AMERICAN SOCIETY FOR MICROBIOLOGY | mBio

Check for updates

| Microbial Genetics | Research Article

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] Stephanus N. Venter,[4] Luis H. Orellana,[5] Rudolf Amann,[5] Ramon Rossello-Mora,[3] Konstantinos T. Konstantinidis[2]

AUTHOR AFFILIATIONS See affiliation list on p. 12.

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. By analyzing 18,123 complete isolate genomes from 330 bacterial species with at least 10 genome representatives each and available long-read metagenomes, we show that another discontinuity exists between 99.2% and 99.8% (midpoint 99.5%) ANI in most of these species. 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.

IMPORTANCE Bacterial strains and clonal complexes are two cornerstone concepts for microbiology that remain loosely defined, which confuses communication and research. Here we identify a natural gap in genome sequence comparisons among isolate genomes of all well-sequenced species that has gone unnoticed so far and could be used to more accurately and precisely define these and related concepts compared to current methods. These findings advance the molecular toolbox for accurately delineating and following the important units of diversity within prokaryotic species and thus should greatly facilitate future epidemiological and micro-diversity studies across clinical and environmental settings.

KEYWORDS ANI, strain definition, micro-diversity, epidemiology, clonal complex

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 reference (2)]. However, the application of these units has commonly been inconsistent between different taxa and studies, for example, 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

Editor Igor B. Jouline, The Ohio State University, Columbus, Ohio, USA

Address correspondence to Konstantinos T. Konstantinidis, kostas.konstantinidis@gatech.edu.

Luis M. Rodriguez-R and Roth E. Conrad contributed equally to this article. Author order was based on seniority.

The authors declare no conflict of interest.

See the funding table on p. 13.

Received 5 October 2023
Accepted 3 November 2023
Published 12 December 2023

Copyright © 2023 Rodriguez-R et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Downloaded from https://journals.asm.org/journal/mbio on 07 February 2024 by 194.95.6.38.

the use of clonal complex (CC) and sequence type (ST), two terms that are commonly employed to catalog intra-species diversity. These terms have been successfully used, especially in medical microbiology and epidemiological studies, to identify an outbreak caused by a specific pathogenic organism (or pure isolate) or groups (complexes) of highly related organisms. A ST is typically defined as a collection of genomes with no nucleotide sequence diversity (zero single-nucleotide polymorphisms) in 6–7 selected genetic loci (3, 4). These loci are typically distributed across the genome to avoid co-selection evolutionary events and represent fragments (PCR amplicons) of genes shared by most members of a species; that is, core genes. While this definition is pragmatic and operational, it has its own inherent limitations. Most notably, different gene markers are often used for different taxa, and core genes tend to be more conserved than the genome average. Thus, it remains somewhat speculative how similar (or not) organisms of the same ST may be in the rest of their genome, and this may also depend, at least partly, on the exact loci used in the analysis since different genes often evolve under varied selective constraints. Several recent efforts have also employed all core genes in the genome, providing a different (higher) resolution level compared to earlier efforts with 6–7 genes, which, nonetheless, makes isolate-typing results based on different sets of genes not directly comparable (4, 5). On the other hand, CCs are defined as closely related STs (complexes) based on phylogenetic analysis of the corresponding sequences but there are no established standards on how closely related STs should be to be grouped under the same CC (1, 3). Rather, CCs are usually defined by the clustering patterns (e.g., monophyly) of STs, and after overlaying epidemiological data (e.g., which STs are associated with related disease outbreaks). That is, CCs are largely defined as *ad hoc*. If the intra-species diversity is organized in consistent, discrete units across different taxa, this would provide for more natural and precise intra-species unit definition(s) compared to the existing practice and thus improved communication about intra-species diversity. While it has been recently recognized that prokaryotic organisms may form such discrete units at the species level (6), it remains unclear whether or not such consistent units across different taxa also exist within species.

A related concept that is also commonly used to catalog intra-species diversity is strain. The "strain" represents a fundamental unit of microbial diversity that is commonly used across medical or environmental studies to designate the smallest distinguishable taxonomic unit (7). Strain is presumably a unit within STs and should be distinguished from "clone" as the latter refers to identical genomes, and thus a strain may contain multiple clones. Unlike the stringent and precise definition of STs, the definition of a strain is more relaxed and often appears context dependent. Specifically, the concept of a strain is primarily based on the notion of pure cultures (7), although the concept is also commonly used in culture-independent studies (8). The International Code of Nomenclature of Prokaryotes Code defines strain as the group "*of the descendants of a single isolation in pure culture*" (9). Accordingly, a strain is expected to represent a genome or a collection of genomes that have no single-nucleotide or gene content differences or, if such differences exist, they are expected to not encode for important phenotypic differences (10). Unfortunately, this definition of a strain—while being commonly used across microbiological fields—is sometimes challenging to apply, which could lead to confusion in communication. Notably, it is not always clear when two distinct genomes or cells should be considered the same or separate strains since cell ancestry information is often missing, such as in environmental (culture-independent) surveys. In addition, the isolation of an organism (wild type) in the laboratory is frequently accompanied by phenotypic changes due to adaptation to laboratory conditions. Yet, the wild-type and the laboratory-adapted cells are often considered the same strain (7), although their observed phenotypes may not fully overlap, with the probable exception that the changes involve a key phenotype of interest, in which cases the organisms might be recognized as different strains. The existence of phenotypic differences among members of the same strain could be confusing because high phenotypic similarity is expected at this (the strain) level. To circumvent some of these limitations, we have recently

Downloaded from https://journals.asm.org/journal/mbio on 07 February 2024 by 194.95.6.38.

proposed a definition threshold for strain at 99.99% ANI based on the high gene-content similarity among genomes related at this level (>99.0% of total genes, typically) and thus, phenotypic relatedness. This study was based on comparisons among 162 coexisting *Salinibacter ruber* isolate genomes recovered from two saltern sites in Spain (11). Testing this definition with a larger collection of genomes and how it precisely relates to STs and CCs has not been performed yet.

Culture-independent (metagenomic) studies of natural microbial populations during the past decade revealed that bacteria and archaea predominantly form sequence-discrete populations with intra-population genomic sequence relatedness typically ranging from ~95% to ~100% ANI depending on the population considered. For example, younger populations since the last population diversity sweep event or older populations with frequent intra-population recombination show lower levels of intra-population diversity. By contrast, ANI values between distinct populations are typically lower than 90% (12). Intermediate identity genotypes, for example, sharing 85%–95% ANI, when present, are generally ecologically differentiated and scarcer in abundance, and thus should probably be considered distinct species (6, 13, 14) rather than representing cultivation or other sampling biases (15). Such sequence-discrete populations have been recovered from many different habitats, including marine, freshwater, soils, human gut, and biofilms, and are usually persistent over time and space [e.g., references (16–20)] indicating that they are not ephemeral but long-lived entities. Furthermore, these sequence-discrete populations commonly harbor substantial intra-population gene content diversity (i.e., they are rarely clonal) (16, 19). Therefore, these populations appear to be "species-like" and may constitute important units of microbial communities. Moreover, the 95% ANI threshold appears to be largely consistent with how isolate genomes have been classified into (named) species in the last couple of decades; that is, ~97% of named species include only organisms with genomes sharing >95% ANI (21). In summary, it appears that a natural gap in ANI values can be used to define prokaryotic species and has been largely consistent with how species are recognized (21). In this context, "discontinuity" or "gap" refers to the dearth of genome pairs showing 85%–95% ANI relative to counts of pairs showing ANI > 95% or <85%. Whether or not a similar ANI gap exists within species has not been evaluated yet.

RESULTS AND DISCUSSION

An ANI gap within species around 99.2%–99.8%

In the process of assessing cultivation biases as a possible explanation for the ANI-based sequence discrete populations previously (6), we observed another discontinuity (or gap) in ANI values that may be used to more reliably and systematically define the units within a species. Specifically, the analysis of 18,123 complete genomes from 330 species available in NCBI's Assembly database with at least 10 genome representatives per species revealed a clear bimodal distribution in the ANI values within named species or 95% ANI-defined groups of genomes (genomospecies). That is, there is a scarcity of genome pairs showing 99.2%–99.8% ANI (average around 99.5% ANI) in contrast to genome pairs showing ANI > 99.8% or <99.2%. Specifically, among the 18,123 complete genomes in our data set, there are 4,280,133 genome pairs showing ANI > 96%, which would translate to about 107,000 pairs per every 0.1% unit of ANI if there was no bimodal distribution and the ANI values among these genome pairs were evenly distributed between 96% and 100% ANI (i.e., a uniform distribution). Our analysis revealed only 235,527 genome pairs between 99.2% and 99.8% ANI, which is threefold fewer data points than expected by chance alone in a uniform ANI value distribution (642,000 pairs expected). As it is also obvious in Fig. 1, many species showed more than one such ANI gap; that is, a multi-modal distribution of ANI values was observed for several species. For instance, the well-known phylogroups within the model bacterial species *Escherichia coli* (22) corresponded to an ANI gap of around 97%–98% for this species. However, the 99.2%–99.8% ANI was the only gap that was consistently observed across many species. In particular, by employing the kernel density estimate or Hartigan's dip test

Downloaded from https://journals.asm.org/journal/mbio on 07 February 2024 by 194.95.6.38.

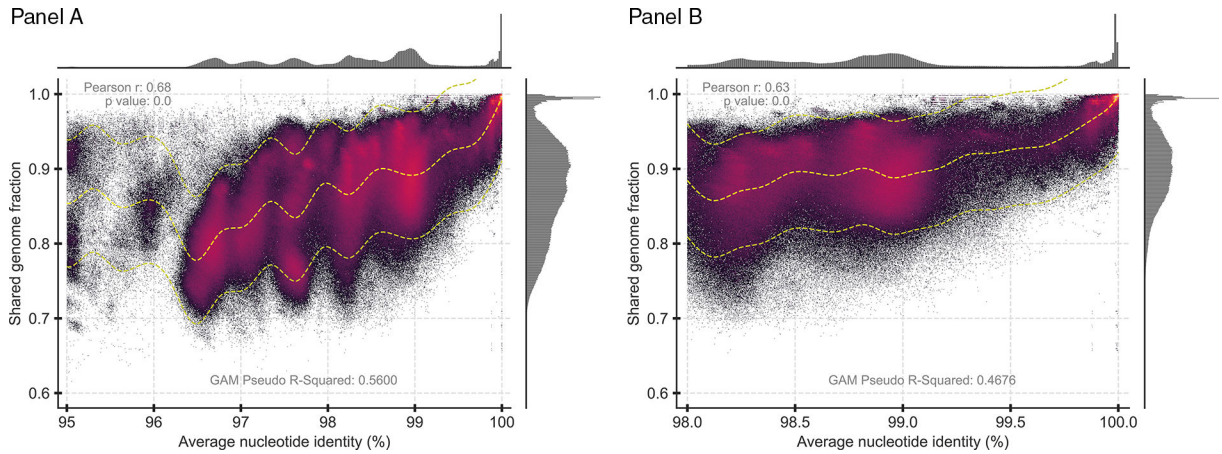

FIG 1 ANI vs shared gene content for the 17,283 complete genomes used in this study. Each datapoint represents a comparison between a pair of genomes. FastANI (21) was used to generate ANI values between the genomes of a pair (x-axis) and their shared genome fraction (y-axis). The shared genome fraction 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 = 4,346,079 datapoints), and that Panel B is a zoomed-in version of Panel A (n = 2,676,181 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 on 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).

to statistically identify significant peaks and gaps in the ANI data distribution of each species (without predefining an ANI location for the gaps), we observed that about 70% of the 330 species showed a clear ANI gap around 99.2%–99.8% in addition to or to the exclusions of other gaps. Indeed, the most likely number (mode) for any intra-specific ANI gaps was 99.5% using several independent techniques for gap detection (see also Supplementary Material and Results). Among the remaining species, half were too clonal to call (meaning, they only encompass genomes showing >99.5% ANI). The other half, including species such as *Listeria monocytogenes* and *Bordetella bronchiseptica*, exhibited an intra-species ANI gap that appears to be shifted compared to the 99.2%–99.8% ANI gap that characterized most species. Figure S1 shows a few representative examples of the actual ANI value distributions observed and all 330 species are available at https://github.com/rotheconrad/bacterial_strain_definition; Fig. S2 shows a frequency-normalized heatmap representation of the ANI value distribution observed for each species. It is important to note that the pattern for the former, clonal species, could be viewed as a much larger ANI gap (e.g., the gap extends below 99.5% ANI) compared to the predominant 99.2%–99.8% ANI gap, which could be driven—for example—by a relatively recent diversity sweep event. Therefore, a pronounced gap in ANI values is observed among very closely related members of a species around 99.2%–99.8% ANI (Fig. 1) albeit with a few exceptions to this pattern (Fig. S2). Importantly, this ANI gap appears to be consistent across phylogenetically diverse species from a dozen of distinct bacterial phyla evaluated, including Gram negative and Gram positive. About half of the 330 species evaluated appear to be environmental species (as opposed to a pathogen or human/animal host associated) or species of biotechnological interest, revealing no major bias based on the (presumed) ecology of the organisms evaluated, albeit the gap of the environmental species appears slightly less pronounced relative to that of the clinical species, on average (Fig. S3; further discussed below). Furthermore, the ANI gap does not seem to be driven by a couple or a few species based on a sub-sampling of all species to the same number of genomes (n = 10) (Fig. S4). Instead, it represents a property of the majority of the 330 species evaluated (see also Fig. S2 for specific species

Downloaded from https://journals.asm.org/journal/mbio on 07 February 2024 by 194.95.6.38.

examples). Therefore, it appears that an important level of genomic differentiation may exist within species.

It is unlikely that this 99.2%–99.8% intra-species ANI gap is due to cultivation or classification biases due to the reasons mentioned previously, such as that cultivation media usually do not distinguish between members of the same or closely related species (i.e., similar organisms) (6) and that random subsampling provided similar patterns (Fig. S4). It is also likely that the intra-species ANI gap is even more pronounced in nature because very closely related genomes (e.g., showing ANI >99.8% to each other) are often selected against for genome sequencing (and thus are likely underrepresented in our collection) based on pre-screening using fingerprinting techniques (e.g., RAPD, MLST) to avoid sequencing of redundant genomes. It is conceivable that this bias might have been stronger for environmental vs clinical species, and thus accounts for the difference noted above between these two groups of genomes, although alternative explanations such as the use of less discriminating methods for pre-screening environmental isolates cannot be excluded at this point. Furthermore, we were able to identify only a few clear exceptions to this 99.2%–99.8% intra-species ANI gap when examining individual species with enough sequenced representatives. For a few species (n < 10), such as the genomes assigned to the closely related ST-10 and ST-167 of *E. coli* (Fig. 4, Panel D) or *Riemerella anatipestifer* (Fig. S2, Panel D), we did not observe a clear gap in the 99%–100% range, and this could be due to the isolation and pre-screening biases mentioned above or reflect their actual natural diversity patterns. However, these were the exceptions rather than the rule of the species examined (e.g., most highly sampled STs of *E. coli* do show the gap; Fig. 4 and S2). Therefore, for future studies, we suggest evaluating the ANI value distribution for the species of interest, and if the data indicate so, to adjust the ANI threshold to match the gap in the observed ANI value distribution. The 99.2%–99.8% ANI range should represent the gap for most species based on the data set evaluated here, and thus a useful reference point.

Support from a high-throughput cultivation environmental study and long-read metagenomes

Our team has recently described a collection of high-draft, isolate genomes of *Salinibacter ruber* (n = 162), chosen for sequencing at random from a larger collection of isolates (n = 257) recovered from two solar saltern sites on the Mallorca and Fuerteventura Islands (Spain) (14). Solar salterns are human-controlled environments used for salt production. They are operated in repeated cycles of feeding with natural saltwater, followed by water evaporation due to ambient sunlight, and finally, salt precipitation. Salterns from different parts of the world have been reported to harbor similar microbial communities, generally consisting of two major lineages, that is, the archaeal *Halobacteria* class and the bacterial family *Salinibacteraceae*, class *Rhodothermia* (23). Within each lineage, the species richness is relatively high (24–26). *Sal. ruber* often makes up 5%–20% of the total microbial community of saltern sites, that is, it is an abundant population *in situ*, and the growth media used in our previous study have been tested and found to not bias the diversity of *Sal. ruber* isolates that can be recovered in culture (14, 23). Notably, our analysis of the 162 genomes revealed that only 0.35% of the total 13,122 ANI comparisons fell between 99.6% and 99.8% vs 1.24% expected if the total ANI values higher than 99% were distributed uniformly at random between 99% and 100% (~4-fold reduction in datapoints; Fig. 2). Therefore, the intra-species ANI values between members of the same *Sal. ruber* species from a single sampling site and year (e.g., Mallorca Island) or two sites separated by about 2,000 km (Mallorca vs Fuerteventura Islands) revealed a remarkably similar ANI gap to that observed with the heterogeneous genome collections available in NCBI or the clinical species subset.

We also examined recently available long-read metagenomes from a variety of habitats (27, 28) to offer a culture-independent assessment of the intra-species diversity patterns. It should be noted that short-read metagenomes are not ideal for our purposes because the shorter sequence fragments show a larger dispersion of identity values

Downloaded from https://journals.asm.org/journal/mbio on 07 February 2024 by 194.95.6.38.

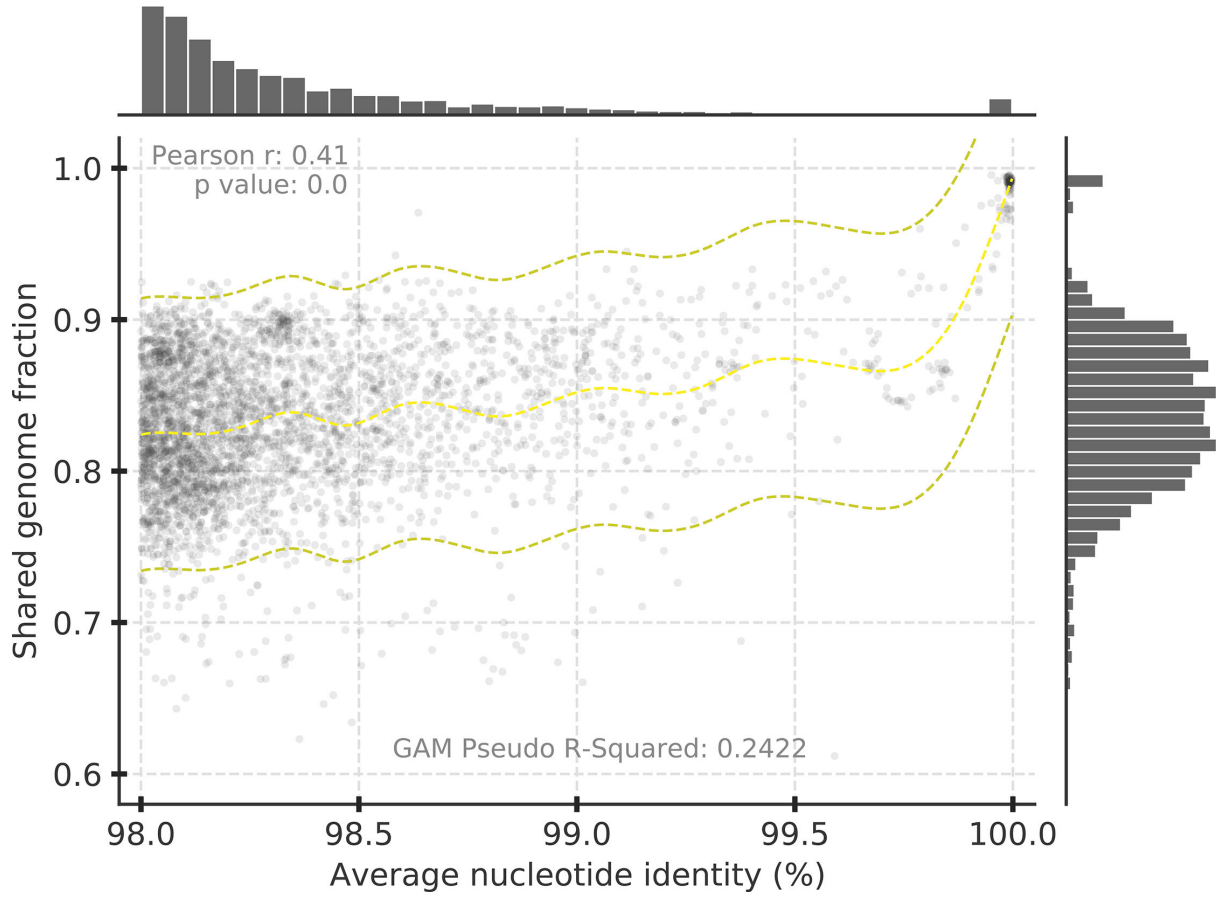

FIG 2 ANI vs shared gene content for *Salinibacter ruber* isolate genomes recovered for the same site. Figure 2 is identical to Fig. 1B except that the underlying data represent 162 *Sal. ruber* isolate genomes (*n* = 3847 datapoints with >98% ANI) recovered from two saltern sites on the Mallorca and Fuerteventura Islands (Spain) as previously described (14). 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).

around the (genome-) average identity value and thus could mask (obscure) the ANI gap we observed in the analysis of both NCBI complete and *Sal. ruber* draft isolate genomes described above. Moreover, direct read mapping of short reads can only offer identity resolutions of 1 divided by the read length (e.g., 1% resolution for 100 bp because one nucleotide mismatch decreases the identity from 100% to 99/100 = 99% or 0.4% resolution for 250 bp), making it impossible to resolve a gap so close to 100% identity. Nonetheless, previous short-read metagenomic surveys that recovered and compared metagenome-assembled genomes (MAGs) of the same species from different samples have revealed a scarcity of MAGs that are related around 99.0%–99.5% ANI (20), although the resolution—and thus, the sharpness of this ANI gap—is not as pronounced as that presented above based on isolates. The latter could be due, at least in part, to the assembly process frequently merging sequences related at 97%–98% nucleotide identity or higher into a consensus sequence based on the most commonly used settings for the assembly software (29).

To avoid the possibility of the assembly obfuscating sequence similarities in the data, we did not assemble the long-read sequences but made direct comparisons against each other or a reference genome or MAG recovered from the same sample, using only reads longer than 10 Kbp and from both Oxford Nanopore and PacBio data sets. We also examined, independently, the diversity patterns based on full sequences of the RNA polymerase subunit B (*rpoB*) gene recovered by a subset of the long-reads to offer higher resolution, since *rpoB* has been shown to represent a reliable marker that reflects well the

Downloaded from https://journals.asm.org/journal/mbio on 07 February 2024 by 194.95.6.38.

whole-genome identity (5). The analysis of *rpoB* metagenomic sequences also circumvented the effect of varied degrees of sequence conservation between different genes on the signature; that is that some genes are more or less conserved than the genome average, and individual (long) reads typically carry different sets of genes, which could therefore introduce noise with respect to gaps in nucleotide diversity patterns (note that the effect of individual genes on sequence identity patterns or ANI value is negligible when whole-genome sequences are used instead). Collectively, the results showed that the intra-species ANI gap is present in most populations that were abundant enough to be adequately sampled (e.g., showing 5–10× coverage or more) by the metagenomic library in the human gut (Fig. 3) as well as soil and ocean habitats (Fig. S5). In a few cases, especially of oceanic populations, the ANI gap was not observed, mostly because either the populations were too clonal to assess (e.g., no reads showed nucleotide identities < 99%; the majority of the cases observed) or harbored extensive intra-population that was—more or less—evenly distributed between 96% and 100% ANI (Fig. S5B and C). For several of the latter cases, however, the ANI gap became (more) obvious when the analysis was restricted to the *rpoB* gene (Fig. S5D), revealing that the effect of varied degrees of sequence conservation between different genes on (blurring) the signature was significant even based on long reads. Collectively, these results showed that while exceptions to the pattern exist, the predominant picture is that the signature is present in natural populations, providing further support for the intra-species ANI gap revealed by the analysis of the isolate genomes described above.

Gene content diversity within the >99.8% ANI clusters

Another notable observation from the data from all species comparisons is that shared gene content generally decreases as ANI distance (or genomic divergence) increases within the 95% ANI clusters, but the decrease is biphasic. That is, shared gene content decreases quickly among genome pairs sharing 99.8%–100% ANI but then, the decrease is less dramatic in genome pairs sharing between 96.0% and 99.8% ANI. In other words, genome pairs sharing between 99.8% and 100% ANI (i.e., one-fourth of a unit of ANI) may not share up to 10% of their total genes in the genome (average values of genome pairs showing ~99.8% ANI) and more divergent genomes of the same species (i.e.,

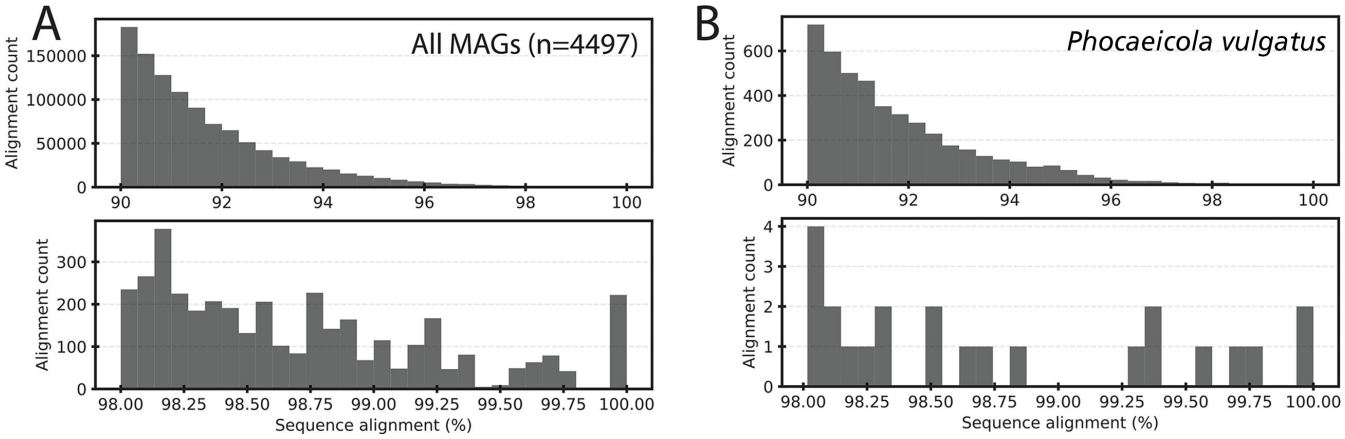

FIG 3 The intra-species ANI gap is present in natural populations recovered by long-read metagenomic sequencing of human fecal samples. The underlying data are Oxford Nanopore long read metagenomes of human fecal samples (109 samples, about 136Gbp, in total, after read quality trimming) described previously (28). 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 1 kb and average quality scores of >90% across 1 kb 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 vs 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 the mapping of all long reads against a single MAG taxonomically identified as *Phocaeicola vulgatus*. Note the sparsity of reads mapping between 99.4% and 99.9% nt identity for Panel A and 98.8%–99.3% for Panel B.

Downloaded from https://journals.asm.org/journal/mbio on 07 February 2024 by 194.95.6.38.

showing 96.0% < ANI < 99.8%) may not share up to 20% of their genes (average values of genome pairs showing ~96.0% ANI), adding another ~10% of gene content differences for 3.75 additional units of ANI (vs 0.25 unit in the 99.8%–100% range; Fig. 1, y-axis). Furthermore, the genes that differed (not shared) between genomes showing >99.8% ANI are more enriched in hypothetical and mobile (e.g., prophage and transposases) functions compared to the functions that differ between more divergent genomes by ~10% of the total genes in the genome (P < 0.001, z-test; Fig. S6), consistent with what was reported earlier for intra-species gene-content diversity (30). Collectively, these results show that genome pairs showing ANI > 99.8% are also expected to be much more similar in gene content compared to more divergent genomes of the same species. Nonetheless, it is important to highlight that even very closely related genomes (showing ANI > 99.8%) often show substantial gene-content differences, up to about 10% of the total genes based on our evaluation, albeit most of these differences are likely ephemeral and metabolically/ecologically not-important genes (e.g., Fig. S6). Hence, members of the same 99.8% ANI-based intra-species unit should be expected to be overall more similar in shared functional gene content and thus phenotype relative to comparison between such intra-species units but not necessarily identical.

Comparison to Sequence Types

We also assessed how consistent the 99.5% ANI threshold (as the midpoint of the 99.2%–99.8% range that corresponds to the gap) is with the assignment of genomes to other intra-species units. We found that the 99.5% ANI threshold is most similar to the Sequence Type (ST), the latter defined as identical sequences (or alleles) for 6–7 genetic loci, among the units evaluated (data not shown). We primarily report on the analysis of the *E. coli* species below because it is a good representative of the ANI patterns observed within other species in our data set, the large number of closed *E. coli* genomes available (n = 2072), and the availability of a robust Multi-Locus Sequence Typing/Analysis (MLST/MLSA) scheme (31) that has been used for at least two decades to provide below-species resolution and identify outbreaks of *E. coli* pathogens. Under the *E. coli* MLST scheme, genomes are assigned to the same ST based on identical sequences in seven *E. coli* core genes (namely, *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA*) (31). Our evaluation showed that the 99.5% ANI threshold is largely consistent with how genomes are assigned to STs; that is, ~79.5% of ST assignments, for the four most abundant STs (n = 615 genomes), were supported by the 99.5% ANI threshold (Fig. 4). In other words, only about 20% of the genomes that were assigned to the same ST showed <99.5% ANI among themselves (low false positives or high precision). The existence of a clear gap around 99.5% ANI in the latter cases indicated that these STs could be split into two (or more) STs for more homogenous STs in terms of overall genomic relatedness (e.g., ST-10 and ST131, Fig. 4). The higher resolution provided by ANI in these cases is due, at least in part, to the fact that the core genes used in current MLST schemes tend to be more highly conserved, at the sequence level, than the genome average (represented by ANI). These results are also consistent with our previous conclusion that the 95% ANI threshold for species demarcation should be adjusted upward if the ANI is based on a few universal or core genes, as opposed to the whole genome (21). Furthermore, 3.5% of the genomes with ANI >99.5% were assigned to different STs, revealing even lower false negatives (or higher recall; Table S1; Fig. S7). Similar results to those reported here for *E. coli* were observed for several of the 14 most-sampled species in our data set with available MLST schemas, albeit recall was slightly better than that observed for the *E. coli* data set (average across all ST comparisons: 83.1%, stdev 0.31) while precision was worse (average 91.6%, stdev 0.19).

The high accuracy of the 99.5% ANI threshold compared to ST assignments is due, at least in part, to the fact that hundreds, if not thousands (e.g., at least 3,000 for the *E. coli* pairs), of genes are used in each ANI calculation at this level of high relatedness. Thus, the ANI threshold is robust against horizontal gene transfer or other evolutionary events that could affect the sequence identity of one or a few loci, a known limitation of traditional

Downloaded from https://journals.asm.org/journal/mbio on 07 February 2024 by 194.95.6.38.

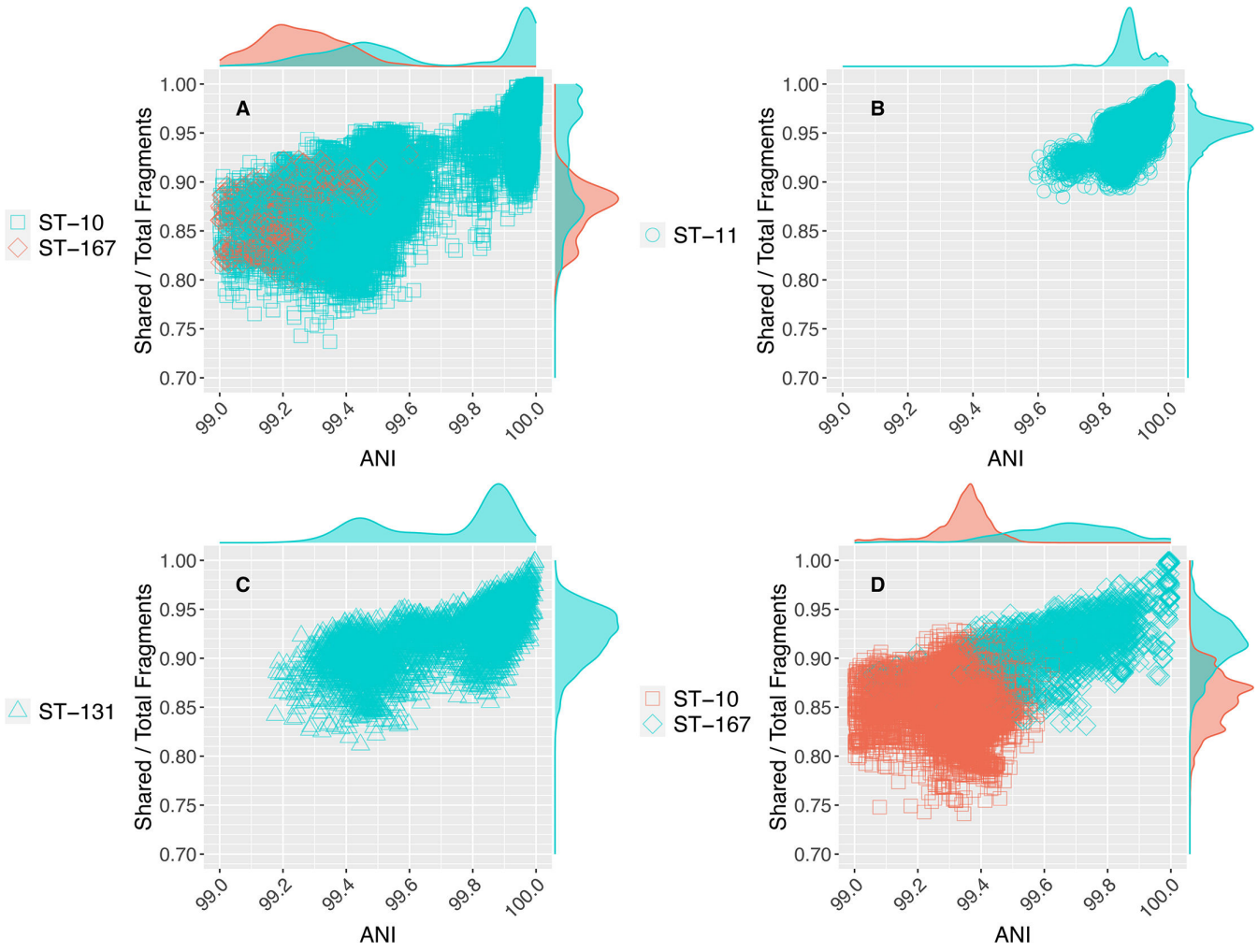

FIG 4 Comparison of the 99.5% ANI threshold to available STs of *E. coli*. All *E. coli* genomes were assigned to an ST using the tool mlst version 2.19.0 (31). The four panels show the four most abundant STs based on the number of genomes assigned to them (see Table S1; Fig. S7 for underlying data and F1 statistics, 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 closely related genomes to genomes assigned to ST-11, 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 into 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 (Figs. S1 and S2, and https://github.com/rotheconrad/bacterial_strain_definition), indicating that exceptions to the main pattern are generally infrequent.

MLST approaches. Importantly, the 99.5% ANI threshold, or a different ANI threshold to better capture intra-species diversity patterns as recommended above, is a property that emerges from the data themselves as opposed to a manmade threshold (such as "identical sequences in seven loci" used in MLST applications) and thus it should capture the natural diversity patterns better. Consistent with this interpretation, our preliminary results from applying the 99.5% ANI threshold to a collection of *E. coli* isolate genomes collected over a period of 18 months in Northern coastal Ecuador as part of the EcoZUR study (for "*E. coli* en Zonas Urbanas y Rurales") (32) shows that the 99.5% ANI-defined STs map well to local outbreaks of pathogenic *E. coli* (Feistel et al., in preparation). Therefore, using the 99.5% ANI to define new or refine existing STs could provide data-informed

Downloaded from https://journals.asm.org/journal/mbio on 07 February 2024 by 194.95.6.38.

groups that encompass genomically more homogeneous organisms compared to the existing practice in some cases. Another important advantage of the 99.5% ANI is that it can be automatically implemented and thus does not require manual curation, which is the case when establishing new ST numbers when novel (meaning, not seen previously) sequences become available (3).

ANI-based definitions for strain, ST, and CC

Our evaluation shows that the 99.5% ANI threshold, or the 99.8% if the upper end of the observed ANI gap is used instead, is directly comparable to how STs have been defined (e.g., Fig. 4) and thus could be used to complement or even substitute in the future the requirement for 6–7 identical loci that have the shortcomings explained above. We propose the midpoint (99.5% ANI) as opposed to the upper value (99.8% ANI) of the gap as a more conservative threshold and because it is the most commonly detected midpoint of the (rightmost) ANI gap observed across different species (e.g., Fig. S2, S8 and S9). We also suggest that the term genomovar could be used to refer to these 99.5% ANI intra-species units instead of STs, in case ST should maintain its original conception of 6–7 identical loci for historic, pragmatic, or other reasons. The term genomovar was originally used to name distinct genomic groups within species that cannot be distinguished phenotypically from each other and therefore cannot be classified as distinct species based on the standard taxonomic practices (33). Hence, genomovar may capture conceptually well the 99.5% ANI groups, as we also proposed recently elsewhere (11).

We think that CCs should continue to be defined based on phylogenetic analysis of STs; in fact, we see the 99.5% ANI threshold proposed here as a complementary, rather than competitive, approach to the phylogenetics because it provides a convenient means to define and/or refine STs, which can then be used in phylogenetic analysis to define CCs, assess evolutionary relationships between STs, etc. Related to this, it is important to note that the computation of ANI is, on average, two orders of magnitude faster compared to the phylogenetic placement of a genome using all core genes (21). Therefore, the ANI-based approach and thresholds proposed here should provide highly efficient and practical means to classify (type) large collections of genomes into STs at the whole-genome level and subsequently perform phylogenetic (or other) analysis of representatives of the resulting STs.

Finally, we did not observe another ANI gap within the 99.5% ANI clusters and thus recommend the use of the term strain only for nearly identical genomes. We recently proposed to define a strain as a collection of genomes sharing ANI > 99.99% based on high gene-content similarity among the *Sal. ruber* genomes related at this level and recovered from the same saltern site (11), an important prerequisite of the current definition for strain (9). Typically, >99.0% gene content is shared at this ANI level based on our previous analysis of the *Sal. ruber* genomes (11) and the data presented here for all highly sampled species (Fig. 1 and 2). Furthermore, our previous study showed that this ANI level encompasses well the typical sequencing and assembly noise observed when splitting the raw reads from a *Sal. ruber* isolate genome project in two halves and assembling the two subsets independently for comparisons of the resulting genome sequences (11). That is, this exercise almost never provided two assemblies with lower than 99.99% ANI and/or 99% gene-content similarity among them. It should be noted, however, that we have not exhaustively evaluated all possible sequencing methods and assembly tools in this respect; hence, higher sequencing noise might be observed in some cases and the strain ANI threshold might need to be adjusted accordingly. Furthermore, in cases where the ANI is lower but close to 99.99%, it is advisable to examine the SNP patterns across the genome to ensure that the lower ANI value is not due to the high sequence divergence of one or a couple of (short) regions of the genome but instead represents sequence diversity that is evenly distributed, more or less, across the whole genome. Finally, the 99.99% ANI threshold represents only practical and convenient means for defining strains, and it could be neglected or adjusted should key

Downloaded from https://journals.asm.org/journal/mbio on 07 February 2024 by 194.95.6.38.

phenotypic differences distinguishing organisms sharing ANI > 99.99% are known/found such as antibiotic resistance or catabolic genes carried by plasmids.

What are the underlying mechanisms for the 99.5% ANI gap?

The mechanism(s) that underly the 99.5% ANI gap (or the earlier 95% ANI gap for the species level) remain essentially speculative and should be the subject of future research to further advance the mechanistic understanding of the microbial diversity patterns observed in nature. Most notable is the idea that members of a population cohere together *via* means of unbiased (random) genetic exchange which is more frequent within vs between populations or CCs (i.e., *the biological* or *sexual species concept*) (34). A competing hypothesis is that several members of the species are functionally differentiated from each other either due to specialization for different growth conditions or different affinities for the same energy substrate and thus selection over time for these functions purge diversity (i.e., *the ecological species concept*) (35–37). It is intriguing to note that the ecological explanation is also consistent with the notion that STs or different strains of the same species are somewhat ecologically and/or functionally distinguishable from each other. Notably, given an estimated mutation rate of ~4×10^{-10} per nucleotide per generation (38) and between 100 and 300 generations per year (39), it would take two distinct *E. coli* lineages or STs at least a couple million years since their last common ancestor to accumulate 0.5% difference (i.e., fixed mutations) in their core genes or 99.5% ANI. Therefore, there is enough time, at least theoretically, for the ecological purging of diversity to take place at around the 99.5% ANI level and thus account for the ANI patterns observed herein. Intriguingly, it has been shown that the explicit inclusion of extinction events in a neutral model of evolution can also result in punctuated distributions of genetic differentiation, opening up a third possibility of historical contingency from stochastic events (40). However, we note that while stochasticity can explain bimodal (or multimodal) distance distributions, a scarcity of ANI values in the same range (i.e., around 99.5% ANI) would be unlikely to repeatedly emerge by chance alone across many different species with distinct lifestyles and evolutionary tempo, as opposed to this range varying between species. In any case, the data available in support of one of these (or another) hypotheses remain sparse and/or anecdotal to date, to the best of our knowledge, and the analysis presented in this study did not aim to advance this issue further but rather to present a highly intriguing observation of patterns of diversity that could have major practical consequences in defining strains, STs, and species more broadly.

Conclusions

Regardless of what the underlying mechanisms are for the 99.5% ANI gap, the results presented here show that the patterns of natural diversity among thousands of sequenced genomes often, but not always, revealed a 99.5% ANI gap that can be used to identify STs more reliably and precisely compared to the current practice. Regarding the use of this threshold to define (or refine) STs vs strains for epidemiological purposes, we believe that the threshold is highly appropriate, as it matches well the intended meaning and use of STs, and thus its application to ST definition is straightforward. However, we favor the alternative term of genomovar to refer to these 99.5% ANI units for epidemiological studies to avoid confusion with historical ST definitions. We do not see any such complications in the application of genomovar to ecological or systematic studies, and thus its use for environmental microbes may be straightforward. Notably, recent large-scale surveys of the human gut microbiome (41) have also revealed distinct intra-species units closely matching the 99.5% ANI threshold and genomovar definition proposed here, although these units were often called strains previously (8). We caution against the use of the term strain in that context since genomes showing ~0.5% or ~0.2% difference in ANI (99.5% or 99.8% ANI, respectively) often show substantial sequence and gene content divergence (e.g., Fig. 1) and thus significant anticipable phenotypic differences. Therefore, multiple (distinct) strains should be expected to be grouped

Downloaded from https://journals.asm.org/journal/mbio on 07 February 2024 by 194.95.6.38.

within the same 99.5% ANI cluster in such cases, and strain, in general, represents a more fine-grained level of resolution than the 99.5% ANI level. The results presented here based on all well-sampled bacterial species further reinforced our recent proposal based on the *Sal. ruber* isolate genomes from the same site to use 99.99% ANI as the threshold to define strains (11). It is important to note, however, that the 99.5% ANI gap is not observed in all species examined, and exceptions are not common but do exist (e.g., Fig. 4D; Fig. S2D). It would be important to understand the underlying reasons for these exceptions, which could include isolation biases as discussed above, toward a more complete understanding of sub-species diversity patterns. Until then, we recommend evaluating the ANI value distribution for the species of interest, and if the data indicate so, adjusting the ANI threshold to match the gap in the observed ANI value distribution. The ANI thresholds proposed here should represent useful reference points for this type of analysis. Collectively, we expect that the findings reported here will advance the molecular toolbox for accurately delineating and following the important units of diversity within prokaryotic species and thus would greatly facilitate future epidemiological and micro-diversity studies.

MATERIALS AND METHODS

Step-by-step methods, including how average trendlines were fit to the data, custom Python code, NCBI Assembly accession numbers for selected genomes, and plots for each selected species are available from: https://github.com/rotheconrad/bacterial_strain_definition. Briefly, all genome sequences were obtained from NCBI's RefSeq Assembly database on 20 April 2022 and were labeled as "complete" and "latest." ANI values and the shared genome fractions were directly obtained from the output of FastANI version 1.32, "One to Many" mode with default settings (21). Results were concatenated to create within species all vs all output. Self-matches were removed, and genome pairs were filtered by minimum ANI values according to the axes of each figure. Selected individual species plots (i.e., all vs all output) of shared genome fraction vs ANI are shown in the Supplementary Material. *E. coli* genomes were assigned to sequence types (ST) using the command-line tool mlst (https://github.com/tseemann/mlst) version 2.19.0 (31) with default settings. The statistical evaluation of the ANI gaps and additional results emerging from this analysis are provided in the Supplementary Material.

ACKNOWLEDGMENTS

This work has been supported by the US National Science Foundation (Award No 1759831 and 2129823) to KTK. L.O. and R.A. acknowledge funding by the Max Planck Society. R.R.M. and T.V. acknowledge the project PID2021-126114NB-C42 of the Spanish Ministry of Science, Innovation and Universities also supported European Regional Development Fund (FEDER) funds. R.R.M. and T.V. research was part of the activities of the "Maria de Maeztu Centre of Excellence" accreditation CEX2021-001198. LMR acknowledges funding from the University of Innsbruck. The computational results presented here have been achieved (in part) using the LEO infrastructure of the University of Innsbruck.

AUTHOR AFFILIATIONS

[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, Georgia, USA
[3]Department of Animal and Microbial Biodiversity, Marine Microbiology Group, 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

Downloaded from https://journals.asm.org/journal/mbio on 07 February 2024 by 194.95.6.38.

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

AUTHOR ORCIDs

Luis M. Rodriguez-R http://orcid.org/0000-0001-7603-3093
Roth E. Conrad http://orcid.org/0000-0001-8155-8441
Dorian J. Feistel http://orcid.org/0000-0001-9478-3170
Konstantinos T. Konstantinidis http://orcid.org/0000-0002-0954-4755

FUNDING

Funder	Grant(s)	Author(s)
National Science Foundation (NSF)	1759831, 2129823	Konstantinos T. Konstantinidis

AUTHOR CONTRIBUTIONS

Luis M. Rodriguez-R, Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – review and editing | Roth E. Conrad, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – review and editing | Tomeu Viver, Investigation, Validation, Writing – review and editing | Dorian J. Feistel, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization | Blake G. Lindner, Data curation, Formal analysis, Investigation, Methodology, Visualization | Stephanus N. Venter, Investigation, Validation, Writing – review and editing | Luis H. Orellana, Data curation, Investigation, Validation, Writing – review and editing | Rudolf Amann, Resources, Supervision, Writing – review and editing | Ramon Rossello-Mora, Resources, Supervision, Writing – review and editing.

DATA AVAILABILITY

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

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental Material (mBio02696-23-S0001.pdf). Supplemental methods, figures, and Tables.

REFERENCES

1. Hanage WP, Fraser C, Spratt BG. 2005. Fuzzy species among recombinogenic bacteria. BMC Biol 3:6. https://doi.org/10.1186/1741-7007-3-6
2. Rossello-Mora R, Amann R. 2015. Past and future species definitions for bacteria and archaea. Syst Appl Microbiol 38:209–216. https://doi.org/10.1016/j.syapm.2015.02.001
3. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA 95:3140–3145. https://doi.org/10.1073/pnas.95.6.3140
4. Baltrus DA, Dougherty K, Beckstrom-Sternberg SM, Beckstrom-Sternberg JS, Foster JT. 2014. 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. https://doi.org/10.1111/mpp.12103
5. Konstantinidis KT, Ramette A, Tiedje JM. 2006. Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. Appl Environ Microbiol 72:7286–7293. https://doi.org/10.1128/AEM.01398-06
6. Rodriguez RL, Jain C, Conrad RE, Aluru S, Konstantinidis KT. 2021. Reply to:"re-evaluating the evidence for a universal genetic boundary among microbial species" Nat Commun 12:4060. https://doi.org/10.1038/s41467-021-24129-1
7. Dijkshoorn L, Ursing BM, Ursing JB. 2000. Strain, clone and species: comments on three basic concepts of bacteriology. J Med Microbiol 49:397–401. https://doi.org/10.1099/0022-1317-49-5-397
8. Yan Y, Nguyen LH, Franzosa EA, Huttenhower C. 2020. Strain-level epidemiology of microbial communities and the human microbiome. Genome Med 12:71. https://doi.org/10.1186/s13073-020-00765-y
9. Parker CT, Tindall BJ, Garrity GM. 2019. International code of nomenclature of Prokaryotes. Int J Syst Evol Microbiol 69:S1–S111. https://doi.org/10.1099/ijsem.0.000778

Downloaded from https://journals.asm.org/journal/mbio on 07 February 2024 by 194.95.6.38.

10. Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, Persing DH, Swaminathan B. 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field GEL electrophoresis: criteria for bacterial strain typing. J Clin Microbiol 33:2233–2239. https://doi.org/10.1128/jcm.33.9.2233-2239.1995
11. Viver T, Conrad RE, Rodriguez-R LM, Ramirez AS, Venter SN, Rocha-Cardenas J, Segura ML, Amann R, Konstantinidis KT, Rossello-Mora R. 2023. Towards estimating the number of strains that make up a natural bacterial population. bioRxiv. https://doi.org/10.1101/2023.02.20.529252
12. Caro-Quintero A, Konstantinidis KT. 2012. Bacterial species may exist, metagenomics reveal. Environ Microbiol 14:347–355. https://doi.org/10.1111/j.1462-2920.2011.02668.x
13. Viver T, Conrad RE, Orellana LH, Urdiain M, Gonzalez-Pastor JE, Hatt JK, Amann R, Anton J, Konstantinidis KT, Rossello-Mora R. 2021. Distinct ecotypes within a natural haloarchaeal population enable adaptation to changing environmental conditions without causing population sweeps. ISME J 15:1178–1191. https://doi.org/10.1038/s41396-020-00842-5
14. Conrad RE, Viver T, Gago JF, Hatt JK, Venter SN, Rossello-Mora R, Konstantinidis KT. 2022. Toward quantifying the adaptive role of bacterial pangenomes during environmental perturbations. ISME J 16:1222–1234. https://doi.org/10.1038/s41396-021-01149-9
15. Murray CS, Gao Y, Wu M. 2021. Re-evaluating the evidence for a universal genetic boundary among microbial species. Nat Commun 12:4059. https://doi.org/10.1038/s41467-021-24128-2
16. Konstantinidis KT, DeLong EF. 2008. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. ISME J 2:1052–1065. https://doi.org/10.1038/ismej.2008.62
17. Bendall ML, Stevens SL, Chan LK, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, Froula J, Kang D, Tringe SG, Bertilsson S, Moran MA, Shade A, Newton RJ, McMahon KD, Malmstrom RR. 2016. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J 10:1589–1601. https://doi.org/10.1038/ismej.2015.241
18. Johnston ER, Rodriguez Rl, Luo C, Yuan MM, Wu L, He Z, Schuur EA, Luo Y, Tiedje JM, Zhou J, Konstantinidis KT. 2016. Metagenomics reveals pervasive bacterial populations and reduced community diversity across the Alaska Tundra ecosystem. Front Microbiol 7:579. https://doi.org/10.3389/fmicb.2016.00579
19. Meziti A, Tsementzi D, Rodriguez-R LM, Hatt JK, Karayanni H, Kormas KA, Konstantinidis KT. 2019. Quantifying the changes in genetic diversity within sequence-discrete bacterial populations across a spatial and temporal riverine gradient. ISME J 13:767–779. https://doi.org/10.1038/s41396-018-0307-6
20. Olm MR, Crits-Christoph A, Diamond S, Lavy A, Matheus Carnevali PB, Banfield JF. 2020. Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. mSystems 5:e00731-19. https://doi.org/10.1128/mSystems.00731-19
21. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9
22. Denamur E, Clermont O, Bonacorsi S, Gordon D. 2021. The population genetics of pathogenic *Escherichia coli*. Nat Rev Microbiol 19:37–54. https://doi.org/10.1038/s41579-020-0416-x
23. Antón J, Rosselló-Mora R, Rodríguez-Valera F, Amann R. 2000. Extremely halophilic bacteria in crystallizer ponds from solar salterns. Appl Environ Microbiol 66:3052–3057. https://doi.org/10.1128/AEM.66.7.3052-3057.2000
24. Viver T, Orellana LH, Díaz S, Urdiain M, Ramos-Barbero MD, González-Pastor JE, Oren A, Hatt JK, Amann R, Antón J, Konstantinidis KT, Rosselló-Móra R. 2019. Predominance of deterministic microbial community dynamics in salterns exposed to different light intensities. Environ Microbiol 21:4300–4315. https://doi.org/10.1111/1462-2920.14790
25. Casamayor EO, Massana R, Benlloch S, Øvreås L, Díez B, Goddard VJ, Gasol JM, Joint I, Rodríguez-Valera F, Pedrós-Alió C. 2002. 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. https://doi.org/10.1046/j.1462-2920.2002.00297.x
26. Gomariz M, Martínez-García M, Santos F, Rodriguez F, Capella-Gutiérrez S, Gabaldón T, Rosselló-Móra R, Meseguer I, Antón J. 2015. From community approaches to single-cell genomics: the discovery of ubiquitous hyperhalophilic bacteroidetes generalists. ISME J 9:16–31. https://doi.org/10.1038/ismej.2014.95
27. K. I. V. Chandni S, Meunier CL, Rick J, Wiltshire KH, Steinke N, Vidal-Melgosa S, Hehemann J-H, Huettel B, Schweder T, Fuchs BM, Amann RI, Teeling H. 2022. 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
28. Gounot J-S, Chia M, Bertrand D, Saw W-Y, Ravikrishnan A, Low A, Ding Y, Ng AHQ, Tan LWL, Teo Y-Y, Seedorf H, Nagarajan N. 2022. Genome-centric analysis of short and long read metagenomes reveals uncharacterized microbiome diversity in Southeast Asians. Nat Commun 13:6044. https://doi.org/10.1038/s41467-022-33782-z
29. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, Gregor I, Majda S, Fiedler J, Dahms E, Bremges A, Fritz A, et al. 2017. Critical assessment of metagenome interpretation-a benchmark of metagenomics software. Nat Methods 14:1063–1071. https://doi.org/10.1038/nmeth.4458
30. Konstantinidis KT, Ramette A, Tiedje JM. 2006. The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci 361:1929–1940. https://doi.org/10.1098/rstb.2006.1920
31. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MCJ, Ochman H, Achtman M. 2006. Sex and virulence in *Escherichia coli* : an evolutionary perspective. Mol Microbiol 60:1136–1151. https://doi.org/10.1111/j.1365-2958.2006.05172.x
32. Peña-Gonzalez A, Soto-Girón MJ, Smith S, Sistrunk J, Montero L, Páez M, Ortega E, Hatt JK, Cevallos W, Trueba G, Levy K, Konstantinidis KT. 2019. Metagenomic signatures of gut infections caused by different *Escherichia coli* pathotypes. Appl Environ Microbiol 85:e01820-19. https://doi.org/10.1128/AEM.01820-19
33. Ursing JB, Rossello-Mora RA, Garcia-Valdes E, Lalucat J. 1995. Taxonomic note: a pragmatic approach to the nomenclature of phenotypically similar genomic groups. Int J SystEvol Microbiol 45:604–604. https://doi.org/10.1099/00207713-45-3-604
34. Power JJ, Pinheiro F, Pompei S, Kovacova V, Yüksel M, Rathmann I, Förster M, Lässig M, Maier B. 2021. Adaptive evolution of hybrid bacteria by horizontal gene transfer. Proc Natl Acad Sci U S A 118:e2007873118. https://doi.org/10.1073/pnas.2007873118
35. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz MF, Alm EJ. 2012. Population genomics of early events in the ecological differentiation of bacteria. Science 336:48–51. https://doi.org/10.1126/science.1218198
36. Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation. Science 315:476–480. https://doi.org/10.1126/science.1127573
37. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43. https://doi.org/10.1038/nature02340
38. Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. Genetics 148:1667–1686. https://doi.org/10.1093/genetics/148.4.1667
39. Gibbons RJ, Kapsimalis B. 1967. Estimates of the overall rate of growth of the intestinal microflora of hamsters, guinea pigs, and mice. J Bacteriol 93:510–512. https://doi.org/10.1128/jb.93.1.510-512.1967
40. Straub TJ, Zhaxybayeva O. 2017. A null model for microbial diversification. Proc Natl Acad Sci U S A 114:E5414–E5423. https://doi.org/10.1073/pnas.1619993114
41. Valles-Colomer M, Blanco-Míguez A, Manghi P, Asnicar F, Dubois L, Golzato D, Armanini F, Cumbo F, Huang KD, Manara S, Masetti G, et al. 2023. The person-to-person transmission landscape of the gut and oral microbiomes. Nature 614:125–135. https://doi.org/10.1038/s41586-022-05620-1

Downloaded from https://journals.asm.org/journal/mbio on 07 February 2024 by 194.95.6.38.