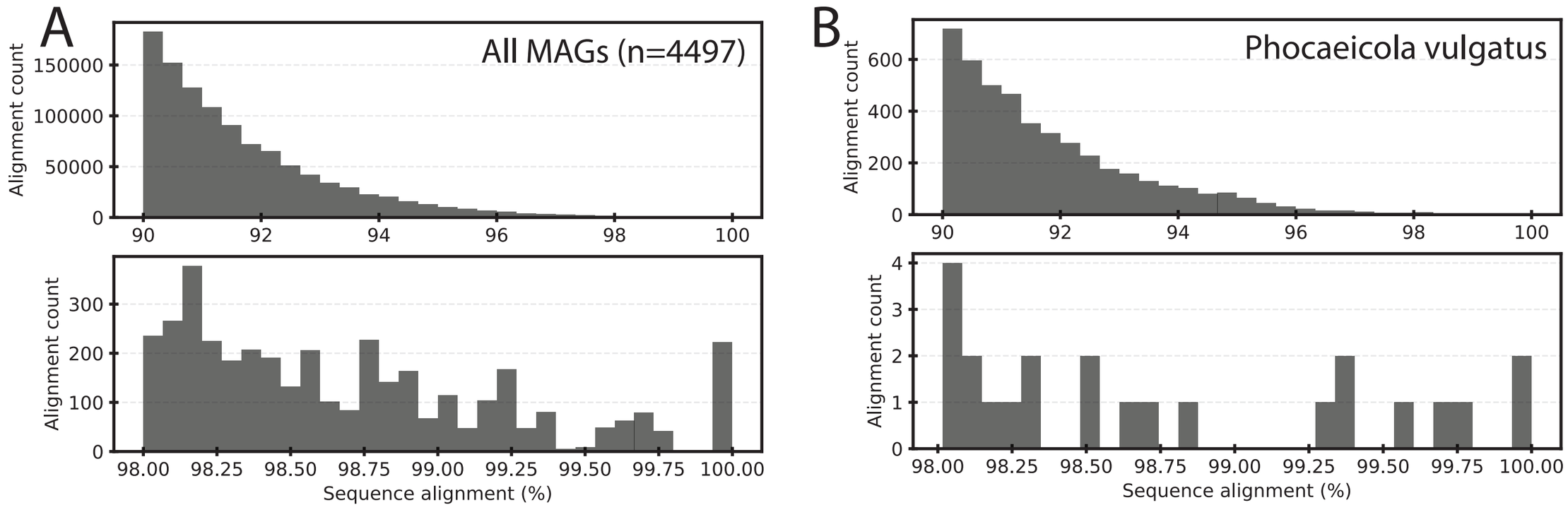
17. L. Dijkshoorn, B. M. Ursing, J. B. Ursing, Strain, clone and species: comments on three basic concepts of bacteriology. *J Med Microbiol* **49**, 397-401 (2000).
18. C. T. Parker, B. J. Tindall, G. M. Garrity, International Code of Nomenclature of Prokaryotes. *Int J Syst Evol Microbiol* 10.1099/ijsem.0.000778 (2015).
19. F. C. Tenover *et al.*, Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *J Clin Microbiol* **33**, 2233-2239 (1995).
20. T. Viver, Conrad, R. E., Rodriguez-R, L. M., Ramirez, A. S., Venter, S. N., Rocha-Cardenas, J., Segura, M. L., Amann, A., Konstantinidis, K. T., and Rossello-Mora, R., Towards estimating the number of strains that make up a natural bacterial population. *bioRxiv* https://doi.org/10.1101/2023.02.20.529252 (2023).
21. J. Anton, R. Rossello-Mora, F. Rodriguez-Valera, R. Amann, Extremely halophilic bacteria in crystallizer ponds from solar salterns. *Appl Environ Microbiol* **66**, 3052-3057 (2000).
22. T. Viver *et al.*, Predominance of deterministic microbial community dynamics in salterns exposed to different light intensities. *Environ Microbiol* **21**, 4300-4315 (2019).
23. E. O. Casamayor *et al.*, Changes in archaeal, bacterial and eukaryal assemblages along a salinity gradient by comparison of genetic fingerprinting methods in a multipond solar saltern. *Environ Microbiol* **4**, 338-348 (2002).
24. M. Gomariz *et al.*, From community approaches to single-cell genomics: the discovery of ubiquitous hyperhalophilic Bacteroidetes generalists. *ISME J* **9**, 16-31 (2015).
25. K. I. V. Chandni S., Meunier C. L., Rick J., Wiltshire K. H., Steinke N., Vidal-Melgosa S., Hehemann J-H., Huettel B., Schweder T., Fuchs B. M., AmannR. I., and Teeling H., Grazers affect the composition of dissolved storage glycans and thereby bacterioplankton composition during a biphasic North Sea spring algae bloom. *BioRxiv* https://doi.org/10.1101/2022.09.22.509014, Preprint (2022).
26. J. S. Gounot *et al.*, Genome-centric analysis of short and long read metagenomes reveals uncharacterized microbiome diversity in Southeast Asians. *Nat Commun* **13**, 6044 (2022).
27. K. T. Konstantinidis, A. Ramette, J. M. Tiedje, The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* **361**, 1929-1940 (2006).
28. T. Wirth *et al.*, Sex and virulence in Escherichia coli: an evolutionary perspective. *Mol Microbiol* **60**, 1136-1151 (2006).
29. A. Pena-Gonzalez *et al.*, Metagenomic Signatures of Gut Infections Caused by Different Escherichia coli Pathotypes. *Appl Environ Microbiol* **85** (2019).
30. J. B. Ursing, R. A. Rossello-Mora, Garcia-Valdes E., J. Lalucat, Taxonomic note: a pragmatic approach to the nomenclature of phenotypically similar genomic groups. *Int J Syst Evol Microbiol.* **45**, 604-604 (1995).
31. J. J. Power *et al.*, Adaptive evolution of hybrid bacteria by horizontal gene transfer. *Proc Natl Acad Sci U S A* **118** (2021).
32. B. J. Shapiro *et al.*, Population genomics of early events in the ecological differentiation of bacteria. *Science* **336**, 48-51 (2012).
33. C. Fraser, W. P. Hanage, B. G. Spratt, Recombination and the nature of bacterial speciation. *Science* **315**, 476-480 (2007).

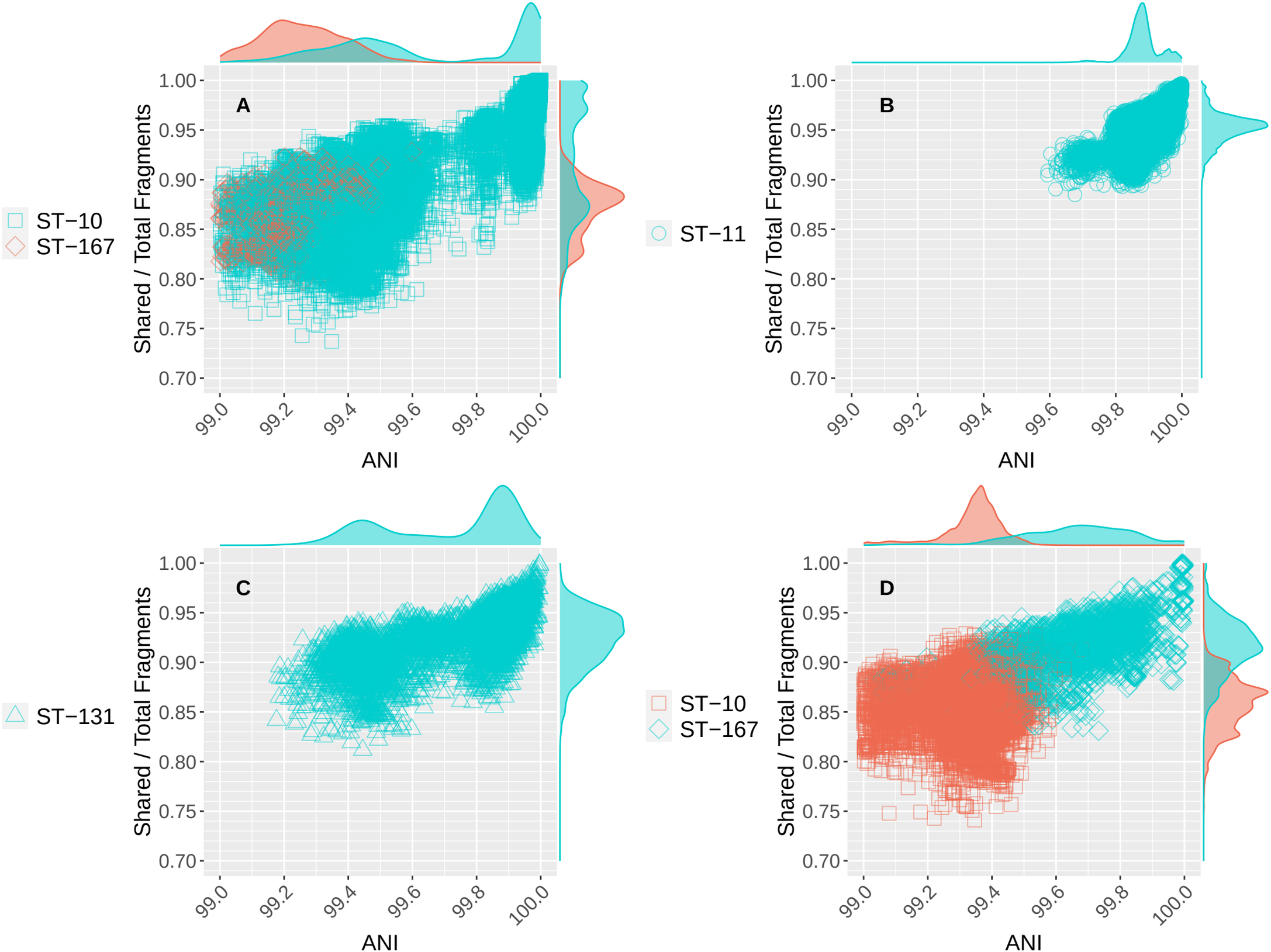
34. G. W. Tyson *et al.*, Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43 (2004).
35. J. W. Drake, B. Charlesworth, D. Charlesworth, J. F. Crow, Rates of spontaneous mutation. *Genetics* **148**, 1667-1686 (1998).
36. R. J. Gibbons, , and B. Kapsimalis Estimates of the overall rate of growth of the intestinal microflora of hamsters, guinea pigs, and mice. *Journal of Bacteriology* **93**, 510–512 (1967).
37. T. J. Straub, O. Zhaxybayeva, A null model for microbial diversification. *Proc Natl Acad Sci U S A* **114**, E5414-E5423 (2017).
38. M. Valles-Colomer *et al.*, The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* **614**, 125-135 (2023).
39. Y. Yan, L. H. Nguyen, E. A. Franzosa, C. Huttenhower, Strain-level epidemiology of microbial communities and the human microbiome. *Genome Med* **12**, 71 (2020).
40. D. Hyatt *et al.*, Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* **11**, 119 (2010).
41. M. Steinegger, J. Soding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028 (2017).
42. C. P. Cantalapiedra, A. Hernandez-Plaza, I. Letunic, P. Bork, J. Huerta-Cepas, eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* **38**, 5825-5829 (2021).


Pearson r: 0.68
p value: 0.0
GAM Pseudo R-Squared: 0.5600
Shared genome fraction
1.0
0.9
0.8
0.7
0.6
95
96
97
98
99
100
Average nucleotide identity (%)


Pearson r: 0.63
p value: 0.0
GAM Pseudo R-Squared: 0.4676
Shared genome fraction
1.0
0.9
0.8
0.7
0.6
98.0
98.5
99.0
99.5
100.0
Average nucleotide identity (%)


Pearson r: 0.41
p value: 0.0
GAM Pseudo R-Squared: 0.2422
Shared genome fraction
1.0
0.9
0.8
0.7
0.6
98.0
98.5
99.0
99.5
100.0
Average nucleotide identity (%)


A
All MAGs (n=4497)
Alignment count
150000
100000
50000
0
90
92
94
96
98
100
Alignment count
300
200
100
0
98.00
98.25
98.50
98.75
99.00
99.25
99.50
99.75
100.00
Sequence alignment (%)
B
Phocaeicola vulgatus
Alignment count
600
400
200
0
90
92
94
96
98
100
Alignment count
4
3
2
1
0
98.00
98.25
98.50
98.75
99.00
99.25
99.50
99.75
100.00
Sequence alignment (%)


A
B
C
D
ST−10
ST−167
ST−11
ST−131
Shared / Total Fragments
ANI
1.00
0.95
0.90
0.85
0.80
0.75
0.70
99.0
99.2
99.4
99.6
99.8
100.0