A natural definition for a bacterial strain and clonal complex.

Luis M. Rodriguez-R[1*], Roth E. Conrad[2,*], Dorian J. Feistel[2], Tomeu Viver[3], Ramon Rosselló-Móra[3], and Konstantinos T. Konstantinidis[2]

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

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

*equal first authors.

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

Abstract

Large scale surveys of prokaryotic communities (metagenomics) as well as isolate genomes have recently revealed that prokaryotic diversity is predominantly organized in sequence-discrete units that may be equated to species. Specifically, genomes of the same unit (or species) commonly show >95% genome-aggregate average nucleotide identity (ANI) to each other and <90% ANI to members of other species, while genomes showing 90-95% ANI are comparatively rare. However, it remains unclear if such "discontinuities" or gaps in ANI values can be observed within species and thus, be used to define strains or clonal complexes; two cornerstone concepts for microbiology that remain ill-defined. By analyzing 18,123 complete isolate genomes from 330 bacterial species with at least ten genome representatives each, we show that such a natural discontinuity exists at around 99.5% ANI. Further, we show that the 99.5% ANI threshold is largely consistent with how clonal complexes have been defined in previous epidemiological studies but provides clusters with ~20% higher accuracy in terms of evolutionary relatedness of the grouped genomes and greater homogeneity in gene content. Collectively, our results should facilitate future micro-diversity studies across clinical or environmental settings because they provide a more natural definition of a clonal complex and strain.

Main

The strain represents a fundamental unit of microbial diversity that is commonly used across medical or environment studies to designate the smallest distinguishable unit. This term is based on the concept of a pure culture (1), and has been defined by the Bacteriological Code as the group of cultures made "*of the descendants of a single isolation in pure culture*" (2). 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 (3). This strain definition, while being commonly used across microbiological fields, remains problematic, however, and often causes confusion in communication. Most notably, it is not clear when two distinct genomes or cells should be considered the same or separate strains since cell ancestry information is often missing, and the existence of phenotypic similarities (or differences) may vary, depending on the growth conditions. Additionally, the isolation of an organism (wild-type) in the laboratory is frequently accompanied by phenotypic changes due to adaptation to the laboratory conditions; yet, the wild-type and the lab-adapted cells or types are typically considered the same strain (1).

A related term that is also commonly used to catalogue intra-species diversity is the clonal complex (CC). The CC has been used, especially in medical microbiology and epidemiological studies, to identify -for instance- an outbreak caused by a specific pathogenic strain or groups (complexes) of highly related strains. Unlike the more relaxed, context-dependent definition of a "strain", a CC is typically defined as a collection of genomes with no nucleotide sequence diversity (zero single-nucleotide polymorphisms) in 6-7 genetic loci (4). These loci are typically distributed across the genome, to avoid co-selection evolutionary events, and represent fragments (amplicons) of genes shared by most members of a species; that is, core genes. While this definition is pragmatic and operational, it is also problematic because it is arbitrary (e.g., sometimes CC definitions allow 0- or 1-point mutations), and core genes tend to be more conserved than the genome average. Thus, it remains somewhat speculative how similar (or not) isolates of the same CC may be in the rest of their genome, and this may also depend, at least partly, on the exact loci used in the analysis (5, 6). If the intra-species diversity was organized in discrete units that can be equated to CC or strains, this would have provided for more natural and meaningful definitions for CCs and strains compared to the existing definitions 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 (7), it remains unclear whether or not such units also exist within species.

Specifically, 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% genome-aggregate average nucleotide identity (or ANI) depending on the population considered (e.g., younger populations since the last population diversity sweep event show lower levels of intra-population diversity). In contrast, ANI values between distinct populations are typically lower than 90% (8). 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 (7, 9, 10) rather than representing cultivation or other sampling biases (11). Such sequence-discrete populations have been recovered from many different habitats, including marine, freshwater, soils, human gut, and biofilms, and were usually persistent over time and space [e.g., (12-16)] indicating that they are not ephemeral but long-lived entities. Further, these sequence-discrete populations commonly harbor substantial intra-population gene content diversity (i.e., they are rarely clonal) (12, 15). 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 genomes have been classified into (named) species in the last couple decades; that is, ~97% of named species include only organisms with genomes sharing >95% ANI (17). 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 (17) [Discontinuity or gap here refers to the small number of genome pairs showing 85–95% ANI relative to counts of pairs showing >95% and <85% ANI]. Whether or not a similar ANI gap exists and can be used to define strains or CC has not been evaluated yet.

Results/Discussion

An ANI gap within species around 99.5%

In the process of assessing cultivation biases as a possible explanation for the ANI-based sequence discrete populations (18), we observed another discontinuity (or gap) in ANI values that may be used to define the smallest unit within a species, that of the strain or CC. Specifically, the analysis of 18,123 complete genomes from 330 species available in NCBI's Assembly database with at least 10 such genome representatives each revealed a clear bimodal distribution in ANI values within named species or 95% ANI-defined groups of genomes (genomospecies). That is, there is a scarcity of genomes pairs showing 99.2-99.8% ANI (average around 99.5% ANI) in contrast to genome pairs showing >99.8% or <99.2% ANI. Specifically, among the 18,123 complete genomes in our dataset, there are 4,280,133 genome pairs showing >96% ANI, which would translate to about 107,000 pairs per every 0.1 percent unit of ANI if there was no bimodal distribution but the ANI values among these genome pairs were evenly distributed between 96% and 100% ANI. Our analysis revealed only 235,527 genome pairs between 99.2% and 99.8% ANI, which is three-fold fewer data points than expected by chance alone in a uniform ANI value distribution (642,000 pairs expected). No other ANI range within 96-100% had such a strong bias based on our dataset. That is, a pronounced gap in ANI values is observed among very closely related members of a species around 99.5% ANI (Fig. 1). 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, and 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. S1). Instead, it represents a universal property of the 330 species evaluated (see also Fig. S2 for specific species examples). Therefore, it appears that another important level of genomic differentiation may exist within species that can be used to define strains and CC.

It is unlikely that this 99.5% intra-species ANI gap is due to cultivation or classification biases due the reasons mentioned previously, such as that cultivation media should not distinguish between members of the same species or closely related groups of organisms (18), and that random subsampling provided similar patterns (Fig. S1). In fact, it is highly likely that the intra-species ANI gap may be even more pronounced in nature because very closely related genomes (e.g., showing >99.8% ANI 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., RARP-PCR, MLST) in order to avoid sequencing of redundant genomes. Despite this bias against very closely related genomes, which in all probability exists but we are currently unable to estimate its magnitude, several species have enough very closely related genomes sequenced in our dataset for robust evaluation of patterns in their ANI value distributions (Fig. 1 and S2). We were not able to identify clear exceptions to this 99.5% intra-species ANI gap when examining individual species with enough sequenced representatives, although such exceptions likely exist. For instance, several species in our collection did not have enough very highly related genome representatives (showing >99% ANI to each other) to assess the critical area of ANI value distribution (i.e., the 99-100% range), and this could be due to the pre-screening biases mentioned above or reflect their actual natural diversity patterns (see also Fig. 2, Panel D for an example within the *E. coli*). Further, for a few species (n < 10) such as *Listeria monocytogenes* and *Bordetella bronchiseptica*, the intra-species ANI gap appears to exist but is shifted compared to the 99.5% ANI that characterized most well-sampled species (Fig. 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 99.5% ANI should work for most species based on the dataset evaluated here.

Gene content diversity within 99.5% ANI clusters

Another notable observation from the data from all species comparisons is that shared gene content decreases, on average, 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 99.0-100% ANI but then, the decrease is less dramatic in genome pairs sharing between 96.0%-99.0% ANI. In other words, genome pairs sharing between 99.0%-100% ANI (i.e., one ANI unit range) may differ in their total gene content by up to 10% (average values of genome pairs showing ~99.0% ANI) and more divergent genomes of the same species (i.e., showing 96.0%>ANI>99.0%) may differ by up to 20% (average values of genome pairs showing ~96.0% ANI), adding another ~10% of gene content differences for 3 additional units of ANI (vs. 1 unit in the 99-100% range). Collectively, these results show that genome pairs showing >99.5% ANI are also expected to be much more similar in gene-content compared to more divergent genomes of the same species. Not only do these results quantify the amount of gene content diversity expected in comparisons of genomes within the same species, but they are also consistent with the notion that members of the same CC should be highly similar in shared functional gene content and thus, phenotype. While the high gene content diversity within species has been observed previously, even based on the very first few bacterial genomes sequenced (19), the analysis presented here provides robust quantification of shared gene content as a function of ANI and CC or strain designation, which should be useful in future studies that aim to quantify the size and value of bacterial species pangenomes.

Comparison to clonal complexes (or Sequence Types, STs)

We also assessed how consistent the 99.5% ANI threshold is with the assignment of genomes to the same CC, the latter defined as identical sequences for 6-7 genetic loci. We focus this analysis on the *E. coli* species because it is a good representative of the ANI patterns observed within other species in our dataset, the large number of *E. coli* genomes available (n=975), and the availability of a robust Multi-Locus Sequence Typing/Analysis (MLST/MLSA) scheme (20) 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 CC, also commonly called Sequence Type (ST), based on identical sequences in seven *E. coli* core genes (namely, *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA*) (20). Our evaluation showed that the 99.5% ANI threshold is largely consistent with how genomes are assigned to STs; that is, ~80% of ST assignments, for the top four most abundant STs (n=615 genomes), were supported by the 99.5% ANI threshold (Fig. 2). In other words, only about 20% of the genomes that were assigned to the same ST showed <99.5% ANI among themselves (high recall). Further, ~2% of the genomes with >99.5% were assigned to different ST, revealing even higher precision (Table S1; Suppl. Fig. S3). The high recall (and precision) of the 99.5% ANI threshold 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), or have been exchanged horizontally very recently among the evaluated genomes. It is also important to note 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 vs. 6-7 genes in MLST schemes and that’s another reason for the higher robustness of the former metric against horizontal gene transfer or other evolutionary events that could affect sequence identity. These results are also consistent with our previous conclusion that the 95% ANI threshold should be adjusted upwards if the ANI is based on a few universal or core genes (17). In any case, the 99.5% ANI threshold is a property that emerges from the data themselves as opposed to a manmade, arbitrary identity threshold (i.e., 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") (21) shows that the 99.5%-ANI-defined CCs map well to local outbreaks of pathogenic *E. coli* (Feistel et al., in preparation). Therefore, using the 99.5% ANI to define CCs will prevent inaccurate calls and provide more biologically informed means to define CCs. 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 (4).

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 in order 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*). 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*) (22-24). It is intriguing to note that the ecological explanation is also consistent with the notion that CCs or different strains of the same species are somewhat ecologically and/or functionally distinguishable from each other. Notably, given an estimated mutation rate of ~4x10^{-10} per nucleotide per generation (25) and between 100 to 300 generations per year (26), it would take two distinct *E. coli* lineages or CCs at least forty thousand years since their last common ancestor to accumulate 1% difference (i.e., fixed mutations) in their core genes or 99% 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 (27). However, we note that while stochasticity can explain bimodal (or multimodal) distance distributions, a scarcity of ANI values in the exact 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.

Conclusion

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 are consistent with a 99.5% ANI threshold that can be used to identify CCs and strains more reliably and precisely compared to the current practice. Regarding the use of this threshold to define CCs vs. strains, we believe that the threshold is highly appropriate, as well as it matches well the intended meaning and use of CCs, and thus its application to CC definition is straightforward. For the strain level, 0.5% or 0.2% difference in ANI (correspondingly, 99.5% and 99.8% ANI) represents substantial, non-trivial, genomic divergence that, in most cases, would likely encompass several genomes with at least some phenotypic differences (due to substantial sequence or gene content differences among the genomes; see Fig. 1). Thus, multiple strains will be likely grouped together under 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. That said, we also expect that the latter would somewhat depend on the context of the study and the existence or not of phenotypic differences. For instance, it is possible that all sequence and gene-content differences within some 99.5% ANI genome clusters to be neutral (or not functional), for at least some growth conditions and habitats. Therefore, in such cases, a strain could be defined at the 99.5% ANI level and thus, include high(er) intra-strain sequence and/or gene-content diversity. Hence, the 99.5% ANI level is also a good starting point in trying to define strains and subsequently assess their gene and phenotypic differences (or lack of). Collectively, we expect that the findings reported here will advance the molecular toolbox for accurately delineating and following the important units of diversity withing prokaryotic species and thus, would greatly facilitate future epidemiological and micro-diversity studies.

Material 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 April 20th, 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 (Jain et al 2018). 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 using the command-line tool mlst (https://github.com/tseemann/mlst) version 2.19.0 (20) with default settings.

Acknowledgments

This work has been supported by the US National Science Foundation (Award No 1759831 and 2129823) to KTK.

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
Ramon Rosselló-Móra: ramon@imedea.uib-csic.es
Konstantinos T. Konstantinidis: kostas.konstantinidis@gatech.edu

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 (17) 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, and that panel B is a zoomed-in version of panel A. The main scatter plot is shaded by density of the 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.

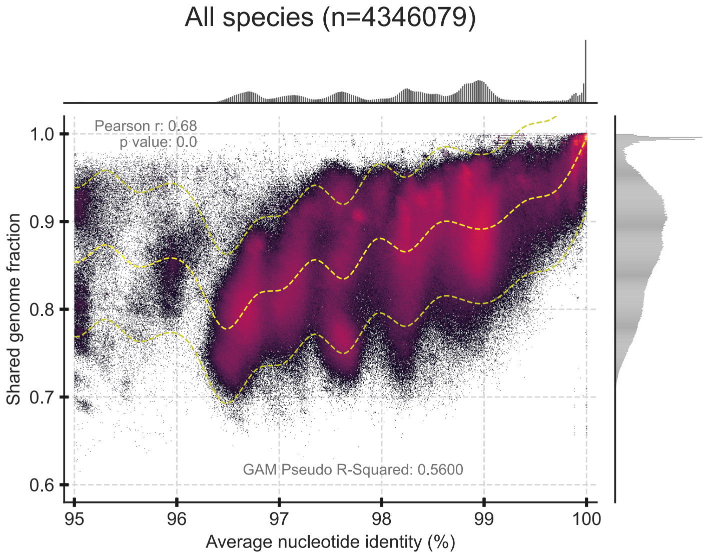


Figure 1, panel B.

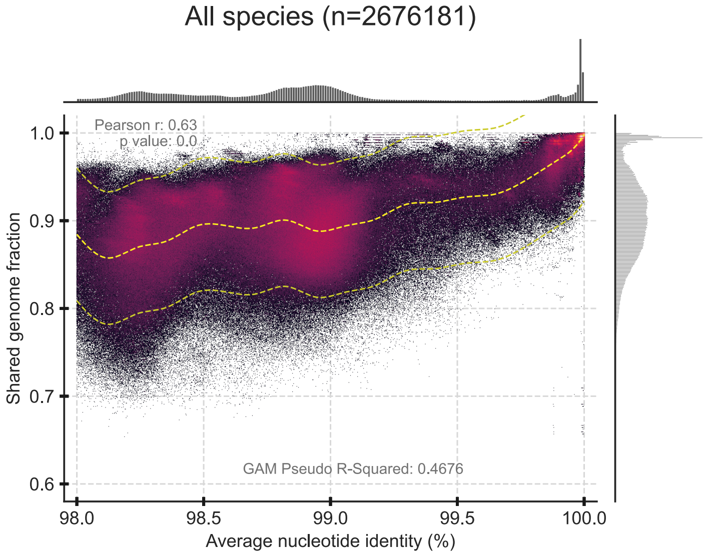
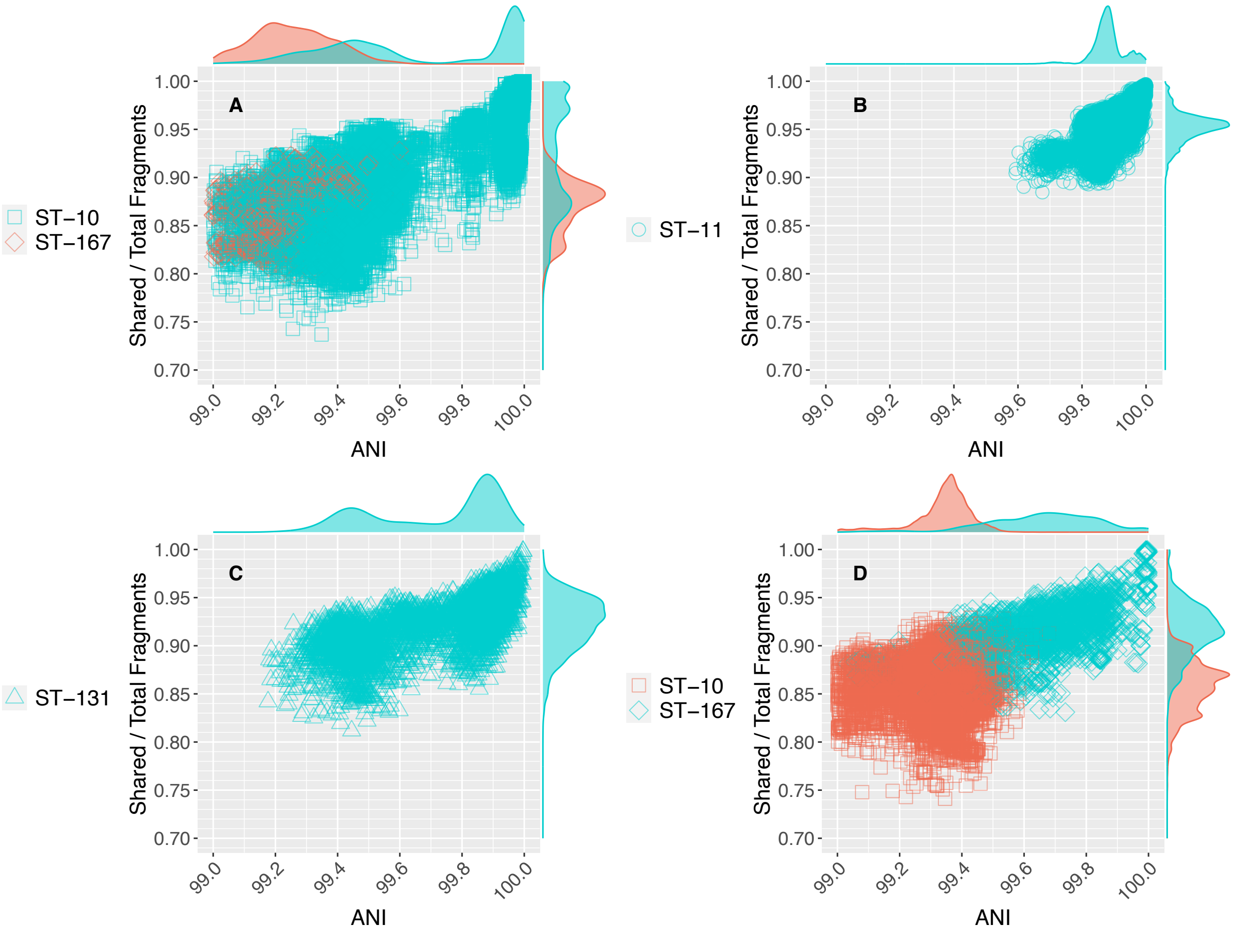


Figure 2. Comparison of the 99.5% ANI threshold to available clonal complexes of *E. coli*. All *E. coli* genomes were assigned to a Sequence Type, ST (or clonal complex) using the command-line tool mlst (https://github.com/tseemann/mlst) version 2.19.0 (20). The four panels show the four most abundant STs based on the number of genomes assigned to them (see Table S1 and Figure S1 for underlying data and F1 statistic, respectively). Each datapoint is a comparison between two genomes, similar to those shown in Figure 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. CC 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 CC based on the 99.5% ANI criterion (and the bimodal ANI value distribution around 99.6% ANI).



References

1. 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).
2. 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).
3. 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).
4. 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).
5. 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).
6. 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).
7. 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).
8. A. Caro-Quintero, K. T. Konstantinidis, Bacterial species may exist, metagenomics reveal. *Environ Microbiol* **14**, 347-355 (2012).
9. 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).
10. R. Conrad *et al.*, Toward quantifying the adaptive role of bacterial pangenomes during environmental perturbations. *The ISME Journal* https://doi.org/10.1101/2021.03.15.435471, Accepted (2021).
11. C. S. Murray, Y. Gao, M. Wu, Re-evaluating the evidence for a universal genetic boundary among microbial species. *Nat Commun* **12**, 4059 (2021).
12. K. T. Konstantinidis, E. F. DeLong, Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J* **2**, 1052-1065 (2008).
13. M. L. Bendall *et al.*, Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J* **10**, 1589-1601 (2016).
14. E. R. Johnston *et al.*, Metagenomics Reveals Pervasive Bacterial Populations and Reduced Community Diversity across the Alaska Tundra Ecosystem. *Front Microbiol* **7**, 579 (2016).
15. 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).
16. M. R. Olm *et al.*, Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems* **5** (2020).

17. 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).
18. L. M. Rodriguez-R, C. Jain, R. E. Conrad, S. Aluru, K. T. Konstantinidis, Rebuttal to Murray and colleagues Nature Communication Matters Arising article. . *Narure Communications*, In press (2021).
19. R. A. Welch *et al.*, Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proc Natl Acad Sci U S A* **99**, 17020-17024 (2002).
20. T. Wirth *et al.*, Sex and virulence in Escherichia coli: an evolutionary perspective. *Mol Microbiol* **60**, 1136-1151 (2006).
21. A. Pena-Gonzalez *et al.*, Metagenomic Signatures of Gut Infections Caused by Different Escherichia coli Pathotypes. *Appl Environ Microbiol* **85** (2019).
22. B. J. Shapiro *et al.*, Population genomics of early events in the ecological differentiation of bacteria. *Science* **336**, 48-51 (2012).
23. C. Fraser, W. P. Hanage, B. G. Spratt, Recombination and the nature of bacterial speciation. *Science* **315**, 476-480 (2007).
24. G. W. Tyson *et al.*, Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43 (2004).
25. J. W. Drake, B. Charlesworth, D. Charlesworth, J. F. Crow, Rates of spontaneous mutation. *Genetics* **148**, 1667-1686 (1998).
26. Gibbons RJ, K. B, Estimates of the overall rate of growth of the intestinal microflora of hamsters, guinea pigs, and mice. *Journal of Bacteriology* **93**, 510–512 (1967).
27. T. J. Straub, O. Zhaxybayeva, A null model for microbial diversification. *Proc Natl Acad Sci U S A* **114**, E5414-E5423 (2017).