Evolutionary Genetic Species Detected in Prokaryotes by Applying the K/θ Ratio to DNA Sequences

Using K/θ to delimit bacterial species

C. William Birky, Jr.

Department of Ecology and Evolutionary Biology, The University of Arizona, Tucson, Arizona, United States of America

Abstract

In previous papers in this series, I described the evolutionary genetic species concept which is based on population genetic theory and should be applicable to any organism. I also described a species criterion, the K/θ ratio, that delimits independently evolving evolutionary species based on single-gene sequences. I then illustrated its application to sexual and asexual eukaryotes. Here, I show how the evolutionary genetic species concept and the K/θ ratio can be applied to bacteria, using the sequences from some genes of the core genome that are rarely, if ever, transferred horizontally between different species. This completes the demonstration that K/θ is a general method for species delimitation, applicable to all organisms. Also, it adds to the evidence that bacteria have species in the most general sense, even though they have the ability to exchange genes across species boundaries. Finally, I show that a published critique of the use of K/θ ≥ 4 as a criterion for independently evolving species rests on errors in the application of population/evolutionary genetic theory.

Citation: Birky CW Jr

Editor:

Received ; Accepted ; Published

Copyright: @ 2020 C. William Birky Jr. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The author's personal funds were the sole support of the research, and the author was solely responsible for study design, data collection and analysis, decision to publish, and preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: birky@u.arizona.edu

Introduction

A prominent feature of the biological world is discontinuous variation, with clusters of similar organisms differing significantly from organisms in other clusters. These clusters, called species and given different binomial names, were originally recognized by their different phenotypes. With the development of population genetic theory, species have been recognized as populations with distinct gene pools underlying their phenotypic differences. When gene sequences are used to reconstruct evolutionary history in the form of phylogenetic trees, different species appear as different clades of similar sequences separated by deep gaps in the tree.

This background is widely, but not universally, accepted. Doubters often focus on the gradual divergence of sister populations from each other, wondering at what point the y should be called different species. Even among those who agree that species are real, distinct populations, there is disagreement on how to define and recognize them, given the gradual divergence of recently separated populations. Many authors simply point to "deep gaps" in a phylogenetic tree, usually of gene sequences, as evidence for different species, but this raises the question: "How deep is deep enough?"

This question can be answered by thinking of species as independently evolving populations. This is arguably the most general model of species (species concept), applicable in principle to all organisms and consistent with population and evolutionary genetic theory [1-8]. Different species are populations that evolve independently of each other, in the sense that they are independent arenas for mutation, natural selection, and random genetic drift. In other words, a new mutant allele arising in one population cannot spread to the other, or if it does move from one species to the other by hybridization, it is very unlikely to be fixed in that population. Moreover, if a shared allele is fixed or lost in one population by drift and/or selection, this does not affect the probability that it will be fixed or lost in the other. I call this the Evolutionary Genetic Species Concept (EGSC) to distinguish it from the Evolutionary Species Concept (EvSC) of Simpson (1). The EvSC as it is usually worded requires that we know, or assume, that the populations in question will continue to evolve independently in the indefinite future. The requirement for future independence should not be part of a species concept because is unknowable; for instance, a trait difference required for adaptation of species A to a different niche from species B could be lost in the future. (However, it should be noted that in 1953 Simpson did allow that speciation might occur in allopatry without ecological differences in "some probably rare cases" [1, p. 171].

In practice, a pair of sister species are recognized as soon as they have diverged to the point where samples of individuals can be shown to come from two independently evolving populations. This can be done using a number of different criteria for

independent evolution. Prominent examples that have been applied to prokaryotes as well as eukaryotes are reproductive isolation and ecological isolation, both of which obviously result in independent evolution of the isolated populations. However, both of these species criteria have conceptual and practical problems. Neither one is generally applicable; reproductive isolation is relevant only for organisms with significant amounts of sexual reproduction, and neither criterion includes the possibility that species may be evolving independently simply because they are physically isolated from each other. Moreover, neither is readily applicable to large-scale biodiversity surveys such as the Barcode of Life based on DNA sequences of one or a few genes isolated from many individuals.

Although DNA sequences are widely used to delimit species in prokaryotes, there is a great deal of controversy about how this should be done, and even about whether bacteria really have species. Some of this controversy mirrors similar debates about species in eukaryotes. However, much of it arises from the phenomenon of horizontal gene transfer (HGT), in which segments of genes, whole genes, and even large segments of chromosomes are transferred between distantly related bacteria, in addition to the more common forms of bacterial gene exchange via conjugation, transformation, and transduction. This problem can be solved by distinguishing between core genes and auxiliary genes [9,10] The core genes are responsible for basic functions such as DNA and protein synthesis and metabolic pathways which are common to all members of closely related populations adapted to similar niches. Auxiliary genes include genes that confer pathogenicity to specific hosts, resistance to certain antibiotics, or adaptation to a

103 novel ecological niche. Auxiliary genes are often exchanged between distantly related
104 bacteria, but core genes are not.
105
106 In this paper, as in previous publications [3-8], I argue that the EGSC can be applied
107 to all organisms, sexual and asexual, prokaryote and eukaryote. I also show how the K/θ
108 ratio, first used to discover and delimit EG species in eukaryotes based on DNA sequence
109 data from a single gene, can be applied to bacteria if a core protein-coding gene is used.
110 In particular, it could be applied to large-scale diversity surveys of bacteria using such a
111 gene. Such surveys are routine for eukaryotes, and can now be done with bacteria
112 regardless of whether they can be cultured in the lab, thanks to the development of
113 methods that allow amplification of genes from single bacterial cells isolated by dilution
114 or flow cytometry from environmental samples (e.g. [11,12]]. The successful
115 identification of evolutionary genetic species in bacteria using the K/θ ratio (this paper)
116 as well as methods such as the Generalized Mixed Yule Coalescent (GMYC; [4,13]) adds
117 to the growing literature that demonstrates that EG species are a general phenomenon
118 found in varying degrees to all organisms, sexual and asexual, prokaryotes and
119 eukaryotes.
120
121 The K/θ test is designed to distinguish between two kinds of clades: (i) those
122 separated by shallow gaps formed within a single species by random genetic drift or
123 transient physical isolation, and (ii) those separated by deep gaps produced by adaptation
124 to different ecological niches and/or by long-term physical isolation, i.e. different species.
125 Simulations [14] and population genetic theory [15] show that 95% of the gaps between

126 clades within a species are expected to have a depth of < $4N_e$ generations, while gaps
127 between species clades can be of any depth. It is thus possible to recognize a pair of
128 clades as different species, with probability ≥ 0.95, if they are separated by gaps of ≥ $4N_e$
129 generations. At that point the sequence difference between the two clades is $K \geq 8N_e\mu$
130 and the ratio K/θ ≥ 8Neu/2Neu ≥ 4 (3-8). (Barraclough's [16] claim to have found errors
131 in my use of K/θ is erroneous, as I show below in Discussion.)
132
133 This result is based on expected probabilities and does not account for uncertainty
134 introduced by finite sample sizes. The relevant sampling theory has been developed by
135 Noah Rosenberg [17, 18], who showed that it holds for any pair of sister clades, i.e.
136 reciprocally monophyletic samples n1 and n2, provided the sample sizes are at least n1 =
137 3 and n2 = 2. The power of these very small samples may seem surprising, but it arises
138 from their first being shown to be reciprocally monophyletic.
139
140 To illustrate how the K/θ ratio is used to detect evolutionary genetic species in
141 prokaryotes, I use the set of sequences of six different housekeeping/core genes
142 previously used by Wertz et al. [19, 20] to delimit species in enteric bacteria. This was
143 one of the first multilocus data sets used to discover bacterial species. It provides an
144 opportunity to apply the K/θ method to single genes and compare the results of using
145 different genes to each other and to the named taxonomic species defined by classical
146 phenotypic traits. Note that these data are used simply to demonstrate the application of
147 the K/θ method and are not intended to be definitive statements about these specific
148 bacterial species.

Materials and Methods

John Wertz kindly provided partial sequences of six housekeeping core genes (*gapA*, *groEL*, *gyrA*, *ompA*, *pgi*, and *16S* rRNA) from seven named species of enteric bacteria (*Citrobacter freundii* (abbreviated CF), *Enterobacter cloacae* (EB), *Escherichia coli* (EC), *Hafnia alvei* (HA), *Klebsiella oxytoca* (KO), *Klebsiella pneumonia* (KP), *Serratia plymuthica* (SP), and the outgroup *Vibrio cholera* (VC).

The aligned sequences for each gene were visualized in MacClade [21] where all the sequences of the gene were trimmed to the same length and all internal sites with gaps in any one sequence were removed. Consequently, all pairwise sequence differences for a given gene are based on the same regions of the gene. Final sequence lengths were: *gapA*, 540 bp; *groEL*, 348 bp; *gyrA*, 634 bp; *ompA*, 433 bp; *pgi*, 242 bp; *16S*, 289 bp. One internal gap of 27 bp in the *ompA* gene of the outgroup *V. cholerae* was retained because to remove it from all sequences would cause an unnecessary loss of sequence information in the ingroup.

To discover species using the trimmed sequences, I followed the procedure outlined by Birky [8] for each gene separately. Briefly,

1. First, I used PAUP* [22] to find a bootstrapped neighbor-joining tree in which all clades were supported by ≥70% of 1000 bootstrap replicas. A Neighbor-joining tree was used to rearrange the sequences in the PAUP* file to reflect the order of sequences in the tree; this greatly simplifies subsequent calculations in Excel, but note that the tree itself is not used in subsequent calculations. Both corrected and uncorrected sequence

differences were then exported from PAUP* to an Excel spreadsheet where all subsequent calculations were performed. Starting at the tips of the tree, each pair of sister clades is tested separately in steps 2-7, to determine the probability that the clades are samples from independently evolving species. This process is continued with deeper and deeper clades until species are found. For clades that are members of a non-bifurcating tree structure such as a polytomy (A, B, C) or ladder ((A,B)C), I compare A and B first, then compare C to whichever one of those is closest to C.

2. For the two candidate clades in a pair, the mean pairwise difference *d* between sequences is multiplied by the sample size correction n/(n-1) where *n* is the number of sequences in the clade, to get the nucleotide diversity π. Note that nucleotide diversity is defined as a function of the uncorrected pairwise distance. When $d = 0$, I used a non-zero estimate of π by assuming that one sequence differs from the others by 1/L where L is the sequence length; then $\pi = 2/Ln(n-1)$.

3. For each clade, I then estimate $\theta = 2Ne\mu$ by $\pi/(1 - 4\pi/3)$.

4. Calculate K = mean pairwise sequence difference between the two clades, corrected for multiple hits.

5. Calculate K/θ for the pair of clades. When the values of θ for the two clades differ, I use the larger value to get a conservative estimate of K/θ. If $K/\theta \geq 4$, I conclude that the individuals in those clades came from different species, with probability ≥ 0.95.

6. The preceding steps are based on expected probabilities and do not account for uncertainty introduced by finite sample sizes. This can be done as follows, using results obtained by Noah Rosenberg [17]. Using the K/θ ratio and the numbers $n1$ and $n2$ of individuals in the two clades, we want to find the probability that the individuals were

195 sampled from populations that have been evolving independently long enough to become
196 reciprocally monophyletic. This is best done using a table available on request from Noah
197 Rosenberg or me; alternatively, the values can be estimated from Figure 6 in Rosenberg
198 [17].
199
200 It is worth repeating that the phylogenetic tree used in step 1 is only used for two
201 purposes: (1) to facilitate the identification of sister clades, and (2) to inform the
202 organization of the Excel spreadsheet of pairwise sequence differences with the most
203 similar sequences together.
204

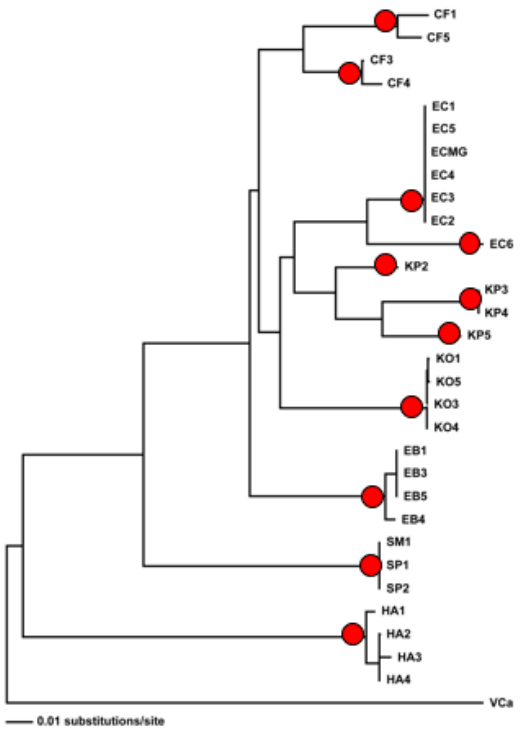
205 Results

206 The Wertz et al. [20] data consist of 5 protein- coding genes plus part of the 16S
207 RNA gene. I will first illustrate the application of the K/θ ratio by showing the detailed
208 analysis for *pgi*, the protein-coding gene that resolves the most species. The high
209 resolution suggests that this gene segregated before most of the other genes in most of the
210 speciation events, by chance. I will then summarize the results obtained with each of the
211 other genes. Following Wertz et al. [20], I consider the single isolate of *S. marcescens* as
212 belonging to the same taxonomic species as *S. plymuthica* because for all the genes in this
213 study it is identical or nearly so to the *S. plymuthica* isolates.

214 Finding species with *pgi*

215 The Neighbor-joining (NJ hereafter) tree in Figure 1 shows that the uncorrected
216 *pgi* sequences form a number of clades that supported that are seen in ≥ 63% of 1000
217 bootstraps; most have 100% support. The taxonomic species *K. oxytoca*, KO; *E. cloacae*,

218 EB; *H. alvea*, HA; and *S. plymuthica* SP, SM all form clades with 100% bootstrap
219 support, but some of these in turn contain two or more strongly supported clades. I
220 applied the K/θ test to each well-supported pair of sister clades, starting at the bottom of
221 the tree in Figure 1. These clades are also well-supported in the NJ tree using corrected
222 sequence differences. The Excel table for *H. alvea* and *S. plymuthica* is shown in Table 1
223 as an example of the data.



224 .

225 Fig 1. NJ tree of *pgi* sequences. CF, EC, KP, KO, EB, SM, and HA are taxonomic
226 species. Closed circles indicate clades that are different EG species delimited by the
227 K/method.

228 1. The Excel table for *pgi* from isolates of *H. alvea* and *S. plymuthica* is shown in
229 Table 1. The full tables for all genes are available on request from the author. To
230 illustrate the procedure, I begin with clade HA2,4, which has weak support (69%
231 of bootstraps; Its sister group is the singlet HA3. Using the formula for a

maximum estimate of π and θ when d = 0, I calculate that θ = 0.00836 and from the table K = 0.0021 so K/θ = 0.25 and these are not different species. This result was obvious from inspection and normally would not warrant a calculation. Next I ask if clade HA1-4 is a different species from clade (SM1,SP1,SP2). The pairwise distances for SM1, SP1, and SP2 are zero while those for HA1-4 are not all zero; I prefer to use the latter because they do not require any correction. For this clade d = 0.00625, π = 0.00833, and θ = 0.00843. K is 0.03032 so K/θ = 36.0 and these are samples from different clades with probability > 0.988. Because K/θ >> 4, these specimens are also samples from different independently species with P >> 0.95. Note that these results are obvious by inspection of the branch lengths in an NJ tree like that in Figure 1 in which branch lengths are proportional to the number of substitutions. The branches connecting the (SM,SP) clade to the HA clade are clearly more than four times the length of the branches connecting any two specimens within the HA clade.

Table 1. Matrix of pairwise sequence differences for *pgi* from *Serratia plymuthica* and *Hafnia alvea*.

	SM1	SP1	SP2	HA1	HA2	HA3	HA4
SM1	-						

SP1	0.0000	-					
SP2	0.0000	0.0000	-				
HA1	**0.3022**	**0.3022**	**0.3022**	-			
HA2	**0.3010**	**0.3010**	**0.3010**	0.0083	-		
HA3	**0.3087**	**0.3087**	**0.3087**	0.0125	0.0042	-	
HA4	**0.3010**	**0.3010**	**0.3010**	0.0083	0.0000	0.0042	-

Uncorrected sequence differences d and corrected differences K (bold type).

2. The taxonomic species *C. freundii* (CF) forms two clades, CF1,5 and CF3,4, each with 100% bootstrap support. Using Excel, I calculate the mean corrected sequence difference between these clades, K = 0.07618. CF1,5 has the larger value of $\theta = 0.0437$, giving $K/\theta = 1.74$. In Rosenberg's table, when $K/\theta = 1.7$ and the sample sizes are 2 and 2, the probability that these two clades are samples from reciprocally monophyletic populations is only 0.57. Moreover, because K/θ is much less than 4, the probability that the samples came from different species is low. This is the more conservative test in terms of splitting species because it uses the larger of the two values of θ. On the other hand, the clade CF3,4 has the smaller $\theta = 0.00167$, giving $K/\theta = 4.5$. From Rosenberg's table, $P = 0.97$ that these are samples from different clades; furthermore, because $K/\theta > 4$, the probability is very high that they are from different independently evolving species. However, this is based on the less conservative estimate of θ. Because only one of the two possible estimates of θ support the hypothesis of different species, I conclude that the *pgi* gene shows weak support for two cryptic species within *C. freundii*. As we will see below, some, but not all, other genes strongly support splitting. This suggests that there is

269 enough recombination in *C. freundii* to cause different genes to complete segregation at
270 different times, by chance.
271 3. The taxonomic species *E. coli* (EC) is also consists of a clade (EC1-5,ECMG) with
272 100% bootstrap support and a singleton EC6. Only the clade can be used to estimate θ,
273 but since the 6 sequences in the clade are identical, π and thus θ cannot be calculated
274 directly. For these cases, one can conservatively estimate these values by assuming that
275 one of the 6 sequences actually differs from the others at a single site. Then using the
276 calculation given in Materials and Methods, one can calculate $\pi = 0.001377$, $\theta = 0.00166$,
277 and $K = 0.07000$ so $K/\theta = 42.3$ and there is strong evidence that the clades EC1-5,MG
278 and EC6 are samples from different clades and from different cryptic species. This
279 procedure is likely to overestimate π and θ and hence underestimate K/θ, so in this sense
280 it is conservative with respect to splitting species.
281 4. Next we consider the clade consisting of *pgi* from KP isolates. Starting at the tips, I
282 compared (KP3,4) with KP5. Using the same procedure as above to get a non-zero value
283 for π, $K/\theta = 8.38$, providing evidence that these are samples from two different clades
284 and species. Comparing (KP3,4) to the other singleton, KP2, I find $K/\theta = 11.6$ and these
285 are also from different species. Thus KP is divided into three species by these gene
286 sequences.
287 5. Finally, the sequence data from this gene confirm that the isolates identified as *K.*
288 *oxytoca* (KO) and E. cloacae (EB) are evolutionary genetic species.
289 To summarize, the sequences of *pgi* confirm that KO, HA, EB, and SM,P are
290 species; split EC into 2 species; and split KP into 3 species.

291 **Finding species with *groEL***

The *groEL* gene (Figure 2) confirms that KO, EB, SM, and HA are species. It splits CF into two species, CF1,5 and CF3,4. EC is confirmed as a separate species from its sister clade CF2-4 if one uses θ from CF2-4 , but not if one uses θ from EC. Although *groEL* sequences split KP into KP1,4 and KP2,6 if one uses θ from KP2,6 they do not if one uses the larger θ from KP1,4.

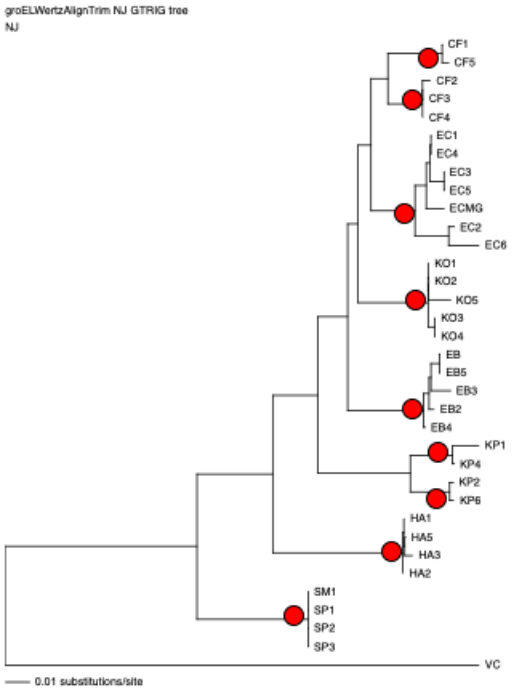
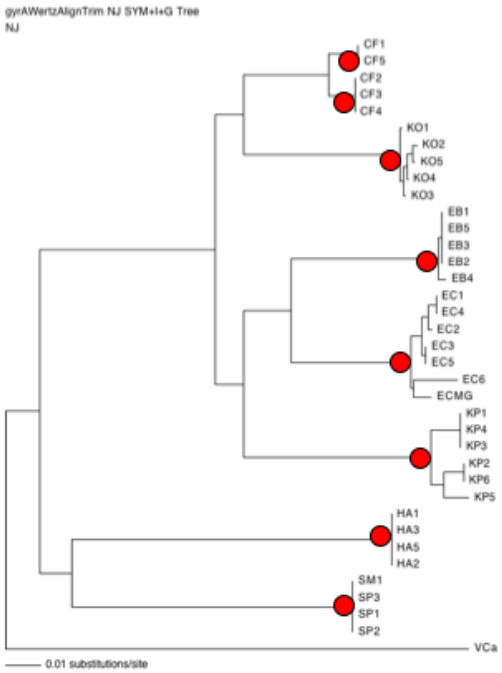


Fig 2. NJ tree of *groEL* sequences. CF, EC, KO, EB, KP, HA, and SM,SP are taxonomic species. Closed circles indicate EG species.

Finding species with *gyrA*

The gene *gyrA* (Figure 3) confirms that KO, EB, EC, HA, and SM are Evolutionary Genetic Species. The sequences of this gene split the taxonomic species CF into two EG species, CF1,5 and CF2-4, and also split KP into KP1,4 and KP2,6.



304

305 Fig 3. NJ tree of *gyrA* sequences. Closed circles indicate EG species.

306 **Finding species with *gapA***

307 The *gapA* gene (Figure 4) confirms that CF, KO, EC, SM,SP, EB, and HA are

308 different species. KP consists of two clades, KP1,3,4 and KP2,6, and a singlet KP5 which

309 are samples from three different species.

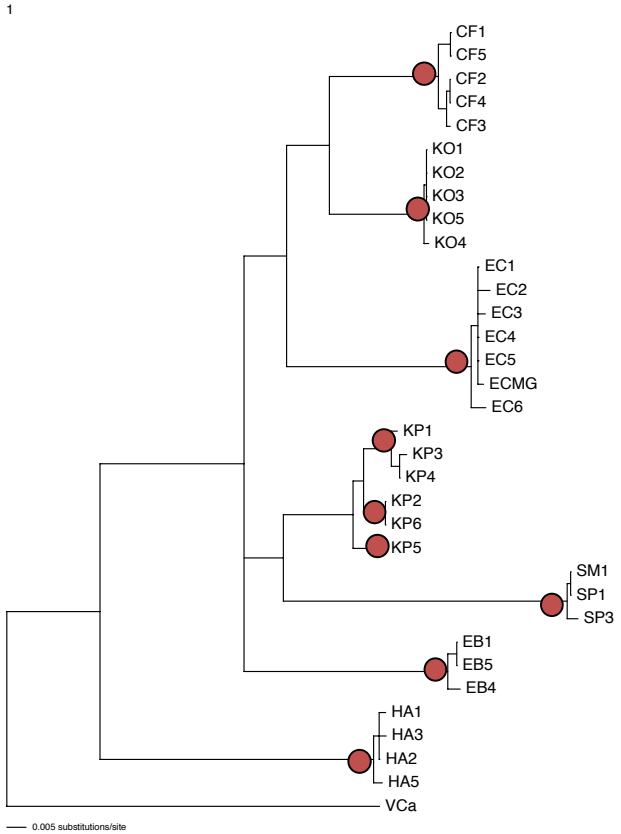


Fig 4. NJ tree of *gapA* sequences. Closed circles indicate EG species.

Finding species with *ompA*

The *ompA* gene (Figure 5) provides evidence that the CF, EC, EB, HA, and SP,SM sequences each represent a single species. Unlike some of the other genes, it does not split any of these species. In contrast, the KO and KP sequences are polyphyletic. Although the KO sequences are very similar with a diversity commensurate with that of other species in these data, the KP sequences are very divergent from each other. Moreover, KP5 is basal to KP1-4 as well as to KO; thus KP is polyphyletic, suggestive of incomplete lineage sorting.

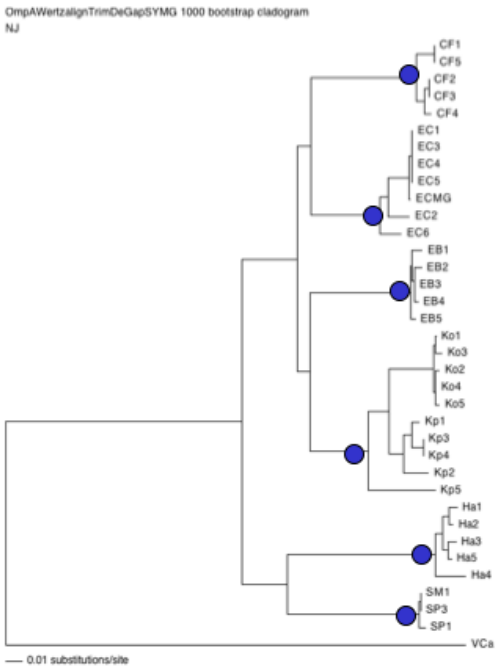


Fig 5. NJ tree of *ompA* sequences. Closed circles indicate EG species (these will be changed to red to match the other figures).

Finding species with *16S*

The *16S* sequences of the taxonomic species *E. coli* and *H. alvia* contained regions that were not present in the other species and for which I could find no explanation, so these two taxonomic species were omitted from the analysis of the *16S* sequences. Also, all pairwise differences among the ingroup sequences were ≤ 0.063, including those between species, so no corrections for multiple hits were applied. The K/θ test verified that the CF1,5, KO, CF2-4, EB, and KP sequence sets, and the SP1 singlet, each represent evolutionary genetic species with probabilities >0.99. However, the CFl and CF5 sequences did not; the clade was supported by only 56% of bootstrap replicas, so no K/θ test was done. The NJ tree is shown in Fig 6.

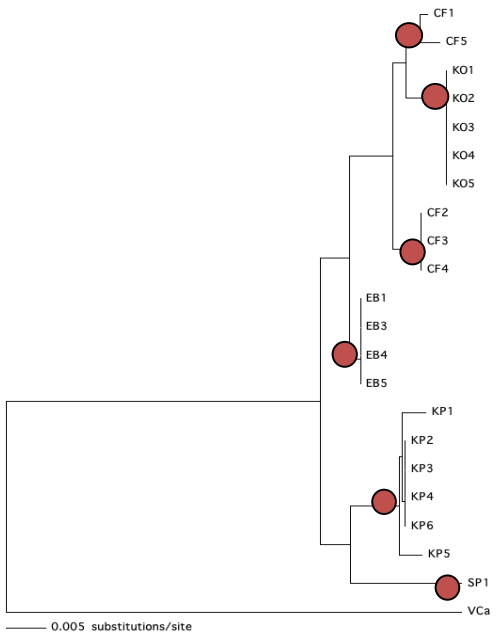


Fig 6. NJ tree of *16S* sequences. Closed circles indicate EG species.

Discussion

A huge literature has accumulated on the nature of species and how they should be defined and discovered (reviewed in [23,24]. Even the very existence of species has been debated, largely because the process of speciation is usually a gradual one in which two populations of one species diverge from each other until they reach the point where they can be recognized as different species. Nowhere has the debate been more intensive than in the prokaryotes. Although the Bacteria and Archaea came later to the party than the Eukaryota, they brought their own entertainment with them, in the form of novel means of sharing genes with close relatives (transformation, transduction, and conjugation). Prokaryotes can even receive new kinds of genes from distant relatives (Horizontal Gene Transfer, hereafter HGT).

My focus here, and in several preceding papers on discovering species in asexual and sexual eukaryotes [3-8] is on two problems: (i) identifying a model of species, i.e. a species concept, that is general enough to cover all organisms, whether prokaryotes or eukaryotes, and all degrees and kinds of sexuality; and (ii) based on the model, testing a general method of assigning specimens to the same or different species (a species criterion), using gene sequences.

16SrRNA genes are found in all organisms and evolve slowly enough to be useful in constructing phylogenetic trees across the tree of life. They have also been used to define Operational Taxonomic Units in bacteria, especially in environmental sequencing projects. Unfortunately, the same evolutionary conservatism that makes rRNA genes useful for deep taxonomy also often means that they have too little variation within and between closely-related species to be useful for species delimitation [25, 26]. Using 16SrRNA sequences from nearly the same set of isolates as in the present study, Riley *et al*. [19, 20] found that all of the taxonomic species formed clades with >50% bootstrap support in a maximum likelihood tree, except CF1,5 which was polyphyletic. My results with the 16S sequences differ; CF2-4 form a species distinct from CF1 and CF5, which are singlets.

Eckert and colleagues [26] used nearly complete 16S sequences to look for a barcode gap in Cyanobacteria, using the Automatic Barcode Gap Discovery (ABGD) program to find gaps in pooled sequences from some but not all taxonomic groups.

However, rather than testing pairs of sister clades, Eckert et al. pooled between 34 and 2448 sequences from each of 16 genera. This will obscure the barcode gap if different pairs of species have gaps in different places, i.e. if some species have values of θ that overlap the values of K in other groups. It is also possible that some of the groups examined by Eckert et al. have undergone a recent adaptive radiation such that K values are very small, as the authors noted.

The results with the protein-coding genes and isolates EB, EC, HA, KO, and SM/P are straightforward: each of the genes shows that these five taxonomic species, discovered by traditional methods of bacterial taxonomy using morphology and various phenotypic tests, are also EG species. The situation with CF is more complex. All five genes show two clades, (CF1,5) and (CF2-4) with 89-100% bootstrap support except *gapA* where the support for the (CF2-4) clade is only 64%. Moreover the K/θ ratio for this pair of clades is greater than 4, signifying different species, for *gyrA*, *pgi*, and *groEL* but not for *gapA* or *ompA*.

Deep gaps in trees separating species: how deep is deep enough?

In the literature of new species descriptions, and in papers proposing the splitting of species, it is common to refer to deep gaps between clades in phylogenetic trees as evidence that the clades should be assigned to different species. The decision as to whether the gap between two clades A and B is deep enough to assign A and B to different species, as opposed to indicating a relatively shallow gap within a single species, is generally made using the systematist's intuition. Occasionally the systematist

394 compares the gap between A and B to the gap between named species C and D,
395 concluding that A and B are different species if the gap between them appears to be as
396 deep or deeper than the gap between C and D. However, this assumes that species C and
397 D are well defined. The K/θ method avoids making these questionable assumptions by
398 providing a decision criterion that is soundly based on population genetic theory.

399 Not All Well-Supported Clades Are Species

400 These analyses illustrate the danger of assuming that all clades are species. For
401 example, in Fig 1 the clade EB1,3,5 has 90% bootstrap support and yet belongs to the
402 same species as EB4. If one studied *E. cloacae* EB only, one could easily draw a tree
403 with the Y axis stretched out, so that the branch connecting the clade to EB4 looks very
404 long and might appear to show two species connected by a long branch.

405 Applicability of the K/θ method

406 It is clear that the Evolutionary Genetic Species Concept is applicable to bacteria.
407 It is also clear that within the framework of the EGSC, species can be discovered using
408 DNA sequences analyzed with the K/θ method. It is important to note that the K/θ test
409 was deliberately designed to work with a single gene or sequence. This is appropriate
410 because sequencing multiple genes from every specimen is too expensive and time-
411 consuming for routine application in species discovery, and will continue to be too
412 expensive in the foreseeable future. It will be vastly more difficult, and often impossible,
413 in the case of environmental sequences, which are of necessity not tied to individual
414 organisms. If the sequences of two or more genes are used, it is important to remember
415 that in a sexual organism different genes may become reciprocally monophyletic at
416 different times. I recommend that the K/θ test be applied to each gene separately; if any

gene places the specimens in two different clades and $K/\theta \geq 4$, this would be strong evidence that the clades represent different species. I do not recommend concatenating gene sequences without first testing whether the signal from a gene that has not yet segregated can obscure the signal from a gene that has become reciprocally monophyletic, i.e. could cause the specimens to be incorrectly assigned to a single species.

Multiple comparisons

I used a probability of 0.05 for hypothesis testing, understanding that when this is applied to many different cases, I expect to make a wrong decision in approximately 5% of the cases. This could be accounted for by using a correction such as the Bonferroni. On the other hand, many of the the results are significant with a much higher probability. I choose not to use a correction and simply accept the risk of a small number of wrong conclusions, less than one in twenty for this work as a whole.

When small sample sizes can provide strong conclusions

Rosenberg [17] showed that small sample sizes can provide strong evidence that the entire populations from which they were taken are reciprocally monophyletic. This is because the probabilities are contingent upon the samples themselves being reciprocally monophyletic. Of course this assumes that the samples are representative of the populations, and do not, for example, come from a local sub-population that happens to be very different from the rest of the population. Thus one must always bear in mind the possibility that the specimens being studied are not representative. On the other hand, the possibility of erroneously splitting a species into two species based on a non-representative sample is likely to be small when K is very large. This is because, if a

440 large K gives a K/θ value less than 4, the θ value must also be much larger than the
441 generally observed values of < 0.05. There are parallels in traditional taxonomy, where
442 new species are often described on the basis of one or a few specimens, especially when
443 the new specimen is very different from the most similar described species. In this case
444 we usually assume that additional new specimens will not have intermediate phenotypes
445 that will cause us to combine the new specimen with a described species. If we had a
446 single specimen of an African lion and two specimens of an ocelot, traditional taxonomy
447 would probably recognize them as different species based on the absence of intermediate
448 phenotypes, even without additional evidence. If gene sequences from these specimens
449 had K/θ ≥ 4, this would indicate an absence of intermediate sequences and we would
450 have similar objective evidence that they were different species.

451 Use of trees and choice of tree-making methods

452 When implementing the K/θ method, the use of phylogenetic trees is limited to (i)
453 finding sister pairs of clades for comparison; (ii) guiding the arrangement of sequences in
454 the Excel spreadsheet to facilitate the calculation of K/θ; and (iii) spotting obvious cases
455 of K/θ based on branch lengths. I use Neighbor-joining for tree construction because it is
456 fast and easy to do with PAUP*, but presumably a different distance-based method could
457 be used.

458 Genes or phenotypes in bacterial species descriptions?

459 The bacterial rules of nomenclature do not specifically require phenotypic data.
460 However, the report of an *ad hoc* committee on species definition in bacteria [27] calls
461 for the use of qualitative phenotypic traits as well as 16S rDNA sequences to delimit

462 species. A quick survey of some recent papers in the *International Journal of Systematic*
463 *and Evolutionary Microbiology* shows that while rDNA sequences are now commonly
464 used in descriptions of new bacterial species, they are used essentially like a collection of
465 traits or to form phylogenetic trees. When these trees, or pairwise sequence differences,
466 or both are used to define trees, there is usually no attempt to apply a theoretically
467 justifiable method to distinguish sequence differences between species from differences
468 within species. Instead, this is based on the judgment or intuition of the investigators.

469 One gene or many?

470 There is a tendency in the recent literature to extoll the virtues of using many
471 genes to discover species. But using more than one gene has several drawbacks. First, it
472 costs more and will not be useful if one gene is sufficient to show that two samples come
473 from different species. Second, if one gene is a core gene, adding another gene which is
474 not a core gene and is involved in LGT can obscure the signal of speciation.

475 Ideally, to discover closely related species we would like to use the gene or genes
476 that segregated earliest in those particular populations; in this way we would catch
477 speciation in its early stages and discover closely related species. This is a real advantage
478 to using multiple gene sequences, which would increase the chance that at least one of
479 them would have segregated early. How important this depends on the amount of sexual
480 reproduction in the bacteria in question. At one extreme, if there is little or no
481 recombination, or extreme inbreeding, all genes will segregate at the same time and any
482 core gene should be as good as any other. At the other extreme, if there is extensive
483 outcrossing and recombination, different genes may segregate at very different times after
484 two populations begin evolving independently. In fact, a given gene may segregate early

485 in one such population and late in the sister population. This is a likely explanation for
486 the observation that CF1 and CF5 form a separate species from CF2, 3, and 4 for all
487 protein-coding genes except *ompA*. Stackebrandt et al. [27] speak favorably of using
488 protein-coding genes, claiming that there is a consensus that at least five such genes are
489 necessary, but there is no such consensus in the sense of a number that is agreed on or at
490 least accepted by nearly everyone in the field. More importantly, Stackebrandt et al. do
491 not consider the possibility that these genes may be at different stages of segregation,
492 which is the simplest explanation of the different trees supported by different genes
493 Wertz et al. data.

494 There is also a potential problem with concatenated gene sequences: a significant
495 K/θ ratio in one gene may be obscured by a small, insignificant K/θ ratio in another gene.
496 In other words, the signal of species divergence in one gene that segregated early after
497 speciation may be swamped by the signal in another gene that has not yet completed
498 segregation with 95% confidence.

499 I suggest that the best method for large-scale diversity studies would be to use a
500 single protein-coding gene to discover species and then, for groups of special economic,
501 medical, ecological, or evolutionary importance or interest, test additional genes to see
502 they separate more species pairs because they segregated earlier in those pairs. Then it
503 would often be useful to look for phenotypic or molecular markers that could be used to
504 identify those species without requiring sequencing.

505 Stability of EG species

506 A nice feature of species discovery using $K/\theta \geq 4$ is that additional specimens are
507 unlikely to fill in the gap between species and require them to be lumped. That is because

508 the large difference between K and θ reduces the probability that other specimens that are
509 intermediate between the two species and fill in the gap between them will ever be
510 discovered. Put another way, it is unlikely that we will find specimens and lineages that
511 join the two clades we identify as species. If this did happen, the single species produced
512 by joining the two clades would have exceptionally large nucleotide diversity. It seems
513 likely that, once two populations are assigned to different species using $K/\theta \geq 4$, those
514 species are unlikely to be lumped in the future. It is perhaps more likely that data from
515 additional genes or from new specimens will cause a species to be split in the future, and
516 this has been seen repeatedly in the literature. But this is true for species delimited by any
517 method, using any kind of data; in science, new information often forces us to change our
518 conclusions.

519 Other methods used to discover bacterial species with single 520 loci

521 Fixed cutoffs of 95%, 99%, or other levels of sequence similarity in rRNA genes
522 are widely used to delimit species or potential species in the bacterial literature. These
523 cutoffs share with $K/\theta \geq 4$ the advantage of allowing one to assign sequences from
524 uncultivated or uncultivatable bacteria, or environmental sequences, to species. They are
525 based mainly on the observed levels of sequence similarity among already described
526 species, and as such they are only good to the extent that the described species are real.
527 Moreover, cutoffs are generally assumed to apply to uniformly over all species, but this
528 assumes that all species have the same within-species variation, i.e. the same values of
529 nucleotide diversity; this is not true.

530 Most of the methods proposed for using DNA sequence data to delimit species are
531 global methods, i.e. in a phylogenetic tree of gene sequences from many specimens, they
532 look for the transition point from species divergence to divergence between individuals
533 within species. One of these, GMYC [13] uses coalescent theory to detect the transition
534 point. Barraclough et al. [28] used GMYC to discover species from environmental 16S
535 rDNA sequences of bacteria but the method has not been applied to sequences from
536 already-described taxonomic species of prokaryotes. Poisson Tree Picking (PTP) [29]
537 attempts to identify significant changes in the numbers of substitutions along branches,
538 which are expected to be significantly higher between species than within species. To
539 my knowledge, PTP has not been applied to prokaryotes.
540 Ecosim [30] detects species produced by selective sweeps during adaptation to
541 different habitats. It assumes a Poisson model of substitutions, in contrast to K/θ and
542 GMYC that allow more complex and presumably more realistic models to be used. The
543 theory behind Ecosim includes random drift and selective sweeps but not less drastic
544 forms of selection. AdaptML [31] is another local method. It identifies clades associated
545 with ecological parameters that might be relevant to ecological species. It detects species
546 only if they are adapted to different niches that have already been identified. These
547 methods will not detect species that evolve independently because they are physically
548 isolated as in the case of *Sulfolobus islandicus* in hot springs [32] unless they are also
549 adapted to different niches. In other words, they may fail to detect some evolutionarily
550 independent species.
551 These global methods are all subject to error if different species have different
552 levels of within-species variation, i.e. different values of nucleotide diversity. This is in

553 contrast to $K/\theta \geq 4$ which is a local method focusing on sister clades and thus on putative
554 sister species. Such pairwise comparisons have the advantage that sister species are more
555 likely to have similar values of θ. In fact, the problem of delimiting species in general
556 really is a local problem. We have no problem assigning specimens of the grizzly bear
557 and the mule deer to different species, but it is more difficult to decide whether the
558 Alaskan brown bear and polar bear are really different species, e.g. [33].

559 Microscopic organisms do have species

560 Rossberg et al. [34] claimed that organisms smaller than about 1mm can not have species,
561 which the present paper clearly refutes for bacteria. Rossberg *et al.* focused on
562 meiofauna, but their computer simulations apply equally to bacteria as well as to asexual
563 and sexual eukaryotes. Remarkably, they failed to cite any of the nearly two centuries of
564 taxonomic literature describing species in microscopic invertebrates, protists, algae,
565 fungi, etc. A likely source of their erroneous conclusion is their use of carrying capacities
566 instead of the effective population size, which is appropriate for the critical parameter
567 $N_e\mu$ and is very much smaller than the carrying capacity. In fact, their Figure 3 shows the
568 signal of species when they used values of the carrying capacity that are more in line with
569 known effective population sizes.

570 Limitations of the K/θ method

571 In order to study the diversity of cyanobacteria from hot springs in Yellowstone
572 National Park, Ward et al. [35] did environmental sequencing of clone libraries from the
573 16SrRNA/ITS region from PCR-amplified DNA. Their Figure 4 shows a neighbor-
574 joining tree of ITS sequences of bacteria identified as *Synechococcus*. They found 10
575 putative ecotypes of *Synechococcus*, which are thus different species adapted to different

576 niches. However, it is evident by inspection of the tree that only the three clades called A,
577 A', and B satisfy the K/θ ≥ 4 criterion. This could happen if most of the ecotypes have
578 diverged too recently to pass the the K/θ ≥ 4 criterion for independent evolution. In this
579 case ecotype analysis can divide species more finely because it delimits species at an
580 earlier stage in the speciation process.

581 Using K/θ in Formal Descriptions of New Species

582 I have shown that the K/θ ratio can be used to delimit evolutionary species in
583 bacteria, illustrating the procedure with an exemplar data set. However, this does not
584 mean that the species so delimited conform to current practice in bacterial systematics, or
585 that the species so detected can be formally described in conformance with the current
586 rules of taxonomic nomenclature for bacteria. Of course the same is true of species
587 delimited with the GMYC method that also detects EG species, or with the method of
588 Cohan [36], which detects a subset of EG species that are separated by ecology.
589 Unfortunately, the rules of bacterial systematics codified in the International Code of
590 Nomenclature of Bacteria are not explicitly based in population genetic theory, even
591 though they pay lip service to the BSC which defines species as reproductively isolated
592 populations in outcrossing sexual organisms.

593 Response to the critique of Barraclough [37]

594 Barraclough [37] claimed to have found two errors in my calculation that a K/θ
595 value of 4 or greater is strong evidence that samples belonging to two different
596 evolutionary species. First, citing Edwards and Beerli [38], he claimed that my "K from
597 the genealogy overestimates the true divergence time, however, due to sampling of the

598 polymorphism from the ancestral common ancestor species" and that K is equal to $2T\mu$
599 plus the θ of the ancestor. However, Edwards and Beerli said that the addition of θ is
600 necessary when K denotes the time elapsed since two populations diverged, but it is not
601 appropriate when estimating the time T since allelic divergence, as I am doing here and in
602 previous papers using K/θ. Second, Barraclough refers to "a mitochondrial marker with
603 $\theta = N\mu$." (In correspondence, he said that in this section he used N to refer to the
604 effective population size rather than the census size.) If this were true, then the upper
605 95% confidence limit of K/θ would be 8. But by definition $\theta = 2N_e\mu$. This is because the
606 sequence difference between two gene copies in two specimens A and B from the same
607 species sharing a most recent common ancestor O is equal to the sequence difference $N_e\mu$
608 between A and O plus the sequence difference between B and O, which is also $N_e\mu$.
609 Consequently, after $4N_e$ generations, $K/\theta = 8N_e\mu/2N_e\mu = 4$, and not 8.

610 Could we use K/θ as a measure of the progress of speciation?

611 If we adopt the evolutionary species concept, then speciation is usually a gradual
612 process. (Exceptions can include rare dispersal to isolated sites such as islands, or
613 speciation by hybridization or polyploidization.) To the extent that speciation is gradual,
614 the K/θ ratio is expected to gradually increase. Consequently, if we have two populations
615 that we suspect are undergoing speciation, K/θ might be used as a measure of the
616 progress of speciation. It would presumably be non-linear, as suggested by the
617 nonlinearity of progress toward reciprocal monophyly of the two populations [17].

618 Species are universal

619 In this paper and previous papers in this series [3-8] I have shown that Evolutionary
620 Genetic species are found in bacteria and in both asexual and sexual eukaryotes. The
621 arguments in this paper should also apply to Archaea. I strongly suspect that viruses also
622 form such species, and that extraterrestrial organisms, if they exist, will likewise form
623 populations that evolve independently from each other.
624

625 # Acknowledgements

626 I thank John Wertz for providing the aligned sequences.
627

628 # References

629 1. Simpson GG The species concept. Evolution 1951; 5:285-298.
630 2. Wiens JJ. What is speciation and how should we study it? Am Nat 2004;163(6):914-
631 23.
632 3. Birky, CWJr, Wolf C, Maughan H, Herbertson L, Henry E. Speciation and selection
633 without sex. Hydrobiologia 2005; 546:29-45.
634 4. Barraclough TG, Birky CWJ, Burt A. Diversification in sexual and asexual
635 organisms. Evolution. 2003;57(9):2166-72.
636 5. Birky, CWJr, Adams J, Gemmel M, Perry J. Using population genetic theory and
637 DNA sequences for species detection and identification in asexual organisms PLoS
638 One 2010; 5:e10609.
639 6. Birky, CWJr, Ricci C, Melone G, Fontaneto D. Integrating DNA and traditional
640 taxonomy to describe diversity in poorly studied microscopic animals: new species of

641 the genus Abrochtha Bryce, 1910 (Rotifera: Bdelloidea: Philodinavidae). Zool J Lin
642 Soc 2011; 161:723-734.
643 7. Schön I, Pinto RL, Halse S, Smith AJ, Martens K, Birky CWJr. Cryptic species in
644 putative ancient asexual darwinulids (Crustacea, Ostracoda). PLoS One 2012;
645 7:e39844.
646 8. Birky CWJr. Species detection and identification in sexual organisms using
647 population genetic theory and DNA sequences. PLoS One 2013; 8:e52544.
648 9. Lan R, Reeves PR. Intraspecies variation in bacterial genomes: the need for a species
649 genome concept. Trends in Microbiology. 2000;8(9):396-401.
650 10. Lan R, Reeves PR. When does a clone deserve a name? A perspective on bacterial
651 species based on population genetics. Trends in Microbiology. 2001;9(9): 419-24.
652 11. Otteson EA, Hong JW, Quake SR, Leadbetter JR. Microfluidic digital PCR enables
653 multigene analysis of individual environmental bacteria. Science. 2006;314:1464-7.
654 12. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Andeerson IJ, Cheng J-F, et al.
655 Insights into the phylogeny and coding potential of microbial dark matter. Nature.
656 2013;499:431-7.
657 13. Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, et al.
658 Sequence-based species delimiation for the DNA taxonomy of undescribed insects.
659 Syst Biol. 2006;55:1-15.
660 14. Avise JC. Molecular Markers, Natural History and Evolution. New York: Chapman
661 & Hall, Inc.; 1994.
662 15. Hudson RR, Coyne JA. Mathematical consequences of the genealogical species
663 concept. Evolution. 2002;56:1557-65.

664 16. Barraclough TG. The Evolutionary Biology of Species. Godfray CJ, Dunne JA,
665 Sheldon BC, editors. Oxford, United Kingdom: Oxford University Press; 2019.
666 17. Rosenberg NA. The shapes of neutral gene genealogies in two species: Probabilities
667 of monophyly, paraphyly, and polyphyly in a coalescent model. Evolution.
668 2003;57(7):1465-77.
669 18. Rosenberg NA. Statistical tests for taxonomic distinctiveness from observations of
670 monophyly. Evolution. 2007;61(2):245-476.
671 19. Riley MA, Lizotte-Waniewski M. Population Genomics and the Bacterial Species
672 Concept. In: Gogarten MB, Gogarten JP, Olendzenski L, editors. Horizontal Gene
673 Transfer: Genomes in Flux. Methods in Molecular Biology. New York City: Humana
674 Press; 2009. p. 367-77.
675 20. Wertz JE, Goldstone C, Gordon DM, Riley MA. A molecular phylogeny of enteric
676 bacteria and implications for a bacterial species concept. J. Evol. Biol. 2003; 16:1236-
677 1248.
678 21. Maddison DR, Maddison WP. 2000. MacClade 4. Sinauer Associates, Sunderland,
679 Massachusetts.
680 22. Swofford DL. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other
681 Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts
682 23. Zachos FA. Species Concepts in Biology: Springer International Publishing;2016.
683 220 p. 32.
684 24. Wilkins JS. Species: University of California Press; 2011. 320 pp.
685 25. Tang CQ, Leasi f, Obertegger U, Kieneke A, Barraclough TG, Fontaneto D. The
686 widely used small subunit 18S rDNA molecule greatly underestimates true diversity

687 in biodiversity surveys of the meiofauna. Proceedings of the National Academy of
688 Sciences USA. 2012;109(40):16208–12.
689 26. Eckert EM, Fontaneto D, Coci M, Callieri C. Does a barcoding gap exist in
690 prokaryotes? Evidences from species delimitation in Cyanobacteria. Life. 2015;5:50-
691 64. doi: 10.3390/life5010050.
692 27. Stackebrandt E, Frederiksen W, Garrity GM, Grimont PAD, Kampfer P, Maiden
693 MCJ, et al. Report of the ad hoc committee for the re-evaluation of the species
694 definition in bacteriology. Int. J. Syst. Evol. Micr. 2001;52:1043-7.
695 28. Barraclough TG, Hughes M, Ashford-Hodges N, Fujisawa T. Inferring evolutionarily
696 significant units of bacterial diversity from broad environmental surveys of single-
697 locus data. Biol. Lett. 2009; 5:425-8.
698 29. Zhang J, Kapli P, Pavlidis P, Stamatakis A. A general species delimitation methode
699 with applications to phylogenetic plasements. Bioinformatics. 2013;29(22): 2869-76.
700 30. Koeppel A, Perry EB, Sikorski J, Krizanc D, Waarner A, Ward DM, et al. Identifying
701 the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology
702 into bacterial systematics. Proceedings of the National Academy of Sciences USA.
703 2008;105(7):2504-9.
704 31. Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource partitioning
705 and sympatric differentiation among closely related bacterioplankton. Science.
706 2008;320:1081-5.
707 32. Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ. Biogeography of the
708 *Sulfolobus islandicus* pan-genome. Proc Nat Acad Sci USA. 2009;l106:8605-10.

709 33. Cahill JG, Green RE, Fulton TL, Stiller M, Jay F. Genomic evidence for island
710 population conversion resolves conflicting theories of polar bear evolution. PLoS
711 Genetics. 2013;9:e31003345.
712 34. Rossberg AG, Rogers T, McKane AJ. Are there species smaller than 1 mm? Proc.
713 Roy. Soc. B. 2013;280:20131248.
714 35. Ward DM, Bateson MM, Ferris MJ, Kuhl M, Wieland A, Koeppel A, et al.
715 Cyanobacterial ecotypes in the microbial mat community of Mushroom Spring
716 (Yellowstone National Park, Wyoming) as species-like units linking microbial
717 community composition, structure and function. Phil. Trans. R. Soc. B.
718 2006;361:1997-2008.
719 36. Cohan FM. Towards a conceptual and operational union of bacterial systematics,
720 ecology, and evolution. Phil. Trans. R. Soc. B. 2006; 361:1985-86.
721 37. Barraclough TG. The Evolutionary Biology of Species. Godfray CJ, Dunne JA,
722 Sheldon BC, editors. Oxford, United Kingdom: Oxford University Press; 2019.
723 38. Edwards SV, Beerli P. Perspective: Gene Divergence, Population Divergence, and the
724 Variance in Coalescence Time in Phylogeographic Studies. Evolution.
725 2000;54(6):1839-54.