

15
16
17
18
19
20
21
22
23
24
25
26
27
28
29

Abstract

Confusion about strain classification and nomenclature permeates modern microbiology. Although taxonomists have traditionally acted as gatekeepers of order, the numbers of and speed at which new strains are identified has outpaced the opportunity for professional classification for many lineages. Furthermore, the growth of bioinformatics and database fueled investigations have placed metadata curation in the hands of researchers with little taxonomic experience. Here I describe practical challenges facing modern microbial taxonomy, provide an overview of complexities of classification for environmentally ubiquitous taxa like *Pseudomonas syringae*, and emphasize that classification and nomenclature need not be the one in the same. A move toward implementation of relational classification schemes based on inherent properties of whole genomes could provide sorely needed continuity in how strains are referenced across manuscripts and data sets.

30

31 **Confusion Abounds in Modern Bacterial Taxonomy**

32

33 Communication between researchers is a foundation of all scientific disciplines, and
34 clarity of the message is therefore essential for moving science forward. Alternatively, confusion
35 of underlying messages leads directly to systemic problems and disagreements. For modern
36 microbiologists perhaps the best example of how systemic confusion can slow research progress
37 involves ongoing disagreements about bacterial classification and nomenclature, a confusion
38 which is only amplified by the traditional entwinement of these two activities. The advent of
39 'big data' has placed microbiology at a crossroads where we can either systematically change
40 the way we think about describing strains or suffer within an ever expanding cloud of
41 uncertainty. Now is the time to divorce classification of strains from any discussions about
42 nomenclature based on species concepts and transition to a system based on genomic
43 information, at least for metadata entry and to ensure continuity across manuscripts.

44

45 The root of this article lies in a frustration that many researchers deal with every day. A
46 frustration born out of the clashes between how taxonomy should proceed in theory and the
47 realities of how it proceeds in practice. Although nomenclatural confusion has always
48 inconvenienced microbiology, the speed and focus of research as well as the dedication of large
49 numbers of taxonomists previously enabled back and forth dialogues to smooth over ongoing
50 disagreements. However, traditional taxonomic schemes have not efficiently dealt with the
51 rapid influx of genomic data and were not designed to account for intrinsic challenges that arise
52 when non-taxonomists publish metadata. Researchers have been arguing about bacterial species
53 concepts since the dawn of microbiology [1-3], but the intent of this article is not to get caught
54 up in discussions about what constitutes a bacterial species nor is it to suggest that the perfect
55 mechanism for classification has been uncovered. My goal is to call attention to conflicts
56 between the philosophy and practice of bacterial taxonomy, which generate much confusion for
57 the classification of environmentally ubiquitous taxa like *Pseudomonas syringae*.

58

59 **The Philosophy of Bacterial Classification and Nomenclature**

60

61 Taxonomy is a branch of microbiology that consists of three fundamental and often
62 intertwined activities: identification, classification, and nomenclature of strains [4]. While these
63 words can often be thought of as synonymous, important yet subtle distinctions can be drawn
64 between these three areas. Whereas classification provides a means to index strains logically, it
65 can exist independently of studies of how to accurately identify or name particular groups of
66 strains. Therefore, taxonomy is not simply the study of organismal classification but a larger
67 field that deals with the implementation of classification for diagnostic or naming purposes.

68

69 Bacterial strain and species names should relay useful information about properties or
70 phenotypes generally shared by the groups that fall under defined nomenclatural umbrellas.
71 One main reason for applying Linnean species names to microbes is that these names represent
72 a shorthand for describing underlying aspects of the organism's biology while also informing

73 about classification. Farmers in the field or doctors in a hospital rely on accurate identification
74 of strains in order to make informed decisions about how to proceed. In an academic context,
75 although names are important for identification's sake, they also provide continuity throughout
76 manuscripts, and enable researchers to investigate their own systems and make general
77 predictions from other work. As science and technology progress, and many researchers
78 become beholden to large curated databases, metadata provides a way to rapidly access and
79 apply information across taxa and studies. For instance, if metadata were perfectly applied, one
80 need only search the words "*Pseudomonas syringae*" and all relevant references and datasets
81 would be retrieved. This is currently not the case, as described below, because such a search
82 would miss a variety of genomes and many publications that truly focus on *P. syringae*. The
83 importance of proper curation of metadata to research is illustrated by the development of lines
84 of investigation that employ text mining algorithms to discover emergent phenomena across
85 systems [5]. How this metadata is curated significantly affects continuity and communication
86 within manuscripts and across research groups.

87
88 In a perfect world, one would be able to see a name, garner something about the biology
89 of that particular strain, and know exactly how other strains within the same assigned species
90 name are related. For instance, *Helicobacter pylori* represents a cluster of strains which largely
91 only lives within human stomachs and which are causative gastric ulcers and cancer [6-8].
92 Classification and nomenclature for *H. pylori* strains is straightforward, despite extensive
93 diversity across strains, because the unique and specialized niche that this organism inhabits
94 enables tight grouping of phenotypic and phylogenetic clusters [9]. In other cases, phenotypic
95 traits of interest exist outside those normally used for classification so that additional modifiers
96 are added. *Escherichia coli* is a common inhabitant of vertebrate gastrointestinal tracts, but
97 certain modifiers are added to this species name to reflect important data about biology [such as
98 uropathogenic *E. coli* (UPEC) enterohemorrhagic *E. coli* (EHEC), etc...] [10]. Naming schemes
99 are by no means perfect and throughout the course of microbiology there have been situations
100 where nomenclature and classification have pointed in opposite directions. These cases are
101 useful because they highlight different requirements for microbial nomenclature between
102 professional and academic settings, and they demonstrate how single classification schemes can
103 function in one context but not the other. Even though all strains of *Shigella* cause similar
104 disease symptoms in humans, there have been multiple independent evolutionary origins due
105 to convergent evolution of phenotypes by different strains of *E. coli* [11]. This underlying
106 evolutionary convergence might not matter much in a hospital, because, all else equal, doctors
107 can prescribe similar treatments for all cases. However, conflation of different evolutionary
108 lineages under the same name could affect how results are interpreted across comparative
109 genomic studies unless researchers explicitly incorporate these phylogenetic nuances into their
110 designs. While this particular example of confusion was clarified quite quickly, largely because
111 of the importance of *Shigella* for human health, many other cases likely exist where
112 nomenclature is wrong or misleading but which will languish in obscurity due to the lack of
113 widespread interest.

114

115 **Nomenclature and Classification Schemes in Practice**

116
117 When thinking about bacterial taxonomy, one cannot set aside historical momentum
118 generated by the requirement of cultureability of strains in the early days of microbiology. The
119 first step for any nomenclatural decision is traditionally the establishment of a “type” strain that
120 is used to set a foothold for new species designations [12]. Following from cultureability,
121 bacterial types are binned by observable properties at micro and macroscopic scales. Features
122 such as staining with certain dyes (i.e. Gram stain), cell shape, flagellar type, nutrient profiles,
123 and pathogenicity on certain hosts are still used to assign newly isolated bacterial strains into
124 broad categories. One of the better known schemes today involves grouping of *E. coli* strains
125 based on their O and H antigens (e.g. O157:H7)[13], while *Salmonella* strains have are typed
126 according to their phage sensitivities [14]. Gradually, diagnostic assays have become
127 increasingly specialized and technologically advanced, enabling finely grained classification of
128 strains [15].

129
130 Although phenotyping schemes continue to work well in many cases, they can be time
131 consuming and often require specialized equipment and expertise. Classification methods that
132 use compositional properties of each genome, such as DNA-DNA hybridization, were
133 developed to overcome these challenges [4]. Indeed, genomic hybridization remains a gold
134 standard for species characterization. The creation of genotyping schemes was further
135 welcomed because genomes were thought to reflect inherent properties of each strain and were
136 potentially less prone to convergence or experimental variability. With the development of PCR,
137 random genotyping methods such as Rep-PCR enabled closer views into the genetic
138 relationships between or differentiation among strains and could be carried out at a lower cost
139 and with less expertise required than phenotypic methods [4]. The DNA sequencing revolution
140 sparked innovation of multiple layers of sequence based characterization, which ultimately
141 focused on targeted sequencing and phylogenetic comparison of the 16s rRNA gene conserved
142 throughout bacteria [16,17]. Paralleling the development of more specialized phenotyping
143 methods, more finely grained comparisons of DNA sequences, such as multilocus sequence
144 analysis (MLSA), promised increased resolution for classification within strains and species
145 [18]. Finally, decreased costs and increased reliability of whole genome sequencing now
146 provides an unprecedented amount of information for classification [3,4].

147
148 A crucial step for any formal taxonomic scheme involves some level of oversight, which
149 today takes the form of dedicated groups of researchers that diligently curate and vet bacterial
150 names [e.g. 19]. When a new bacterial species is proposed, vetting largely takes place by
151 publishing a characterization of the type strain in the *International Journal of Systemic and*
152 *Evolutionary Microbiology* [12]. Although there have been countless disagreements over what
153 constitutes a bacterial species, it has been possible for these experts to take the time to reconcile
154 the nomenclatural record. This system has held up quite well to scrutiny when phenotypes
155 were of dire interest (as is the case of human pathogens), when there were numerous paid
156 taxonomic positions whose purpose was to vet new entries into the nomenclatural arena, and
157 when researchers could spend the time to learn ‘a feel’ for their organisms. With funding

158 cutbacks, and millions of new strains being identified yearly, it is unclear how sustainable this
159 system will be going forward.

160

161 **Blindspots in Microbial Taxonomy**

162

163 The history of the field has established a legacy where many firmly believe that
164 phenotypic characterization should figure prominently in bacterial taxonomy [1,3,4]. This belief
165 has culminated in what is referred to as the polyphasic approach to classification and
166 nomenclature, where phenotypic characterizations are blended with genotypic and
167 phylogenetic information to place strains into nomenclatural groups. This reliance on
168 phenotypes for taxonomic purposes persists even though there are numerous cases where
169 phenotyping is impossible for certain strains, such as those that cannot be cultured [17]. Instead
170 of completely redesigning classification schemes to only incorporate molecular methods and
171 sequencing, exceptions to nomenclatural rules have been created to account for outliers. For
172 instance, when type strains cannot be cultured, a "Ca." is appended before the Linnean name
173 representing the idea that these are only candidate taxa which have not been validated
174 according to the established rules [12,20]. Furthermore, the ease of sequencing has
175 democratized the taxonomic process, and anyone that can perform PCR can generate new data.
176 New 'species' are discovered by researchers who are not necessarily focused on taxonomy and
177 have not been trained to classify these organisms. In these cases, when strains have not been
178 fully vetted and assigned proper species names or where there is no type strain, the tradition is
179 to simply place "sp." after the genus [12]. Lastly, with the increasing number of strains that are
180 sequenced and cultured, the more exceptions that arise to the established phenotypic rules for
181 any particular group. What may have traditionally fit into a nice taxonomic box based on
182 phenotyping methods may falter now, even for well described taxa, because exceptions or
183 complications frequently arise with greater sampling [21]. Furthermore, with greater scrutiny,
184 some already established phenotyping schemes fall apart due to inherent variability of the
185 phenotypes themselves within single strains [22].

186

187 Taxonomists continue to struggle with how to incorporate phenotypic and genotypic
188 methods for bacterial classification and to how to name bacterial species [1,3,4,23]. This reliance
189 on human subjectivity builds in an inherent weakness for continuity across data sets and
190 manuscripts. Every time a new phenotype of interest has emerged, or every time a lineage is
191 split due to actual differences in evolutionary signals, there is a possibility that names of species
192 or even genera change. While this practice is no doubt important from a taxonomic standpoint,
193 the plasticity of nomenclature presents a significant obstacle for chasing down relevant data in
194 historical literature because information about strains of interest is often hidden by changes in
195 name for each strain. Advances in publishing technologies also work against the traditional
196 nomenclature schemes. Although every new species name should be thoroughly documented
197 and vetted with publication in designated journals, with the proliferation of modern journals it
198 is easier than ever to publish papers regardless of their conclusions. Reviewers' time is limited,
199 and many may be unaware of current nomenclatural challenges, so that papers may be
200 published where strain descriptions conflict with previously agreed upon taxonomic standards.

201 Such situations amplify confusion in the literature because lineages are then officially referred
202 to in multiple ways without a clear way to delineate the official taxonomic status.

203

204 **The *P. syringae sensu lato* Species Complex as an Example**

205

206 There is no better system to illustrate the nomenclatural challenges of modern day
207 microbiology, and to give examples for confusion arising from the points mentioned above,
208 than *P. syringae*. Pseudomonads can be found ubiquitously across environments, but are also
209 well known as pathogens of humans, animals, insects, fungi, and plants [24]. Therein lies one of
210 the greatest challenges of pseudomonad taxonomy, that nomenclature within this group has
211 often been biased by relying on phenotypes related to pathogenicity across hosts. It is possible,
212 and even likely, that pathogenic contexts represent only a small sample of the selective
213 pressures faced by strains in nature [25]. Therefore, phenotypes that might be of interest in
214 defining species names may not be those that play a large role in the lifestyle of the strains and
215 might therefore not lead to the most cohesive groupings. In a way, *Pseudomonas* is the
216 taxonomic opposite of *H. pylori*.

217

218 The name *Pseudomonas syringae* began as a formal way to classify pseudomonad
219 pathogens of plants, with the name “syringae” coined because the original type strain was a
220 pathogen of lilacs within the genus *Syringa* [23]. Eventually, the LOPAT diagnostic scheme was
221 created to classify strains as *P. syringae* based upon five different phenotypes; levan production,
222 oxidase, potato soft rot, arginine dihydrolase, tobacco hypersensitivity. Species level
223 nomenclature was layered onto this scheme, usually defined by pathogenic reactions and
224 symptoms on host plants. Given the intrinsically changeable nature of species names, strains
225 broadly classified within *P. syringae* are also currently informally grouped into the designations
226 *sensu lato* (the largest sense) and *sensu stricto* (the smallest sense). *P. syringae sensu lato* includes
227 all species that could be classified as part of *P. syringae* according to the LOPAT scheme
228 genomic comparisons but in the absence of any other phytopathogenic information. *P. syringae*
229 *sensu stricto* includes only those strains that are closely related to and phytopathogenically
230 similar to the original type strain. After bacterial species nomenclature was overhauled in 1980,
231 this system was modified to include “pathovar” or “pathogenic variety” status as a way to
232 reflect previously established boundaries between strains with distinct phytopathogenic
233 profiles [26]. Although the causes are outside the bounds of this piece, strains within *P. syringae*
234 *sensu lato* have been split up into multiple species with subsets assigned pathovar designations.
235 Some pathovars themselves are polyphyletic due to convergence in host range (i.e. pv.
236 *avenellae*), whereas others are tightly grouped (i.e. pv. *phaseolicola*) [26,27]. There are even some
237 lineages known to cause disease and which are model systems for host pathogen interactions,
238 but which are not part of a monophyletic species cluster (i.e. *P. syringae* pv. *tomato* DC3000)[23].
239 As with many other bacteria, taxonomy of *P. syringae* has increasingly relied upon on genomic
240 properties and DNA sequences. Strains within the group *P. syringae* were first categorized and
241 split up into nine different genomospecies based upon DNA-DNA hybridization [24]. Later,
242 strains were classified into 13 different groups by multilocus sequence type (MLST) and
243 multiple subgroups, with a reduced yet accurate scheme including the loci *rpoD*, *gltA*, *cit*, and

244 *gyrB* [28,29]. Whole genome phylogenies can now be built using a variety of methods and with
245 draft genomes, all approximate the earlier classification schemes (although there are notable
246 exceptions) [30-32]. Given the current state of the field, one can reasonably determine where
247 within *P. syringae sensu lato* their strains of interest fall given a sequence for only the *gltA* locus
248 (Figure 1 and [28]).

249
250 Perhaps the greatest challenge with taxonomic classification of *P. syringae* is the broad
251 phenotypic versatility and variability that strains display. This bacterium is an environmentally
252 ubiquitous hemibiotroph that can be isolated from environmental sources including rivers,
253 lakes, leaf litter, and clouds [33-35]. For any given strain, host range can be difficult to
254 categorize because the occurrence and severity of symptoms is dependent on the precise
255 environmental conditions and sensitivity to these conditions is itself influenced by phylogenetic
256 placement [36]. Moreover, just because strains fail to cause disease does not mean that they will
257 fail to grow *in planta* or survive epiphytically on those hosts. Adding to the confusion,
258 numerous strains have been isolated from environmental resources and from asymptomatic
259 plants [21]. Many of these environmental strains are closely related to known pathogens of
260 plants, and may even be capable of causing disease under laboratory conditions, but these are
261 not given pathovar designations because they are non-pathogenic or their host range has not
262 been determined [39,40]. To these points, the main reason that genomospecies have not been
263 granted official taxonomic status is because there are no consistent phenotypic differences that
264 differentiate them.

265
266 Since *P. syringae sensu lato* has become a model system for environmental genomics, the
267 genome sequences of hundreds of individual strains have been deposited in publicly accessible
268 databases [41,42]. The inconsistency of metadata across these strains makes it difficult, in the
269 least, to mine much of the information that would be useful for comparative genomic analyses.
270 Numerous cases currently exist within Genbank of strains that could be classified in the *P.*
271 *syringae* species complex that are arguably mischaracterized as an incorrect or obsolete species
272 name. Therefore, if one is interested in accessing genomes for all of these strains, they either
273 have to search and find which species name the genome is listed under or perform repetitive
274 sequence based searches to gauge how strains are related. These inconsistencies hamper
275 research because they prevent researchers from discussing overlapping data sets and lead to
276 wasteful redundancy when multiple groups are unaware that they are independently
277 investigating the same strains.

278
279 This already confusing taxonomic situation has been compounded by a variety of
280 published papers which refer to strains with old names, rename strains incorrectly, or
281 differentially name the same strain because of grey areas in nomenclature. One of the best
282 examples involves a strain from pathovar *phaselicola*, which is a pathogen of green beans [26].
283 This strain has variably been referred to as *P. syringae* pv. *phaseolicola* or *Pseudomonas savastanoi*
284 pv. *phaseolicola* even though it could properly be named as *Pseudomonas amygdali* pv. *phaseolicola*
285 according to current rules [23,26,43]. There are also numerous cases where strains currently
286 exist in taxonomic limbo as *Pseudomonas* sp. because they have not been formally described (i.e.

287 UB246 in Figure 1) [21,28]. The widespread challenges of properly referring to strains within
288 this species complex are demonstrated in Figure 1. Even though most of the strains represented
289 in the figure can be referred to as *P. syringae*, as shown by the highlighted names, multiple other
290 species names are currently used by different research groups. Despite the best intentions of all
291 researchers, nomenclature remains inconsistently applied across these strains. It would be one
292 thing if the taxonomic confusion were limited to one particular strain within *P. syringae sensu*
293 *lato*, but such confusion is seemingly universal and widespread across lineages. I have no
294 doubts that the situation is equally as confusing across other environmentally ubiquitous
295 groups that contain phytopathogens (i.e. *Burkholderia*, *Erwinia*, *Rhizobium* [44-46]), or across
296 groups like *Vibrio* where conflicts arise between traditional nomenclatural schemes and in depth
297 analyses of population genetics signals [47].

298

299

300 **Divorcing Strain Classification from Species Names**

301

302 The ever-increasing flood of genomic data will lead to an increase in nomenclatural
303 confusion across taxa. New DNA sequencing technologies are continuing to emerge and mature
304 so that, very soon, direct sequencing of nucleotides and single cell genomics may be possible
305 under field conditions [47]. Complete genome sequencing will eventually be cost efficient and
306 straightforward enough to use for rapid classification across all taxa, even the unculturable
307 majority. Along these lines, it is worth noting that genome sequencing has potentially overtaken
308 phage phenotyping as the preferred method for *Salmonella* classification in the United Kingdom
309 [48]. Although phenotyping will certainly still be useful for more thorough studies, these tests
310 will always be likely used in combination with genotypic data. Furthermore, it is inevitable that
311 researchers who are only casually interested in proper taxonomy will continue to submit
312 sequences to databases and to publish papers. In this light, what can be done to minimize future
313 confusion?

314

315 A key to clearing these challenges is the realization that taxonomic classification and
316 nomenclature are not commutative. Although the ability to name strains inherently requires
317 classification, classification itself can take place sequentially based on inherent genomic
318 properties and through comparison to strains that are already indexed without the need to
319 group strains together under potentially subjective nomenclatural umbrellas. All that is
320 necessary for such a system is agreement as to an algorithm for classification, and the ability to
321 identify the most closely related strain according to this scheme. Furthermore, although
322 discussions about how to name bacterial strains inevitably deteriorate into disagreements about
323 how to define bacterial “species”, establishment of an independent classification scheme would
324 separate these discussions and disagreements from indexing these strains in practice for
325 posterity.

326

327 There will never be a perfect classification system, but we can develop a scheme
328 centered around whole genome sequences which ensures continuity of research messages
329 across databases and papers. To achieve any level of success this scheme must be relational in

330 that new designations are assigned solely through genotypic similarity to previously indexed
331 strains whether or not they have been formally characterized as a type strain, and it must exist
332 independently of disagreements about species concepts. An added benefit of relational
333 classification is that identifiers could also provide information about relationships between
334 strains. This system must be expandable so that discoveries of new clades do not force a
335 complete rewriting of previous classifications, and must be applicable to uncultured strains.
336 Classification by this system must be automatable and function well regardless of level of
337 taxonomic expertise, and could even be retroactively applied across all published genome
338 sequences and manuscripts. Ideally, these classifications would already be implementable given
339 the infrastructure of Genbank and could be added to keywords of journal articles to streamline
340 search strings for text mining software.

341
342 Numerous options for such a classification scheme already exist, they just need to be
343 universally implemented. MLSA/MLST or 16s rRNA schemes work well except they can be
344 confounded by recombination, can be biased in the subsets of loci used, and ultimately just
345 contain a lower number of informative sites than whole genomes [3,4,17,18]. Moreover,
346 assigning group numbers based on static numbers significantly decreases the ability to
347 efficiently add intermediate taxa at later times and limits the capability to relay information
348 about relatedness between strains through relational classification. For instance, MLST group 2
349 in *P. syringae* has already been expanded to include MLST groups 2a/2b/2c/2d but what would
350 happen if group 2a were further split? Likewise, *P. syringae* MLST group 10 is more closely
351 related to group 5 than to group 9 according to the current typing scheme [28]. Life
352 Identification Numbers (LINS) have been suggested as an alternative scheme that fits this bill,
353 whereby the average nucleotide identity (ANI) of genomes of interest are sequentially
354 compared [49,50]. After one genome is designated as a reference point (denoted by inclusion of
355 0s at all positions in its identifier), new genomes are compared and assigned standard
356 numerical classifications based on overall genome similarity to the most closely related genome
357 already assigned a LIN. Each position within the LINS identifier (denoted by subscript letters as
358 in Figure 1) represents an increasing threshold of similarity for ANI between two strains.
359 According to this scheme, strains that differ in number at position "A" have a much greater
360 difference in ANI than strains that differ in number at position "E". Assigning identifiers to new
361 genomes is akin to a first order Markov process, because classification is memoryless and based
362 solely on the most similar genome already classified. In order to assign LIN values to a strain of
363 interest one only needs to identify the most closely related previously categorized strain and
364 compare ANI values. Due to the sequential nature of this scheme, increasing thresholds can be
365 utilized where necessary in order to delineate closely related taxa, which would be manifest as
366 additional positions added to the identifier (i.e. 0_A,0_B,0_C,0_D,0_E,1_F/0_A,0_B,0_C,0_D,0_E,0_F instead of
367 0_A,0_B,0_C,0_D/0_A,0_B,0_C,0_D). Firm borders between species groups need not exist because schemes like
368 LINS provide progressive classification where one simply includes more finely grained
369 comparisons of genome similarity. LINS is a computational example of the goal of moving to
370 classification based on DNA hybridization, except that classification can be automated and take
371 a fraction of the time and expertise. Notably, this system has been applied to strains within *P.*
372 *syringae* [51].

373
374 Some have highlighted the weaknesses of classification solely through genome
375 sequences [1,3,4,50]. Substantive challenges include instances where whole genome data does
376 not exist, but strains can be partially assigned based on a limited set of genomic characteristics
377 (such as MLST or 16s rRNA loci) where necessary. There will also be challenges for strains
378 which have experienced extensive genomic diversification, either through selection or drift, as is
379 the case with genome minimization across many intracellular parasites [51]. Bull *et al.* also bring
380 up the idea that strains are often mislabeled or mixed up during storage or sequencing [4].
381 However, such problems exist regardless of classification scheme and an automated system of
382 classifications linked to publication would actually enable interested parties to track down
383 incorrectly named strains. Lastly, one of the greatest weaknesses to new classification schemes
384 is not the most rational objection, but is nonetheless important to consider. Many researchers
385 have become comfortable with and have grown accustomed to assigning binomial names to
386 new microbes. Numerical codes are a less elegant way to describe new taxa and will be
387 inherently unpalatable to some, regardless of whether this is implemented in parallel with the
388 binomial system.

389 390 **Concluding Remarks**

391
392 The overall message of this piece is not to throw out all previous taxonomic systems and
393 start anew, but that we must move to implement a sequence-based classification system that is
394 unambiguous. We can create a retroactive and expandable system that could be used by
395 regulatory agencies, in publication keywords, and with metadata that exists independently of
396 species nomenclature or concepts. This system would enable microbial classification to remain
397 unchanged going forward. These systems can be expandable, with algorithms that can be
398 created to automatically classify or group genomes by similarity when they are submitted to
399 databases or to journals. This suggestion parallels both the call for universal DNA Barcoding
400 across species [52] as well as the creation of ORCID identifiers (orcid.org) to trace publications
401 by individual authors. Researchers from across disciplines will likely admit that something has
402 to be done to change the way strains are classified, challenges which are exemplified quite well
403 by *P. syringae sensu lato* (Figure 1). There may not be one single answer to the problem of
404 nomenclatural confusion in microbes, but the today's challenges will only worsen unless we
405 rethink taxonomic strategies.

406

407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426

Figure 1. Current and Potential Classification of *Pseudomonas syringae sensu lato*. A Bayesian phylogeny was built using the full-length sequence of *gltA* from a diverse subset of *P. syringae sensu lato* genomes. Although each of these strains can be referred to as *P. syringae* in published papers and in metadata, strains that are currently referred to by species names other than *P. syringae* are highlighted in different colors. Strains which are known pathogens are labelled as either “pv.” for pathovar or with alternative species names in the absence of pathovars. Strains which have been isolated from environmental sources or from plants in the absence of official disease classification have been labeled only with the strain number. To the right of each strain name is a mock numerical classification, based upon an example scheme described in [49], but in the absence of defined ANI thresholds or actual genome comparisons. Each column within this identifier represents an increasing threshold for ANI between that particular strain and the most similar genome already assigned a LIN. For instance column A could represent an ANI of 40%, followed by column B at 50%, C at 75%, D at 90%, E at 99%. Different numbers within these columns represents that particular strains differ from a common reference for that threshold ANI. Abbreviations: WGS, whole genome sequencing; LINS, Life Identification Numbers; ANI, Average nucleotide identity.

- 427
428 1 Oren, A. and Garrity, G.M. (2013) Then and now: a systematic review of the systematics
429 of prokaryotes in the last 80 years. *Antonie Van Leeuwenhoek* 106, 43–56
430 2 Shapiro, B.J. and Polz, M.F. (2015) Microbial Speciation. *Cold Spring Harb Perspect Biol* 7,
431 3 Zhi, X.-Y. *et al.* (2011) Prokaryotic systematics in the genomics era. *Antonie Van*
432 *Leeuwenhoek* 101, 21–34
433 4 Bull, C.T. and Koike, S.T. (2015) Practical Benefits of Knowing the Enemy: Modern
434 Molecular Tools for Diagnosing the Etiology of Bacterial Diseases and Understanding
435 the Taxonomy and Diversity of Plant-Pathogenic Bacteria. *Annu Rev Phytopathol* 53, 157–
436 180
437 5 Blaby-Haas, C.E. and de Crécy-Lagard, V. (2011) Mining high-throughput experimental
438 data to link gene and function. *Trends Biotechnol* 29, 174–182
439 6 Salhi, A. *et al.* (2015) DESM: portal for microbial knowledge exploration systems. *Nucleic*
440 *Acids Research* DOI: 10.1093/nar/gkv1147
441 7 Linz, B. and Schuster, S.C. (2007) Genomic diversity in *Helicobacter* and related
442 organisms. *Res Microbiol* 158, 737–744
443 8 Baltrus, D. *et al.* (2009) *Helicobacter pylori* Genome Plasticity. *Genome dynamics* 6, 75–90
444 9 Correa, P. and Piazuelo, M.B. (2008) Natural history of *Helicobacter pylori* infection. *Dig*
445 *Liver Dis* 40, 490–496
446 10 Solnick, J.V. *et al.* (2006) The Genus *Helicobacter*. In *The prokaryotes* pp. 139–177, Springer
447 New York
448 11 Welch, R.A. (2006) The Genus *Escherichia*. In *The prokaryotes* pp. 60–71, Springer New
449 York
450 12 Pupo, G.M. *et al.* (2000), Multiple independent origins of *Shigella* clones of *Escherichia*
451 *coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci*
452 12;97(19):10567-72
453 13 Lapage, S.P. *et al.* (1992) *International Code of Nomenclature of Bacteria: Bacteriological Code,*
454 *1990 Revision - PubMed - NCBI,*
455 14 Chui, H. *et al.* (2015) Rapid, Sensitive and Specific *E. coli* H Antigen Typing by Matrix-
456 Assisted Laser Desorption/Ionization Time-of-Flight (MALDI-TOF)-Based Peptide Mass
457 Fingerprinting. *J. Clin. Microbiol.*
458 15 Baggesen, D.L. *et al.* (2010) Phage typing of *Salmonella Typhimurium* - is it still a useful
459 tool for surveillance and outbreak investigation? *Euro Surveill.* 15, 19471
460 16 Fournier, P.-E. *et al.* (2015) From culturomics to taxonomogenomics: A need to change
461 the taxonomy of prokaryotes in clinical microbiology. *Anaerobe* 36, 73–78
462 17 Tringe, S.G. and Hugenholtz, P. (2008) A renaissance for the pioneering 16S rRNA gene.
463 *Curr Opin Microbiol* 11, 442–446
464 17 Yarza, P. *et al.* (2014) Uniting the classification of cultured and uncultured bacteria and
465 archaea using 16S rRNA gene sequences. *Nat Rev Micro* 12, 635–645
466 18 Almeida, N.F. *et al.* (2010) PAMDB, a multilocus sequence typing and analysis database
467 and website for plant-associated microbes. *Phytopathology* 100, 208–215
468 19 Bull, C.T. *et al.* (2010) Comprehensive list of names of plant pathogenic bacteria, 1980-
469 2007. *J Plant Pathol* 92:3 551-592

- 470 20 Murray, R.G. and Stackebrandt, E. (1995) Taxonomic note: implementation of the
471 provisional status Candidatus for incompletely described procaryotes. *Int J Syst Bacteriol*
472 45, 186–187
- 473 21 Diallo, M.D. *et al.* (2012) *Pseudomonas syringae* naturally lacking the canonical type III
474 secretion system are ubiquitous in nonagricultural habitats, are phylogenetically diverse
475 and can be pathogenic. *ISME J*
- 476 22 Bartoli, C. *et al.* (2014) The *Pseudomonas viridiflava* phylogroups in the *P. syringae* species
477 complex are characterized by genetic variability and phenotypic plasticity of
478 pathogenicity-related traits. *Environ Microbiol* 16, 2301–2315
- 479 23 Young, J.M. (2010) Taxonomy of *Pseudomonas syringae*. *Journal of Plant Pathology* 92, S1.
480 5–S1. 14
- 481 24 Gardan, L. *et al.* (1999) DNA relatedness among the pathovars of *Pseudomonas syringae*
482 and description of *Pseudomonas cannabina* sp. Nov. (ex Sutic and Dowson 1959). *Int J Syst*
483 *Bacteriol* 49 Pt. 2: 469-78
- 484 25 Peix, A. *et al.* (2009) Historical evolution and current status of the taxonomy of genus
485 *Pseudomonas*. *Infect Genet Evol* 9, 1132–1147
- 486 26 Morris, C.E. *et al.* (2008) The life history of the plant pathogen *Pseudomonas syringae* is
487 linked to the water cycle. *ISME J* 2, 321–334
- 488 27 Young, J.M. *et al.* (1992) Changing concepts in the taxonomy of plant pathogenic
489 bacteria. *Annu Rev Phytopathol* 30:67-105
- 490 28 Arnold, D.L. *et al.* (2011) *Pseudomonas syringae* pv. *phaseolicola*: from “has bean” to
491 supermodel. *Mol Plant Pathol* 12(7):617-27
- 492 29 MARCELLETTI, S. and Scortichini, M. (2015) Comparative Genomic Analyses of
493 Multiple *Pseudomonas* Strains Infecting *Corylus avellana* Trees Reveal the Occurrence of
494 Two Genetic Clusters with Both Common and Distinctive Virulence and Fitness Traits.
495 *PLoS ONE* 10, e0131112
- 496 30 Berge, O. *et al.* (2014) A User's Guide to a Data Base of the Diversity of *Pseudomonas*
497 *syringae* and Its Application to Classifying Strains in This Phylogenetic Complex. *PLoS*
498 *ONE* 9, e105547
- 499 31 Hwang, M.S.H. *et al.* (2005) Phylogenetic characterization of virulence and resistance
500 phenotypes of *Pseudomonas syringae*. *Appl Environ Microbiol* 71, 5182–5191
- 501 32 Baltrus, D.A. *et al.* (2011) Dynamic Evolution of Pathogenicity Revealed by Sequencing
502 and Comparative Genomics of 19 *Pseudomonas syringae* Isolates. *PLoS Pathog* 7, e1002132
- 503 33 Baltrus, D.A. *et al.* (2014) Ecological Genomics of *Pseudomonas syringae*. in D.C. Gross, A.
504 Lichens-Park and C. Kole, eds. *Genomics of plant-associated bacteria*. Springer, Berlin,
505 Heidelberg. 59-77
- 506 34 Baltrus, D.A. *et al.* (2014) Incongruence between multi-locus sequence analysis (MLSA)
507 and whole-genome-based phylogenies: *Pseudomonas syringae* pathovar *pisi* as a
508 cautionary tale. *Mol Plant Pathol* 15, 461–465
- 509 35 Hirano, S. and Upper, C. (2000) Bacteria in the leaf ecosystem with emphasis on
510 *Pseudomonas syringae*---a pathogen, ice nucleus, and epiphyte. *Microbiol Mol Biol Rev* 64,
511 624

- 512 36 Morris, C.E. *et al.* (2009) Expanding the paradigms of plant pathogen life history and
513 evolution of parasitic fitness beyond agricultural boundaries. *PLoS Pathog* 5, e1000693
- 514 37 Morris, C.E. *et al.* (2010) Inferring the Evolutionary History of the Plant Pathogen
515 *Pseudomonas syringae* from Its Biogeography in Headwaters of Rivers in North America,
516 Europe, and New Zealand. *MBio* 1, e00107–10–e00107–20
- 517 38 Mazarei, M. and Kerr, A. (1990) Distinguishing pathovars of *Pseudomonas syringae* on
518 peas: nutritional, pathogenicity and serological tests. *Plant Pathology* 39, 278–285
- 519 39 Elvira-Recuenco, M. *et al.* (2003) Differential responses to pea bacterial blight in stems,
520 leaves and pods under glasshouse and field conditions. *Eur Journal Plant Pathol* 109, 555-
521 564
- 522 40 Hockett, K.L. *et al.* (2014) *Pseudomonas syringae* CC1557: A Highly Virulent Strain With
523 an Unusually Small Type III Effector Repertoire That Includes a Novel Effector. *Mol*
524 *Plant Microbe Interact* 27, 923–932
- 525 41 Bartoli, C. *et al.* (2014) A framework to gauge the epidemic potential of plant pathogens
526 in environmental reservoirs: the example of kiwifruit canker. *Mol Plant Pathol* 16, 137–
527 149
- 528 42 Mohr, T.J. *et al.* (2008) Naturally occurring nonpathogenic isolates of the plant pathogen
529 *Pseudomonas syringae* lack a type III secretion system and effector gene orthologues. *J*
530 *Bacteriol* 190, 2858–2870
- 531 43 O'Brien, H.E. *et al.* (2011) Evolution of Plant Pathogenesis in *Pseudomonas syringae*: A
532 Genomics Perspective. *Annu Rev Phytopathol* 49, 269–289
- 533 44 McCann, H.C. *et al.* (2013) Genomic Analysis of the Kiwifruit Pathogen *Pseudomonas*
534 *syringae* pv. *actinidiae* Provides Insight into the Origins of an Emergent Plant Disease.
535 *PLoS Pathog* 9, e1003503
- 536 45 Ramos, C. *et al.* (2012) *Pseudomonas savastanoi* pv. *savastanoi*: some like it knot. *Mol Plant*
537 *Pathol.* 13(9), 998-1009
- 538 46 Pritchard, L. *et al.* (2015) Genomics and Taxonomy in Diagnostics for Food Security: soft-
539 rotting enterobacterial plant pathogens. *Anal. Methods* DOI: 10.1039/C5AY02550H
- 540 47 Ormeño-Orrillo, E. *et al.* (2015) Systematic and Applied Microbiology. *Syst. Appl.*
541 *Microbiol.* 38, 287–291
- 542 48 Gupta, R.S. (2014) Molecular signatures and phylogenomic analysis of the genus
543 *Burkholderia*: proposal for division of this genus into the emended genus *Burkholderia*
544 containing pathogenic organisms and a new genus *Paraburkholderia* gen. nov. harboring
545 environmental species. *Front Genet.* doi: 10.389/fgene.2014.00429
- 546 49 Preheim, S.P. *et al.* (2011) Merging taxonomy with ecological population prediction in a
547 case study of *Vibrionaceae*. *Appl Environ Microbiol* 77:7195-7206
- 548 50 Loman, N.J. and Pallen, M.J. (2015) Twenty years of bacterial genome sequencing. *Nat*
549 *Rev Micro* 13, 787–794
- 550 51 Ashton, P. *et al.* (preprint) Revolutionising public health reference microbiology
551 using whole genome sequencing, *Salmonella* as an exemplar. bioRxiv
552 doi: <http://dx.doi.org/10.1101/033225>
553

- 554 52 Marakeby, H. *et al.* (2014) A System to Automatically Classify and Name Any Individual
555 Genome-Sequenced Organism Independently of Current Biological Classification and
556 Nomenclature. *PLoS ONE* 9, e89142
- 557 53 Weisberg, A.J. *et al.* (2015) Similarity-based codes sequentially assigned to ebolavirus
558 genomes are informative of species membership, associated outbreaks, and transmission
559 chains. *Open Forum Infect Dis* 2:ofv024
- 560 54 Vinatzer, B.A. *et al.* A new exclusively genome-based species-independent taxonomic
561 framework for all life forms applied to *Pseudomonas syringae*. Poster session presented at:
562 APS Annual Meeting; 2015 August 1-5; Pasadena, CA
- 563 55 Staley, J.T. (2006) The bacterial species dilemma and the genomic-phylogenetic species
564 concept. *Philos Trans R Soc Lond B Biol Sci* 361, 1899–1909
- 565 56 Klasson, L. (2004) Evolution of minimal-gene-sets in host-dependent bacteria. *Trends*
566 *Microbiol* 12, 37–43
- 567 57 Goldstein, P.Z. and DeSalle, R. (2010) Integrating DNA barcode data and taxonomic
568 practice: Determination, discovery, and description. *Bioessays* 33, 135–147
569

Also Currently Known As:

Pseudomonas savastanoi

Pseudomonas amygdalii

Pseudomonas avenellae

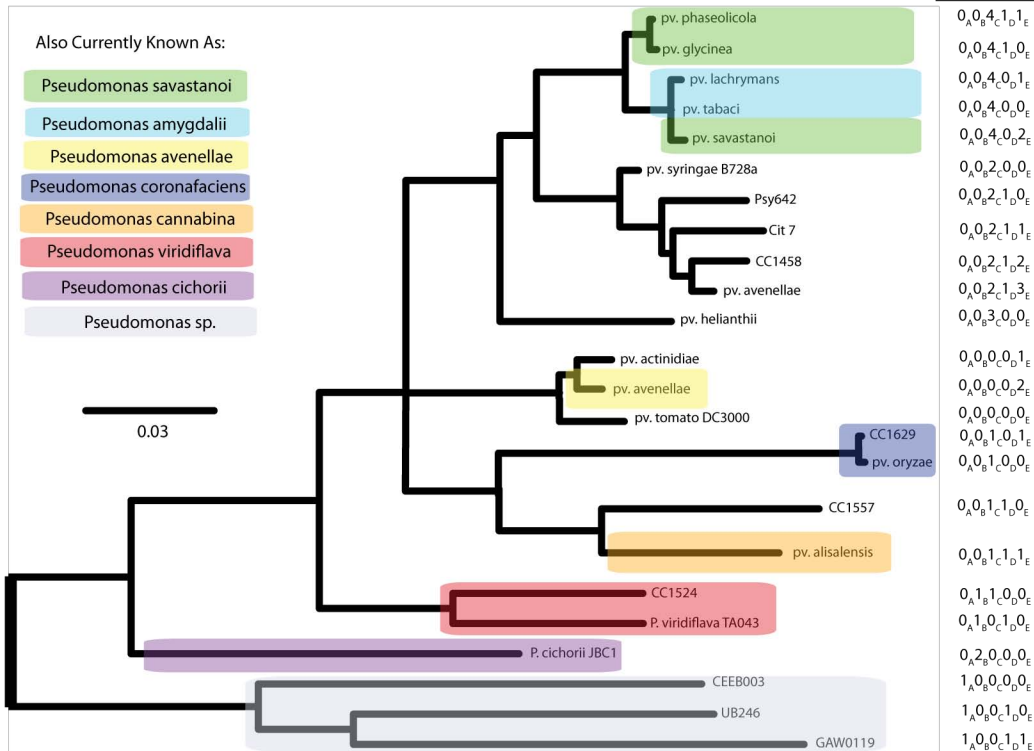
Pseudomonas coronafaciens

Pseudomonas cannabina

Pseudomonas viridiflava

Pseudomonas cichorii

Pseudomonas sp.



Also Currently Known As:

Pseudomonas savastanoi

Pseudomonas amygdalii

Pseudomonas avenellae

Pseudomonas coronafaciens

Pseudomonas cannabina

Pseudomonas viridiflava

Pseudomonas cichorii

Pseudomonas sp.

