Phylogenetic and population structure analyses uncover pervasive misclassification and help assessing the biosafety of *Pseudomonas putida* for biotechnological applications

Hemanoel Passarelli-Araujo[1,2,*], Sarah H. Jacobs[2], Glória R. Franco[1], Thiago M. Venancio[2,*]

[1]Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

[2]Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes, RJ, Brazil.

*Corresponding authors

Short running title: Population structure of *Pseudomonas putida*

Av. Alberto Lamego 2000, P5 sala 217; Parque Califórnia
Campos dos Goytacazes, RJ, Brazil
CEP: 28013-602
HPA: hemanuel.passarelli@gmail.com; TMV: thiago.venancio@gmail.com

Summary

Pseudomonads inhabit a variety of environments. *Pseudomonas putida* is a saprophytic soil bacterium, known to bioremediate toxic compounds and to promote plant growth. Despite the ability to occupy different niches, the population structure of *P. putida* remains unclear. Here, we report its population structure for the first time. We retrieved 11,025 *Pseudomonas* genomes and used *P. putida* KT2440 as a type strain to define *P. putida*. We found that 90 of the 138 deposited *P. putida* genomes were misclassified, including some reference genomes. The population structure of *P. putida* is composed of at least 7 clonal complexes (CCs). Clinical isolates are not restricted to a specific CC and acquired resistance genes are present at low frequency in plasmids from clinical isolates. Virulence profiles support the potential of CC7 members to outcompete other plant or human pathogens through a type VI secretion system. *P. putida* lacks key genes for indole-3-acetic acid production, indicating an incomplete or non-functional pathway. Finally, we found that horizontal gene transfer had an important role in shaping the ability of *P. putida* to bioremediate aromatic compounds such as toluene. Our results provide the grounds to understand *P. putida* genetic diversity and its safety for environmental applications.

Keywords: *Pseudomonas alloputida*, *Pseudomonads*, cgMLST

35 **INTRODUCTION**

36 The genus *Pseudomonas* is diverse and composed of Gram-negative bacteria
37 (Stanier et al., 1966; Palleroni, 2010). Since the first description of *Pseudomonas*, in
38 1894, the taxonomy of the genus has evolved in parallel with new methodologies to
39 discriminate species (Palleroni, 2010). The taxonomy was inconsistent until the
40 division of *Pseudomonas* into rRNA homology groups (Stanier et al., 1966). However,
41 because of the low-resolution of 16S rRNA gene for intra-specific classification, other
42 housekeeping genes were used to improve taxonomic resolution (Ait Tayeb et al.,
43 2005; Gomila et al., 2015).

44 The *Pseudomonas putida* group includes other species such as *P. monteilii*, *P.*
45 *fulva*, *P. plecoglossicida*, and *P. putida sensu stricto* (*P. alloputida*). *P. putida* belongs
46 to a wide group of fluorescent species from the *Pseudomonas* genus. It is ubiquitous
47 in soil and water and several strains have been isolated from various environments,
48 such as polluted soils and plant roots (Volke et al., 2020). *P. putida* is well known to
49 perform many functions such as plant growth promotion, bioremediation, and
50 protection from plant pathogens (Timmis, 2002).

51 Because of its high metabolic diversity, applications of *P. putida* have been
52 proposed for environmental, industrial, and agricultural uses (Volke et al., 2020). *P.*
53 *putida* KT2440 was the first strain of this species to have its genome sequenced, and
54 it is considered a model organism to study plant-microbe interactions (Nelson et al.,
55 2002; Belda et al., 2016). Therefore, KT2440 can be thought as a type strain for *P.*
56 *putida*. KT2440 was isolated from garden soil in Japan and it is certified as a host-
57 vector system safety level 1 (Kampers et al., 2019). This strain has become a
58 workhorse for industrial biotechnology because of its high metabolic versatility, genetic

accessibility, and stress-resistance (Nikel and de Lorenzo, 2018). In addition to the biotechnological application of *P. putida*, some clinical isolates have been described as reservoirs of resistance genes (Fernandez et al., 2015; Molina et al., 2016), raising concerns about its biosafety.

Although *P. putida* exhibit enormous potential for bioremediation, biocontrol, and plant growth promotion, its population structure remains poorly understood. In this work, we report the population structure of *P. putida sensu stricto* (or *P. alloputida*, as recently proposed by Keshavarz-Tohid et al. (2019)), which was used to estimate the diversity and distribution of bioremediation and plant growth promotion genes. Further, we also used the inferred population structure to better understand the prevalence of antibiotic resistance and virulence genes and to assess, in detail, the potential biosafety concerns regarding this species.

RESULTS AND DISCUSSION

Phylogeny and classification of *Pseudomonas putida* (*P. alloputida*)

We obtained 11,025 *Pseudomonas* genomes available in RefSeq in June 2020, out of which 10,457 had completeness greater than 90% according to BUSCO (Seppey et al., 2019). We computed the pairwise distances between each isolate using mashtree (Katz et al., 2019) to compute the distance tree of the genus, which is highly diverse (Figure 1). We mapped each genome deposited in the NCBI RefSeq as *P. putida* in the tree and found that the highest density of genomes falls within a monophyletic group of 439 isolates with average nucleotide identity (ANI) values between 84% and

100% (Figure S1a); this clade corresponds to the *P. putida* group (Figure 1) and comprises other species such as *P. plecoglossicida*, *P. monteilli,* and *P. fulva*.

ANI analysis provides a raw estimate of bacterial species (Richter and Rossello-Mora, 2009). A minimum threshold of 95% ANI has been used to attain species membership, a value that has been empirically defined based on correlations with DNA-DNA hybridization and 16S rRNA thresholds (Richter and Rossello-Mora, 2009; Bobay, 2020). We used *P. putida* KT2440 as a type-strain to evaluate the ANI values from other isolates in the *P. putida* group. The sorted distribution of ANI values from KT2440 showed an abrupt break around 95%, supporting its effectiveness as a threshold to delineate *P. putida* (Figure S1b). The isolate previously classified as *P. monteilli* IOFA19 (GCA_000633915.1) had the lowest ANI value within the predicted species (95.61%), followed by a drop to 91.19% (Figure S1b).

We conducted a network analysis to assess the species composition according to ANI > 95% with members from the *P. putida* group. We observed a discrete number of cohesive clusters that would correspond to the expected number of species within the *P. putida* group (Figure 2). We retrieved other species such as *P. asiatica*, *P. soli*, and *P. monteilli*, as well as some potentially novel species (Table S1). In this analysis, a species was defined as a cluster with a type strain or with at least ten connected isolates. We found clusters that most likely correspond to new species, as also noticed elsewhere (Keshavarz-Tohid et al., 2019). Further, these values are probably underestimated, as new genomes would increase the number of connections in the network.

The poor classification of *P. putida* isolates is a subject of concern. Isolates that are well known for their ability to promote plant growth (W618) (Taghavi et al., 2009), to

105 oxidize manganese (GB-1) (Rosson and Nealson, 1982), and to damage human
106 tissues (HB3267) (Fernandez et al., 2015), are actually not *P. putida* strains, as they
107 belong to different groups in the network. For example, HB3267 was classified as *P.*
108 *putida* because of its close phylogenetic relationship with the nicotine degrader S16
109 (Molina et al., 2014). HB3267, along with DLL-E4, SF1, and S11, were proposed as
110 members of a new species, *Pseudomonas shirazica* (Keshavarz-Tohid et al., 2019).
111 However, we found that these strains, along with S16, grouped with *Pseudomonas*
112 *asiatica* type strains. Importantly, both *P. shirazica* and *P. asiatica* were proposed
113 around the same time (Keshavarz-Tohid et al., 2019; Tohya et al., 2019) as attempts
114 to describe the same species. Therefore, the nomenclature of the group containing
115 HB3267 warrants a more careful reconciliation. Keshavarz-Tohid et al. proposed to
116 rename *P. putida sensu stricto* to *P. alloputida* based on the proposed type-strain Kh7
117 (Keshavarz-Tohid et al., 2019). Nevertheless, Kh7 is also a member from *P. putida*
118 *sensu stricto* group containing the type strain KT2440.

119 The precise taxonomy of the *P. putida* group still warrants a deeper analysis.
120 Although the strain NBRC 14164 has been proposed as a *P. putida* type strain (Ohji et
121 al., 2014), it belongs to a group containing 16 other genomes, named in our network
122 as *Pseudomonas spp1*. Moreover, the phylogenetic analysis of the *P. putida* group
123 also revealed the phylogenetic isolation of NBRC 14164 from the other main *P. putida*
124 strains (Yonezuka et al., 2017). We considered the group *Pseudomonas spp1* as a
125 potentially novel species and kept KT2440 as the type strain because of its pioneering
126 isolation, description, and genome sequencing (Nelson et al., 2002). We excluded
127 strains from potentially different species that were classified as *P. putida* (Table S2)
128 and focused on *P. putida sensu stricto* (*P. alloputida*), represented as the red group in
129 the network with 68 isolates (Figure 2).

130

131 **Pangenome analysis**

132 An effective way to investigate the evolution of a given population is through
133 pangenome analysis. A pangenome is defined as the total set of genes in a given
134 species (Tettelin et al., 2005), which is subdivided into core genes, when present in all
135 isolates; accessory genes, when present in at least two (but not in all) isolates or;
136 unique genes, when exclusive to an isolate. The pangenome of *P. putida* has been
137 previously estimated with only 9 isolates (Udaondo et al., 2016). However, among
138 those isolates, GB1, HB3267, S16, and W619 are likely from other species and were
139 removed from our analysis (Table S2). Hence, we re-assessed the pangenome of *P.*
140 *putida* using 68 isolates with well-supported taxonomic delineation.

141 The *P. putida* pangenome comprises 25,782 gene families, of which 3803 (14.75%)
142 constitute the core genome (i.e. genes present at least 95% of the isolates). The core-
143 genome is associated with intrinsic physiology and tends to decrease in size as new
144 isolates are added. Although with a much lower number of isolates, the core-genome
145 estimated earlier (3386 genes (Udaondo et al., 2016)) is smaller than that described
146 here. This result corroborates the misclassification of *P. putida* strains, since the
147 inclusion of genomes from different species would reduce the number of core genes.

148 Regarding the pangenome openness, the slope of the curve (α= 0.417) supports
149 that *P. putida* has an open pangenome (Tettelin et al., 2005) (Figure S2). Our
150 estimated α value is much lower than the maximum threshold used to define an open
151 pangenome ($\alpha < 1$), corroborating the previous prediction (Udaondo et al., 2016).
152 Further, this low α value implies that, if more isolates are included in the analysis, new

153 gene families will be detected at a high rate. The high number of gene families in the
154 pangenome of *P. putida* is dominated by unique and low-frequency genes. Only 6,373
155 genes (24.71%) are found between 5% to 95% of the isolates, while 15,606 genes
156 (60.53%) are found in less than 5% of the isolates, including 10,917 unique genes. The
157 high number of low-frequency genes could be partially attributed to fragmented
158 genomes. However, reference genomes such as DOT-T1E showed 267 unique genes,
159 far more than the average of 160.5 unique genes found across the dataset. Although
160 the number of low-frequency genes may be overestimated, the high prevalence of
161 unique genes among closely-related genomes indicates a high turnover of unstable
162 genes that might be adaptive under transient selective pressures in their environments.

163 Once determining the exact number of pan and core genomes remains intractable,
164 we also estimated the genomic fluidity (φ) of *P. putida*. The φ estimator is a robust
165 metric that represents the ratio of unique gene families to the sum of gene families in
166 genomes pairs (Kislyuk et al., 2011). The smaller the φ, the greater the genes shared
167 by a pair of randomly selected genomes. Analyzing φ instead of the core genome
168 proportion provides a more realistic measure of cohesiveness within a species
169 because low-frequency genes directly affect the pangenome size. *P. putida* has φ =
170 0.20 ± 0.04, indicating that random pairs of *P. putida* genomes have an average 20%
171 and 80% of unique and shared genes, respectively.

172

173 **Population structure**

174 Determining relationships between isolates can provide novel insights into the
175 metabolic diversity of a given species. The Multilocus Sequence Typing (MLST)

176 analysis is a technique based on single-nucleotide polymorphisms (SNPs) within a few
177 housekeeping genes. MLST schemes are available for several species (Jolley et al.,
178 2018). In an MLST analysis, each combination of SNPs defines a Sequence Type (ST)
179 that can be linked to form Clonal Complexes (CC) (Jolley et al., 2018). A variation of
180 classical MLST is the core genome MLST (cgMLST), which provides greater resolution
181 by using SNPs from the entire core genome (de Been et al., 2015). Here, we used
182 225,009 SNPs obtained from the *P. putida* core genome to reconstruct the
183 phylogenetic tree and the cgMLST profile.

184 The cgMLST tree unveiled 7 CCs (Figure 3a), out of which CC1 is the most distant
185 from the rest of the population, a trend that is also supported by the ANI analysis
186 (Figure S3). Six out of the 9 clinical isolates were found in CC4 and three in CC7. We
187 also checked whether the same clustering pattern could be obtained by analyzing the
188 presence/absence patterns of accessory genes. We estimated the Jaccard distance
189 for genes present between 5% - 95% of the isolates to perform a Principal Coordinate
190 Analysis (PCoA), which allowed us to resolve all the main groups, particularly CC1,
191 CC4, and CC7 (Figure 3b).

192 We also performed a Discriminant Analysis of Principal Components (DAPC)
193 (Jombart et al., 2010) to recover the genes that contribute most to separate the
194 population based on their presence/absence profiles (Figure S4). The top 50
195 discriminating genes separate CC5 and CC7 from the rest of the population, but fail to
196 resolve the relationships between other CCs (Figure S5). Next, we used branch
197 lengths from cgMLST tree as an indirect estimator of diversity within the CCs (Figure
198 3c). CC7 is the most clonal group, comprising 9 isolates including KT2440 and three
199 clinical isolates (GTC_16482, GTC_16473, and NBRC_111121).

200 The previously published MLST scheme for *P. putida* comprises eight housekeeping
201 genes (*argS, gyrB, ileS, nuoC, ppsA, recA, rpoB,* and *rpoD*) (Ogura et al., 2019). We
202 assigned isolates to STs and grouped those into CCs (Table S3). This analysis
203 revealed some incongruences with the previous reports (Ogura et al., 2019). For
204 example, by using allele combinations with perfect-match to predict the STs, KT2440
205 was detected as ST69 and not as ST58 (Ogura et al., 2019), indicating that a revision
206 in the public database is necessary. The predicted number of CCs was also supported
207 by the admixture model from STRUCTURE (Pritchard et al., 2000). This model
208 assumes that each isolate has ancestry from one or more *K* genetically different
209 sources, which we referred to as CCs. The number of CCs corresponds to the
210 estimated number of clusters represented by parameter *K*. Instead of using the highest
211 raw marginal likelihood, we followed the protocol for estimating the best *K* value
212 suggested by Evanno et al (2005). The *ad hoc* statistics ΔK indicates that, according
213 to our dataset, the population structure of *P. putida* is composed of at least 7 CCs
214 (Figure 3d).

215 Once data from eight *loci* may be insufficient to accurately describe the population
216 structure of *P. putida*, the populations identified by STRUCTURE were only considered
217 if they matched cgMLST results. When correlating the cgMLST tree topology with
218 ancestry proportion predicted for each isolate, we observed a clear delimitation of
219 genetic blocks for each CC (Figure 4). However, there are a few inconsistencies. For
220 example, isolate B4 has a greater ancestry proportion with CC2, although cgMLST
221 indicates its greater proximity to CC3. In addition, KCJK7916, FF4, and B6-2 were not
222 assigned to a CC because, although they have a higher proportion of ancestry with
223 CC3, they are paraphyletic to CC3 and CC4. Since CCs are assumed to be
224 monophyletic, these strains were designated as No Clonal Complex (NOCC). We

225 expect that a greater number of *P. putida* genomes from isolates from various sources
226 will improve the resolution of the *P. putida* population structure, including the CC
227 assignment to isolates described here as NOCC.

228 The cgMLST phylogenetic tree and the distance tree support CC1 as the basally
229 branching group of *P. putida* (Figure 4). CC1 comprises isolates from deep-sea
230 sediments from Indian Ocean, costal water from the Pacific Ocean, lotus field, and
231 arthropods (Table 1). The main difference of CC1 is the lack of 229 gene families in
232 the accessory genome, which are present in at least one isolate from all other CCs
233 (Table S4). Among these absent gene families, there is a genomic island with nearly
234 46 Kbp and harboring 38 genes. Some of those genes are involved in sugar transport,
235 as previously identified in KT2440 (CC7) (Wu et al., 2011), as well as genes encoding
236 hypothetical proteins (coordinates 3,126,465-3,172,496 in KT2440).

237

238 **Resistance profiles**

239 We evaluated the composition of antibiotic resistance genes using the CARD
240 database (Alcock et al., 2020). All 15 different genes in the core resistome encode
241 MDR efflux pumps (Table 2, Table S5). We found MDR efflux complexes such as
242 MexAB-OprM, MexEF-OprN, and MexJK, from the resistance-nodulation-cell division
243 (RND) efflux pumps family. These efflux pumps are associated with intrinsic and
244 acquired multidrug resistance in *P. aeruginosa* (Schweizer, 2003; Lister et al., 2009).
245 However, these RND efflux pumps may play an alternative role in *P. putida* by pumping
246 out toxic substrates such as toluene (Rojas et al., 2001). We also identified *cpxR*, which
247 encodes a protein that promotes MexAB-OprM expression in *P. aeruginosa* in the

248 absence of the MexR repressor (Tian et al., 2016), which is absent in *P. putida*. The
249 presence of MexAB-OprM in the core genome, under CpxR regulation, supports its
250 involvement with intrinsic physiology in addition to drug resistance, because this
251 complex can be involved in both quorum-sensing and host interaction mediation in *P.*
252 *aeruginosa* (Evans et al., 1998; Sanchez et al., 2002).

253 Regarding the acquired resistome, we found 45 different genes that confer
254 resistance by pumping out or inactivating antibiotics, as well as by interacting with the
255 antibiotic targets (Table S5). These genes are distributed at low-frequency (Figure 5a),
256 indicating that most of them are strain-specific or acquired through horizontal gene
257 transfer. In general, there is no clear correlation between acquired resistome and
258 population structure (Figure S6), although CC7 has more acquired resistance genes
259 than other CCs (Figure 5c, Figure S6).

260 Along with efflux pumps, the acquired resistome includes a range of antibiotic
261 inactivating enzymes genes that confer resistance to beta-lactams (e.g. bla_{CARB-3},
262 bla_{IMP-1}, bla_{OXA-2}, bla_{PDC-7}, bla_{TEM-1}, and bla_{VIM-2}); aminoglycosides (e.g. *aac(6')-IIa*,
263 *aadA*, *aph(3'')-Ia*, and *aph(6')-Id*; chloramphenicol (*cat*) and; fosfomycin (*fosA*) (Table
264 S3). These genes were distributed in few strains, mostly clinically relevant (Figure S6,
265 Table S5); the top four strains with more acquired resistance genes were GTC_16473
266 (19 genes), GTC_16482 (16 genes), DZ-F23 (14 genes), and 15420352 (11 genes).
267 Importantly, all of these strains (except DZ-F23) are clinical.

268 We evaluated the presence of acquired genes in plasmids predicted with the PLSDB
269 database (version 2020_06_29) (Galata et al., 2019). *P. putida* GTC_16473 contained
270 the genes *aac(6')-IIa*, *aadA23*, and bla_{CARB-3} located in a scaffold with high identity with
271 the pJR2 plasmid from *Pasteurella multocida* (NC_004772.1). We also identified

272 *bla*$_{OXA-2}$, *aadA22*, *aac(6')-Ia*, *aac(6')-IIc*, *aph(3'')-Ib*, *aph(6')-Id*, *bla*$_{IMP-1}$, and *sul1* genes
273 in plasmid-like sequences in *P. putida* GTC_16473, although they are in contigs that
274 have not reached the coverage thresholds to be reliably classified as plasmids,
275 supporting an underestimation of plasmids in *P. putida* isolates. *P. putida* XWY-1
276 (CC6) also contains a plasmid, pXWY-1 (NZ_CP026333.1), which harbors the
277 resistance genes *sul1*, *aadA2,* and *qacH*. This strain was isolated from rice fields in
278 China, and has environmental importance. Finding non-clinical strains harboring
279 plasmids with such relevant resistance genes warrants further investigation.

280 Our results also highlight clinical strains harboring a range of resistance genes
281 (Fernandez et al., 2015; Mustapha et al., 2016). However, although these strains fall
282 in the *P. putida* group (Figure 1), they do not meet the ANI criteria to be classified as
283 *P. putida sensu stricto (P. alloputida)*. The only clinical strains that were isolated and
284 properly classified as *P. putida* were FDAARGOS_409 and 15420352. In addition,
285 other seven clinical strains included in our work were previously classified as
286 *Pseudomonas sp.* and were hitherto unrecognized as *P. putida sensu stricto* (Table 1).
287 Therefore, while we found clinical strains misclassified as *P. putida*, there are also
288 genuine *P. putida* isolates (e.g. GTC_16473) with multiple resistance genes that could
289 be considered outbreak candidates. In total, 9 out of 68 (13.2%) *P. putida* genomes
290 analyzed here belong to clinical strains.

291

292 **Virulence profiles**

293 We used the VFDB database (Liu et al., 2019) to assess the *P. putida* virulence
294 profiles. Although strains such as KT2440 have been approved as a host-vector safety

295 strain (Kampers et al., 2019), the genetic proximity with animal (*P. aeruginosa*) and
296 plant (*P. syringae*) pathogens reinforces the need to evaluate the biosafety of *P. putida*
297 species as a whole. The core virulome of *P. putida* contains genes associated with
298 twitching motility, siderophore production (pyoverdine), and alginate biosynthesis
299 (Table S6). Importantly, *P. putida* lacks key virulence genes usually found in *P.*
300 *aeruginosa*, such as those encoding exotoxin A, alkaline protease, elastase,
301 rhamnolipid biosynthesis components, phospholipase C, plant cell wall-degrading
302 enzymes, and type III secretion system. Although we have detected genes associated
303 with the type III secretion system in both core (HopAJ2 and HopJ1) and acquired
304 resistome (HopAN1), these genes were experimentally ruled out (Macho et al., 2009).

305 Most of the alginate regulatory and biosynthesis systems are present in the core
306 virulome of *P. putida* (Table S7). The overproduction of alginate plays an important
307 role in inducing the mucoid morphotype in *P. aeruginosa* during chronic infection
308 (Henry et al., 1992). However, the transcriptional regulatory gene *algM* (*mucC*),
309 associated with genotypic switching, is absent in *P. putida*. Since the lack of AlgM
310 leads to the non-mucoid phenotype in *P. aeruginosa* (Nelson et al., 2002), it can
311 account for the non-mucoid phenotype in *P. putida* under no water stress conditions.
312 Although the genotypic switching is unlikely to occur in *P. putida*, alginate production
313 is important under water stress conditions by promoting biofilm development and
314 protecting from desiccation (Chang et al., 2007).

315 The acquired virulome comprised genes for type II and VI secretions systems,
316 adherence, and iron uptake (Table S6). While most isolates had a low frequency of
317 resistance genes, the virulence factors displayed a bimodal distribution (Figure 5b), a
318 pattern that has been previously observed for *Klebsiella aerogenes* (Passarelli-Araujo

319 et al., 2019). Genes from the acinetobactin gene cluster and HSI-I type VI secretion
320 system were differentially distributed across CCs (Figure 6a), although it remains
321 unclear whether these patterns emerged mainly from gene gain or loss. *Acinetobacter*
322 *baumannii* has acinetobactin as the major siderophore associated with iron uptake
323 (Song and Kim, 2020), which is synthesized through the *bauABCDE* operon. We found
324 this operon in *P. putida* strains within CC5, CC6, and CC7 (Figure 6a). Further, this
325 operon was surrounded by genes coding for proteins from chemotaxis sensory
326 transducer (PP_2599) and aminotransferases (PP_2588) families (Figure 6b).
327 Siderophore-mediated iron acquisition has been explored in *P. putida* KT2440 (Matthijs
328 et al., 2009) (CC7), but the role played by *bauABCDE* in this species warrants further
329 experimental investigation.

330 Present in all strains from other CCs, the type VI secretion system (T6SS) HSI-I is
331 absent in both CC5 and CC6 (except XWY-1) isolates (Figure 6a). In KT2440 (CC7),
332 HSI-I is a potent weapon against other bacteria such as phytopathogens, increasing
333 the competitiveness of *P. putida* (Bernal et al., 2017). The absence of T6SS virulence
334 genes was observed in BIRD-1 (CC5) (Bernal et al., 2017), and our work generalizes
335 this observation for all CC5 members. In KT2440, T6SS is crucial to kill
336 phytopathogens such as *Xanthomonas campestris* (Validov et al., 2007). These results
337 indicate that BIRD-1 and other members from CC5 are likely less efficient than KT2440
338 as biocontrol agents. Moreover, all clinical *P. putida* strains harbor T6SS, indicating
339 their potential ability to outcompete other bacteria during infections.

340

341 **Plant growth-promotion and bioremediation properties**

342 The ability of *P. putida* to promote plant growth and bioremediate toxic compounds
343 have been explored (Nelson et al., 2002; Roca et al., 2013; Gong et al., 2016; Nikel
344 and de Lorenzo, 2018). We searched for genes involved in plant growth promotion and
345 bioremediation through a literature search of genes that have already been described
346 for *Pseudomonas* species (Table S7). We found genes in the core genome such as
347 the pyrroloquinoline quinone-encoding operon *pqqBCDEFG,* associated with mineral
348 phosphate solubilization in *Serratia marcescens* (Matteoli et al., 2018) and
349 *Pseudomonas fluorescens* (Miller et al., 2010) (Table S8). Mineral solubilization has
350 already been reported experimentally for BIRD-1 (Roca et al., 2013) and KT2440
351 (Miller et al., 2010), indicating that all *P. putida* isolates are genetically equipped to
352 solubilize inorganic phosphate.

353 Another key feature that can enhance plant growth is the colonization of seeds by
354 *P. putida* (Duque et al., 2013). In KT2440, genes associated with surface adhesion
355 (e.g. *lapA*, *lapBCD*), flagellum biosynthesis (e.g. *flhB*, *fliF*, *fliD*, *fliC*), and virulence
356 regulation (e.g. *rpoN* and *gacS*) have been experimentally shown to be important for
357 attachment to corn seeds (Duque et al., 2013). All these genes, except *lapA* and *fliC*,
358 belongs to the *P. putida* core genome (Table S8). Although often described as
359 virulence genes, flagellum genes also play roles in the association of *P. putida* with
360 plants. Other plant growth-promoting genes were also found in the *P. putida* accessory
361 genome (Table S8), except for CC1, which lacks all putative plant growth-promoting
362 genes.

363 Many plant growth-promoting bacteria can synthesize phytohormones such as
364 indole-3-acetic acid. *P. putida* lacks the main genes involved in this process (e.g. *ipdC*,
365 *iaaM*, and *iaaH*), indicating an incomplete or nonfunctional pathway. Although *P. putida*

366 W619 (GCA_000019445.1) is considered one of the most efficient producers of IAA in
367 comparison with other endophytic bacteria (Taghavi et al., 2009), this strain was
368 excluded from our analysis (Table S2) because it was likely misclassified as *P. putida*.
369 In addition, *P. putida* also lacks AcdS, an enzyme that counteract ethylene stress
370 response, a result that has been experimentally confirmed in KT2440 (Glick et al.,
371 2007).

372 Besides the ability to promote plant growth, *P. putida* can tolerate or degrade an
373 array of compounds including heavy metals and hydrocarbons. We identified genes in
374 the core genome that allow *P. putida* to resist various heavy metals such as copper
375 (*cop* genes) and cobalt/zinc/cadmium (*czcABC*) (Table S9). The copper/silver
376 resistance operon, *cusABC*, was present in the accessory genome. Although we found
377 a wide range of genes associated with bioremediation, there is no clear correlation
378 between the population structure and presence/absence profiles of such genes.

379 *P. putida* is known for its capacity to metabolize aromatic hydrocarbons such as
380 toluene, benzene, and *p*-cymene (Nelson et al., 2002; Wu et al., 2011). The toluene-
381 degrading pathway includes the *todABCDE* operon and the *todST* regulator. The *p*-
382 cymene compound can be degraded by means of the *cymAaAbBCDER* or the
383 *cmtAaAbAcAdBCDEFGHI* operon (Eaton, 1997). We found a genomic island of
384 approximately 48kb long harboring all these genes in F1, DOT-T1E, UV4, UV4/95,
385 YKD221, and NBRC_111125 genomes (Figure S9). All these isolates, except
386 NBRC_111125, were experimentally confirmed to degrade toluene. *P. putida* F1 is
387 best known to grow on toluene (Wu et al., 2011); YKD221 was isolated from
388 contaminated industrial soil and degrades *cis*-dichloroethene (Yonezuka et al., 2016);
389 DOT-T1E is an isolate known to grow on different carbon sources (Daniels et al., 2010)

390 and; UV4 and UV4/95 conduct important industrial biotransformation of arenes,
391 alkenes, and phenols (Skvortsov et al., 2018). Interestingly, genes in this genomic
392 island presented a very similar genetic context (Figure S7), with an upstream arm-type
393 integrase associated with bacteriophages. We were unable to precisely define the *att*
394 viral sites, indicating a deterioration of the original structure of the putative
395 bacteriophage. Further, the lack of correspondence between population structure and
396 the presence of an integrase upstream the genomic island indicates that this region
397 was likely acquired via independent horizontal gene transfers in distinct *P. putida* CCs.

398 We also identified RND efflux pumps involved with solvent tolerance in both core
399 (TtgABC) and accessory genomes (TtgDEF and TtgGHI). TtgABC, TtgDEF, and
400 TtgGHI are required for DOT-T1E to efficiently tolerate toluene (Rojas et al., 2001). We
401 observed that TtgABC is the same protein-complex predicted as MexAB-OprM,
402 associated with antibiotic resistance in the core genome. This complex extrudes both
403 antibiotics and solvents such as toluene in *P. putida* DOT-T1E (Rojas et al., 2001),
404 corroborating the additional and important function to extrude antibiotics and organic
405 solvents in all *P. putida* isolates. TtgDEF is located in the genomic island co-occurring
406 with *tod* genes. This complex can expel toluene, but not antibiotics (Mosqueda and
407 Ramos, 2000), reinforcing the broad different molecules that can be used as substrates
408 for RND efflux pumps, and the need to explore what determines this specificity in *P.*
409 *putida* isolates and other bacteria.

410

411 **CONCLUDING REMARKS**

412 Through a remarkable metabolic versatility, *P. putida* can thrive in a wide variety of
413 niches. In this work, we explored the genetic diversity of *P. putida sensu stricto* (*P.*
414 *alloputida*) and characterized its population structure for the first time. Through a large-
415 scale genomic analysis, we identified a major problem with *P. putida* classification,
416 including several reference strains that likely belong to new species, as suggested
417 elsewhere (Keshavarz-Tohid et al., 2019). *P. putida* has an open pangenome
418 dominated by low-frequency genes. The population structure of this species has at
419 least 7 clonal complexes that were verified by cgMLST and STRUCTURE ancestry
420 simulations. We found no association of clinical isolates with a specific CC. Together,
421 our results indicate that *P. putida* clinical isolates are mainly opportunistic and do not
422 pose considerable health concerns.

423 We analyzed genes of clinical and biotechnological interest. *P. putida* has several
424 RND-family efflux pumps to tolerate antibiotics and other toxic compounds such as
425 toluene. The low-frequency acquired resistance genes are predominant in plasmids
426 from a few clinical strains. The detected virulence genes allow *P. putida* to synthesize
427 pyoverdine, attach to seed surfaces, and kill other bacteria (including pathogens)
428 through a type VI secretion system. KT2440 is a member of CC7 that has the T6SS
429 along with an operon for acinetobactin biosynthesis. *P. putida* lacks key genes for the
430 production of indole-3-acetic acid. We also observed that the genes for the degradation
431 of some aromatic compounds, including toluene, were likely horizontally acquired. Our
432 results provide an opportunity for the development of biotechnological applications as
433 well as insights into the genomic diversity of *P. putida*.

434

435 **MATERIALS AND METHODS**

436 **Dataset and genomic features**

437 We recovered 11,025 genomes from the *Pseudomonas* genus in June 2020. To
438 assess the quality of the genomes, we used BUSCO v4.0.6 (Seppey et al., 2019) with
439 a threshold of 90% completeness. The KT2440 (GCA_900167985.1) was used as a
440 reference with mash v.2.2.2 (Ondov et al., 2016) to obtain genomes with a distance
441 lower than 0.05. We used mashtree (Katz et al., 2019) to generate the distance tree.
442 The ANI analysis was performed with pyani 0.2.10 (Pritchard et al., 2016). Network
443 analysis was conducted in R with the igraph package (https://igraph.org). We removed
444 both S12 (GCA_000287915.1) and KT2440 (GCA_000007565.2) because they were
445 duplicated genomes. Gene prediction in all isolates was conducted with prokka v1.12
446 (Seemann, 2014) to avoid bias in the identification of protein families. Plasmids were
447 analyzed PLSD v2020_06_29 (Galata et al., 2019).

448 **Pangenome characterization**

449 We inferred the *P. putida* pangenome using Roary 3.13.0 (Page et al., 2015), with
450 a minimum threshold of 85% identity to cluster proteins. Core genes were defined as
451 those present in more than 95% of the isolates. Jaccard distances were computed by
452 using accessory genes with prevalence between 5% and 95%. Multivariate analysis
453 was performed to assess gene content variations between ecotypes from *P. putida*
454 through DAPC analysis using the *ade4* and *adegenet* packages (Jombart et al., 2010),
455 retaining the 30 principal components and 3 discriminant functions. Pangenome
456 openness and fluidity were conducted with micropan (Snipen and Liland, 2015) with
457 500 and 1000 permutations, respectively.

458 **Population structure analysis**

459 We used *in-house* scripts to extract the genes present in all isolates, which were
460 aligned with MAFFT v7.467 (Katoh and Standley, 2013). SNPs were retrieved with
461 snp-sites v2.3.3 (Page et al., 2016) and SNP alignment was used as input to RAxML
462 v8 (Stamatakis, 2014) to reconstruct the phylogenetic tree using the general time-
463 reversible model and gamma correction. Since we used only variable sites as input,
464 we used ASC_GTRGAMMA to correct ascertainment bias with the Paul Lewis
465 correction. One thousand bootstrap replicates were generated to assess the
466 significance of internal nodes. We inferred the cgMLST scheme using the core genome
467 SNP phylogenetic tree. The phylogenetic tree was visualized with iTOL v4 (Letunic and
468 Bork, 2019).

469 We downloaded the *P. putida* MLST scheme (on June, 2020) containing 116
470 different STs (https://pubmlst.org/databases/). This scheme was designed for the
471 whole *P. putida* group, not only *P. putida* sensu stricto (Ogura et al., 2019). We used
472 BLASTN (Altschul et al., 1997) to determine the best-matching MLST allele to access
473 STs. The allelic profile associated with each ST in our dataset was used to conduct
474 population assignment with STRUCTURE v2.3.4 (Pritchard et al., 2000) with admixture
475 model. The length of Markov chain Monte Carlo (MCMC) was 50,000, discarding
476 20,000 iterations as burn-in. The simulations to calculate the parameter K ranged from
477 2 to 20, with 20 replicates for each K to estimate confidence intervals. Instead of using
478 raw posterior probability to get the best K, we followed the protocol suggested by
479 Evanno et al. (2005). Briefly, we calculated the first and second derivatives to estimate
480 a ΔK of 7. Therefore, we used K = 7 to analyze predicted ancestry probabilities.

481

482 **Detection of genes associated with resistance, virulence, plant growth**
483 **promotion, and bioremediation**

484 We used the Comprehensive Antimicrobial Resistance Database (CARD) database
485 v3.0.9 (Alcock et al., 2020) to predict antibiotic resistance genes. The virulence factor
486 database (VFDB) (Liu et al., 2019) was used to determine virulence genes. This
487 database was downloaded on July 30 2020 and comprises 28,639 proteins associated
488 with virulence in several pathogens. We used virulence genes previously described for
489 the *Pseudomonas* genus. We clustered proteins based on 70% identity to retrieve a
490 non-redundant database using uclust v1.2.22q (Edgar, 2010). We built the database
491 with plant growth promotion and bioremediation through literature searches (Table S7).
492 All predicted proteins were globally aligned against these databases using usearch
493 v11.0.667 (Edgar, 2010) with 50% coverage for both query and subject and minimum
494 60% identity.

495 **CONFLICT OF INTEREST**

496 The authors declare no conflict of interest.

497 **ACKNOWLEDGEMENTS**

498 This work was supported by Fundação Carlos Chagas Filho de Amparo à Pesquisa do
499 Estado do Rio de Janeiro (FAPERJ; grants E-26/203.309/2016 and E-
500 26/203.014/2018), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior -
501 Brasil (CAPES; Finance Code 001), and Conselho Nacional de Desenvolvimento
502 Científico e Tecnológico. The funding agencies had no role in the design of the study
503 and collection, analysis, and interpretation of data and in writing.

504

505 **REFERENCES**

506 Ait Tayeb, L., Ageron, E., Grimont, F., and Grimont, P.A. (2005) Molecular phylogeny of the genus
507 Pseudomonas based on rpoB sequences and application for the identification of isolates. *Res Microbiol*
508 **156**: 763-773.
509 Alcock, B.P., Raphenya, A.R., Lau, T.T.Y., Tsang, K.K., Bouchard, M., Edalatmand, A. et al. (2020)
510 CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database.
511 *Nucleic Acids Res* **48**: D517-D525.
512 Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997)
513 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids*
514 *Res* **25**: 3389-3402.
515 Belda, E., van Heck, R.G., Jose Lopez-Sanchez, M., Cruveiller, S., Barbe, V., Fraser, C. et al. (2016)
516 The revisited genome of Pseudomonas putida KT2440 enlightens its value as a robust metabolic
517 chassis. *Environ Microbiol* **18**: 3403-3424.
518 Bernal, P., Allsopp, L.P., Filloux, A., and Llamas, M.A. (2017) The Pseudomonas putida T6SS is a
519 plant warden against phytopathogens. *ISME J* **11**: 972-987.
520 Bobay, L.M. (2020) The Prokaryotic Species Concept and Challenges. In *The Pangenome: Diversity,*
521 *Dynamics and Evolution of Genomes*. Tettelin, H., and Medini, D. (eds). Cham (CH), pp. 21-49.
522 Chang, W.S., van de Mortel, M., Nielsen, L., Nino de Guzman, G., Li, X., and Halverson, L.J. (2007)
523 Alginate production by Pseudomonas putida creates a hydrated microenvironment and contributes to
524 biofilm architecture and stress tolerance under water-limiting conditions. *J Bacteriol* **189**: 8290-8299.
525 Daniels, C., Godoy, P., Duque, E., Molina-Henares, M.A., de la Torre, J., Del Arco, J.M. et al. (2010)
526 Global regulation of food supply by Pseudomonas putida DOT-T1E. *J Bacteriol* **192**: 2169-2181.
527 de Been, M., Pinholt, M., Top, J., Bletz, S., Mellmann, A., van Schaik, W. et al. (2015) Core Genome
528 Multilocus Sequence Typing Scheme for High- Resolution Typing of Enterococcus faecium. *J Clin*
529 *Microbiol* **53**: 3788-3797.
530 Duque, E., de la Torre, J., Bernal, P., Molina-Henares, M.A., Alaminos, M., Espinosa-Urgel, M. et
531 al. (2013) Identification of reciprocal adhesion genes in pathogenic and non-pathogenic Pseudomonas.
532 *Environ Microbiol* **15**: 36-48.
533 Eaton, R.W. (1997) p-Cymene catabolic pathway in Pseudomonas putida F1: cloning and
534 characterization of DNA encoding conversion of p-cymene to p-cumate. *J Bacteriol* **179**: 3171-3180.
535 Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*
536 **26**: 2460-2461.
537 Evanno, G., Regnaut, S., and Goudet, J. (2005) Detecting the number of clusters of individuals
538 using the software STRUCTURE: a simulation study. *Mol Ecol* **14**: 2611-2620.
539 Evans, K., Passador, L., Srikumar, R., Tsang, E., Nezezon, J., and Poole, K. (1998) Influence of the
540 MexAB-OprM multidrug efflux system on quorum sensing in Pseudomonas aeruginosa. *J Bacteriol* **180**:
541 5443-5447.
542 Fernandez, M., Porcel, M., de la Torre, J., Molina-Henares, M.A., Daddaoua, A., Llamas, M.A. et al.
543 (2015) Analysis of the pathogenic potential of nosocomial Pseudomonas putida strains. *Front Microbiol*
544 **6**: 871.
545 Galata, V., Fehlmann, T., Backes, C., and Keller, A. (2019) PLSDB: a resource of complete bacterial
546 plasmids. *Nucleic Acids Res* **47**: D195-D202.
547 Glick, B.R., Todorovic, B., Czarny, J., Cheng, Z., Duan, J., and McConkey, B. (2007) Promotion of
548 Plant Growth by Bacterial ACC Deaminase. *Critical Reviews in Plant Sciences* **26**.
549 Gomila, M., Pena, A., Mulet, M., Lalucat, J., and Garcia-Valdes, E. (2015) Phylogenomics and
550 systematics in Pseudomonas. *Front Microbiol* **6**: 214.
551 Gong, T., Liu, R., Che, Y., Xu, X., Zhao, F., Yu, H. et al. (2016) Engineering Pseudomonas putida
552 KT2440 for simultaneous degradation of carbofuran and chlorpyrifos. *Microb Biotechnol* **9**: 792-800.
553 Henry, R.L., Mellis, C.M., and Petrovic, L. (1992) Mucoid Pseudomonas aeruginosa is a marker of
554 poor survival in cystic fibrosis. *Pediatr Pulmonol* **12**: 158-161.

555 Jolley, K.A., Bray, J.E., and Maiden, M.C.J. (2018) Open-access bacterial population genomics:
556 BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* **3**: 124.
557 Jombart, T., Devillard, S., and Balloux, F. (2010) Discriminant analysis of principal components: a
558 new method for the analysis of genetically structured populations. *BMC Genet* **11**: 94.
559 Kampers, L.F.C., Volkers, R.J.M., and Martins Dos Santos, V.A.P. (2019) Pseudomonas putida
560 KT2440 is HV1 certified, not GRAS. *Microb Biotechnol* **12**: 845-848.
561 Katoh, K., and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7:
562 improvements in performance and usability. *Mol Biol Evol* **30**: 772-780.
563 Katz, L.S., Griswold, T., Morrison, S.S., Caravas, J.A., Zhang, S., den Bakker, H.C. et al. (2019)
564 Mashtree: a rapid comparison of whole genome sequence files. *Journal of Open Source Software* **4**.
565 Keshavarz-Tohid, V., Vacheron, J., Dubost, A., Prigent-Combaret, C., Taheri, P., Tarighi, S. et al.
566 (2019) Genomic, phylogenetic and catabolic re-assessment of the Pseudomonas putida clade supports
567 the delineation of Pseudomonas alloputida sp. nov., Pseudomonas inefficax sp. nov., Pseudomonas
568 persica sp. nov., and Pseudomonas shirazica sp. nov. *Syst Appl Microbiol* **42**: 468-480.
569 Kislyuk, A.O., Haegeman, B., Bergman, N.H., and Weitz, J.S. (2011) Genomic fluidity: an integrative
570 view of gene diversity within microbial populations. *BMC Genomics* **12**: 32.
571 Letunic, I., and Bork, P. (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new
572 developments. *Nucleic Acids Res* **47**: W256-W259.
573 Lister, P.D., Wolter, D.J., and Hanson, N.D. (2009) Antibacterial-resistant Pseudomonas
574 aeruginosa: clinical impact and complex regulation of chromosomally encoded resistance mechanisms.
575 *Clin Microbiol Rev* **22**: 582-610.
576 Liu, B., Zheng, D., Jin, Q., Chen, L., and Yang, J. (2019) VFDB 2019: a comparative pathogenomic
577 platform with an interactive web interface. *Nucleic Acids Res* **47**: D687-D692.
578 Macho, A.P., Ruiz-Albert, J., Tornero, P., and Beuzon, C.R. (2009) Identification of new type III
579 effectors and analysis of the plant response by competitive index. *Mol Plant Pathol* **10**: 69-80.
580 Matteoli, F.P., Passarelli-Araujo, H., Reis, R.J.A., da Rocha, L.O., de Souza, E.M., Aravind, L. et al.
581 (2018) Genome sequencing and assessment of plant growth-promoting properties of a Serratia
582 marcescens strain isolated from vermicompost. *BMC Genomics* **19**: 750.
583 Matthijs, S., Laus, G., Meyer, J.M., Abbaspour-Tehrani, K., Schafer, M., Budzikiewicz, H., and
584 Cornelis, P. (2009) Siderophore-mediated iron acquisition in the entomopathogenic bacterium
585 Pseudomonas entomophila L48 and its close relative Pseudomonas putida KT2440. *Biometals* **22**: 951-
586 964.
587 Miller, S.H., Browne, P., Prigent-Combaret, C., Combes-Meynet, E., Morrissey, J.P., and O'Gara, F.
588 (2010) Biochemical and genomic comparison of inorganic phosphate solubilization in Pseudomonas
589 species. *Environ Microbiol Rep* **2**: 403-411.
590 Molina, L., Udaondo, Z., Duque, E., Fernandez, M., Bernal, P., Roca, A. et al. (2016) Specific Gene
591 Loci of Clinical Pseudomonas putida Isolates. *PLoS One* **11**: e0147478.
592 Molina, L., Udaondo, Z., Duque, E., Fernandez, M., Molina-Santiago, C., Roca, A. et al. (2014)
593 Antibiotic resistance determinants in a Pseudomonas putida strain isolated from a hospital. *PLoS One*
594 **9**: e81604.
595 Mosqueda, G., and Ramos, J.L. (2000) A set of genes encoding a second toluene efflux system in
596 Pseudomonas putida DOT-T1E is linked to the tod genes for toluene metabolism. *J Bacteriol* **182**: 937-
597 943.
598 Mustapha, M.M., Marsh, J.W., Ezeonwuka, C.D., Pasculle, A.W., Pacey, M.P., Querry, A.M. et al.
599 (2016) Draft Genome Sequences of Four Hospital-Associated Pseudomonas putida Isolates. *Genome*
600 *Announc* **4**.
601 Nelson, K.E., Weinel, C., Paulsen, I.T., Dodson, R.J., Hilbert, H., Martins dos Santos, V.A. et al.
602 (2002) Complete genome sequence and comparative analysis of the metabolically versatile
603 Pseudomonas putida KT2440. *Environ Microbiol* **4**: 799-808.
604 Nikel, P.I., and de Lorenzo, V. (2018) Pseudomonas putida as a functional chassis for industrial
605 biocatalysis: From native biochemistry to trans-metabolism. *Metab Eng* **50**: 142-155.

606 Ogura, K., Shimada, K., and Miyoshi-Akiyama, T. (2019) A multilocus sequence typing scheme of
607 Pseudomonas putida for clinical and environmental isolates. *Sci Rep* **9**: 13980.
608 Ohji, S., Yamazoe, A., Hosoyama, A., Tsuchikane, K., Ezaki, T., and Fujita, N. (2014) The Complete
609 Genome Sequence of Pseudomonas putida NBRC 14164T Confirms High Intraspecies Variation.
610 *Genome Announc* **2**.
611 Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy,
612 A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*
613 **17**: 132.
614 Page, A.J., Taylor, B., Delaney, A.J., Soares, J., Seemann, T., Keane, J.A., and Harris, S.R. (2016) SNP-
615 sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* **2**: e000056.
616 Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T. et al. (2015) Roary: rapid
617 large-scale prokaryote pan genome analysis. *Bioinformatics* **31**: 3691-3693.
618 Palleroni, N.J. (2010) The Pseudomonas story. *Environ Microbiol* **12**: 1377-1383.
619 Passarelli-Araujo, H., Palmeiro, J.K., Moharana, K.C., Pedrosa-Silva, F., Dalla-Costa, L.M., and
620 Venancio, T.M. (2019) Genomic analysis unveils important aspects of population structure, virulence,
621 and antimicrobial resistance in Klebsiella aerogenes. *FEBS J* **286**: 3797-3810.
622 Pritchard, J.K., Stephens, M., and Donnelly, P. (2000) Inference of population structure using
623 multilocus genotype data. *Genetics* **155**: 945-959.
624 Pritchard, L., Glover, H.R., Humphris, S., Elphinstone, J.G., and Toth, I.K. (2016) Genomics and
625 taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods*
626 **8**: 10-24.
627 Richter, M., and Rossello-Mora, R. (2009) Shifting the genomic gold standard for the prokaryotic
628 species definition. *Proc Natl Acad Sci U S A* **106**: 19126-19131.
629 Roca, A., Pizarro-Tobias, P., Udaondo, Z., Fernandez, M., Matilla, M.A., Molina-Henares, M.A. et
630 al. (2013) Analysis of the plant growth-promoting properties encoded by the genome of the
631 rhizobacterium Pseudomonas putida BIRD-1. *Environ Microbiol* **15**: 780-794.
632 Rojas, A., Duque, E., Mosqueda, G., Golden, G., Hurtado, A., Ramos, J.L., and Segura, A. (2001)
633 Three efflux pumps are required to provide efficient tolerance to toluene in Pseudomonas putida DOT-
634 T1E. *J Bacteriol* **183**: 3967-3973.
635 Rosson, R.A., and Nealson, K.H. (1982) Manganese binding and oxidation by spores of a marine
636 bacillus. *J Bacteriol* **151**: 1027-1034.
637 Sanchez, P., Linares, J.F., Ruiz-Diez, B., Campanario, E., Navas, A., Baquero, F., and Martinez, J.L.
638 (2002) Fitness of in vitro selected Pseudomonas aeruginosa nalB and nfxB multidrug resistant mutants.
639 *J Antimicrob Chemother* **50**: 657-664.
640 Schweizer, H.P. (2003) Efflux as a mechanism of resistance to antimicrobials in Pseudomonas
641 aeruginosa and related bacteria: unanswered questions. *Genet Mol Res* **2**: 48-62.
642 Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**: 2068-2069.
643 Seppey, M., Manni, M., and Zdobnov, E.M. (2019) BUSCO: Assessing Genome Assembly and
644 Annotation Completeness. *Methods Mol Biol* **1962**: 227-245.
645 Skvortsov, T., Hoering, P., Arkhipova, K., Whitehead, R.C., Boyd, D.R., and Allen, C.C.R. (2018) Draft
646 Genome Sequences of Pseudomonas putida UV4 and UV4/95, Toluene Dioxygenase-Expressing
647 Producers of cis-1,2-Dihydrodiols. *Genome Announc* **6**.
648 Snipen, L., and Liland, K.H. (2015) micropan: an R-package for microbial pan-genomics. *BMC*
649 *Bioinformatics* **16**: 79.
650 Song, W.Y., and Kim, H.J. (2020) Current biochemical understanding regarding the metabolism of
651 acinetobactin, the major siderophore of the human pathogen Acinetobacter baumannii, and outlook
652 for discovery of novel anti-infectious agents based thereon. *Nat Prod Rep* **37**: 477-487.
653 Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
654 phylogenies. *Bioinformatics* **30**: 1312-1313.
655 Stanier, R.Y., Palleroni, N.J., and Doudoroff, M. (1966) The aerobic pseudomonads: a taxonomic
656 study. *J Gen Microbiol* **43**: 159-271.

657 Taghavi, S., Garafola, C., Monchy, S., Newman, L., Hoffman, A., Weyens, N. et al. (2009) Genome
658 survey and characterization of endophytic bacteria exhibiting a beneficial effect on growth and
659 development of poplar trees. *Appl Environ Microbiol* **75**: 748-757.
660 Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L. et al. (2005) Genome
661 analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial
662 "pan-genome". *Proc Natl Acad Sci U S A* **102**: 13950-13955.
663 Tian, Z.X., Yi, X.X., Cho, A., O'Gara, F., and Wang, Y.P. (2016) CpxR Activates MexAB-OprM Efflux
664 Pump Expression and Enhances Antibiotic Resistance in Both Laboratory and Clinical nalB-Type Isolates
665 of Pseudomonas aeruginosa. *PLoS Pathog* **12**: e1005932.
666 Timmis, K.N. (2002) Pseudomonas putida: a cosmopolitan opportunist par excellence. *Environ*
667 *Microbiol* **4**: 779-781.
668 Tohya, M., Watanabe, S., Teramoto, K., Uechi, K., Tada, T., Kuwahara-Arai, K. et al. (2019)
669 Pseudomonas asiatica sp. nov., isolated from hospitalized patients in Japan and Myanmar. *Int J Syst*
670 *Evol Microbiol* **69**: 1361-1368.
671 Udaondo, Z., Molina, L., Segura, A., Duque, E., and Ramos, J.L. (2016) Analysis of the core genome
672 and pangenome of Pseudomonas putida. *Environ Microbiol* **18**: 3268-3283.
673 Validov, S., Kamilova, F., Qi, S., Stephan, D., Wang, J.J., Makarova, N., and Lugtenberg, B. (2007)
674 Selection of bacteria able to control Fusarium oxysporum f. sp. radicis-lycopersici in stonewool
675 substrate. *J Appl Microbiol* **102**: 461-471.
676 Volke, D.C., Calero, P., and Nikel, P.I. (2020) Pseudomonas putida. *Trends Microbiol* **28**: 512-513.
677 Wu, X., Monchy, S., Taghavi, S., Zhu, W., Ramos, J., and van der Lelie, D. (2011) Comparative
678 genomics and functional analysis of niche-specific adaptation in Pseudomonas putida. *FEMS Microbiol*
679 *Rev* **35**: 299-323.
680 Yonezuka, K., Shimodaira, J., Tabata, M., Ohji, S., Hosoyama, A., Kasai, D. et al. (2017) Phylogenetic
681 analysis reveals the taxonomically diverse distribution of the Pseudomonas putida group. *J Gen Appl*
682 *Microbiol* **63**: 1-10.
683 Yonezuka, K., Araki, N., Shimodaira, J., Ohji, S., Hosoyama, A., Numata, M. et al. (2016) Isolation
684 and characterization of a bacterial strain that degrades cis-dichloroethenein the absence of aromatic
685 inducers. *J Gen Appl Microbiol* **62**: 118-125.
686

687 ## Tables

688 **Table 1.** *Pseudomonas putida* isolates used in this study with average nucleotide
689 identity (ANI) from KT2440.

Strain	Classified as	ANI (%)	CC	Location	Source	Accession
15420352	*P. putida*	97.12	CC4	China	urine	GCA_013305625.1
B2017	*P. putida*	97.25	CC3	Spain	root	GCA_007279645.1
B4	*P. putida*	97.26	CC3	China	soil	GCA_003671955.1
B6-2	*P. putida*	97.12	NOCC	-	soil	GCA_000226035.3
BIRD-1	*P. putida*	97.49	CC5	Spain	rhizosphere	GCA_000183645.1
CA-3	*P. putida*	97.22	CC4	Ireland	waste material	GCA_002810225.1
CY06	*P. monteilii*	95.72	CC1	China	shrimp	GCA_002835905.1
DOT-T1E	*P. putida*	97.28	CC3	Spain	wastewater	GCA_000281215.1
DPA1	*P. putida*	97.22	CC4	Greece	soil	GCA_002891885.1
DZ-F23	*P. putida*	95.68	CC1	China	fly	GCA_002094775.1

F1	*P. putida*	97.19	CC2	USA	soil	GCA_000016865.1
FDAARGOS_409	*P. putida*	97.11	CC4	USA	blood	GCA_002554535.1
FF4	*Pseudomonas sp.*	97.14	NOCC	Chile	wastewater	GCA_007049805.1
FW305-E2	*P. putida*	97.16	CC2	USA	groundwater	GCA_900095365.1
FW305-E2_1	*Pseudomonas sp.*	97.16	CC2	USA	groundwater	GCA_002901725.1
GTC_16473	*Pseudomonas sp.*	99.87	CC7	Japan	*Homo sapiens*	GCA_001753855.1
GTC_16482	*Pseudomonas sp.*	99.83	CC7	Japan	*Homo sapiens*	GCA_001319995.1
H	*P. putida*	99.99	CC7	Germany	soil	GCA_001077495.1
Idaho	*P. putida*	98.29	CC6	China	-	GCA_000226475.2
INSali382	*P. putida*	98.20	CC6	Portugal	vegetable	GCA_001653615.1
IOFA1	*P. putida*	95.73	CC1	Indian Ocean	deep-sea sediment	GCA_001293025.1
IOFA19	*P. monteilii*	95.62	CC1	Indian Ocean	deep sea sediment	GCA_000633915.1
JB	*P. putida*	97.44	CC5	Czech Republic	soil	GCA_001767335.1
JLR11	*P. putida*	99.97	CC7	Spain	wastewater	GCA_001183585.1
JR16	*P. putida*	97.46	CC5	India	soil	GCA_004519745.1
JY-Q	*Pseudomonas sp.*	97.20	CC4	China	tabaco extract	GCA_001655295.1
KB3	*P. putida*	95.73	CC1	Poland	soil	GCA_004614175.1
KBS0802	*Pseudomonas sp.*	99.98	CC7	USA	soil	GCA_005937845.2
KCJK7911	*P. putida*	97.25	CC3	USA	water	GCA_003053335.1
KCJK7916	*P. putida*	97.20	NOCC	USA	water	GCA_003053385.1
KH-18-2	*P. putida*	95.67	CC1	Pacific Ocean	water	GCA_002906815.1
KH-20-11	*P. putida*	95.69	CC1	Pacific Ocean	water	GCA_002906795.1
Kh14	*Pseudomonas sp.*	97.40	CC5	Iran	rhizosphere	GCA_900291005.1
Kh7	*Pseudomonas sp.*	97.39	CC5	Iran	rhizosphere	GCA_900291035.1
KT-90	*P. putida*	95.74	CC1	Pacific Ocean	coastal water	GCA_002906755.1
KT2440	*P. putida*	100.00	CC7	Japan	rhizosphere	GCA_900167985.1
LD6	*P. putida*	97.45	CC5	China	rhizosphere	GCA_003586135.1
LF54	*P. putida*	95.67	CC1	Japan	lotus field	GCA_000390005.2
LS46	*P. putida*	97.18	CC4	Canada	water	GCA_000294445.2
N1R	*P. putida*	97.39	CC5	EUA	soil	GCA_900156185.1
NBRC_111118	*Pseudomonas sp.*	97.23	CC4	Japan	*Homo sapiens*	GCA_001320085.1
NBRC_111121	*Pseudomonas sp.*	99.91	CC7	Japan	sputum	GCA_001320165.1
NBRC_111125	*Pseudomonas sp.*	97.20	CC4	Japan	urine	GCA_001320295.1
NBRC_111136	*Pseudomonas sp.*	97.17	CC4	Japan	urine	GCA_001320745.1
NBRC_111139	*Pseudomonas sp.*	97.15	CC4	Japan	eye discharge	GCA_001753955.1
NCTC13185	*P. putida*	99.99	CC7	-	-	GCA_901482375.1
NCTC13186	*P. putida*	99.99	CC7	-	-	GCA_900636645.1
ND6	*P. putida*	97.10	CC2	China	wastewater	GCA_000264665.2
ODNR4SY	*P. putida*	97.18	CC2	USA	water	GCA_009905395.1
OR45a	*P. putida*	97.44	CC5	Poland	activated sludge	GCA_004614155.1
P11	*P. hunanensis*	97.12	CC4	China	soil	GCA_002910975.1
PaW85-2019	*P. putida*	99.99	CC7	Estonia	-	GCA_011750655.1
PaW85-d13TA	*P. putida*	99.99	CC7	Estonia	-	GCA_011750675.1
PCL1760	*P. putida*	97.45	CC5	Spain	rhizosphere	GCA_001282125.1
PD1	*P. putida*	97.27	CC3	USA	root	GCA_000799625.1
RW4053	*Pseudomonas sp.*	97.44	CC5	Germany	river sediments	GCA_003184135.1

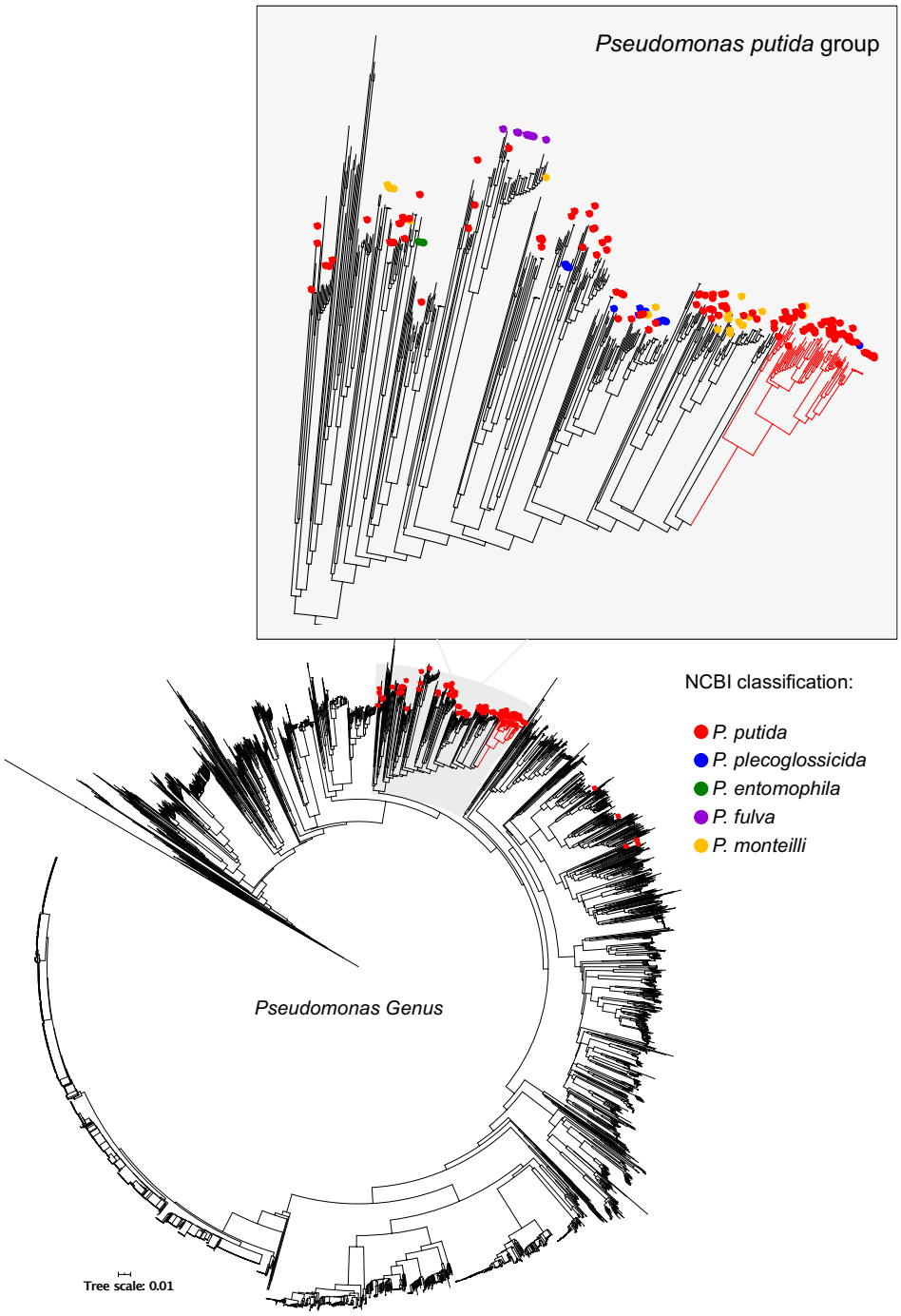
S12	*P. putida*	97.48	CC5	Netherlands	soil	GCA_000495455.2
SJTE-1	*P. putida*	97.17	CC2	China	soil	GCA_000271965.2
SMT-1	*Pseudomonas sp.*	99.99	CC7	China	soil	GCA_003204195.1
SWI36	*Pseudomonas sp.*	98.16	CC6	USA	soil	GCA_002948105.1
SWI36_1	*Pseudomonas sp.*	98.22	CC6	USA	soil	GCA_004153505.1
SWI36_2	*Pseudomonas sp.*	98.22	CC6	USA	soil	GCA_004153435.1
TRO1	*P. putida*	97.17	CC4	Denmark	activated sludge	GCA_000367825.1
UV4	*P. putida*	97.17	CC4	UK	laboratory strain	GCA_002165695.1
UV4/95	*P. putida*	97.17	CC4	UK	laboratory strain	GCA_002165665.1
XWY-1	*Pseudomonas sp.*	98.16	CC6	China	rice fields	GCA_002953115.1
YKD221	*P. putida*	97.23	CC2	Japan	soil	GCA_000787655.1
ZKA3	*P. plecoglossicida*	98.18	CC6	Greece	water	GCA_003633555.1

Table 2. Frequency of resistance mechanisms categories in both core and accessory resistome.

Pangenome division	**Resistance mechanism**	**Frequency**
Core	Antibiotic efflux	15 genes* (100%)
Accessory	Antibiotic efflux	126 genes (67.74%)
	Antibiotic inactivation	47 genes (25.27%)
	Antibiotic target alteration	3 genes (1.61%)
	Antibiotic target protection	1 gene (0.54%)
	Antibiotic target replacement	9 genes (4.84%)

*Number of different genes.

694 **Figures**



695

696 **Figure 1.** Distance tree of 10,457 *Pseudomonas* genomes from *Pseudomonas* genus.
697 Genomes classified as *P. putida* according to NCBI are marked as red circle. *P. putida*
698 group was highlighted to assess the distribution of misclassified genomes. *P. putida*
699 (*P. alloputida*) branches are colored in red.

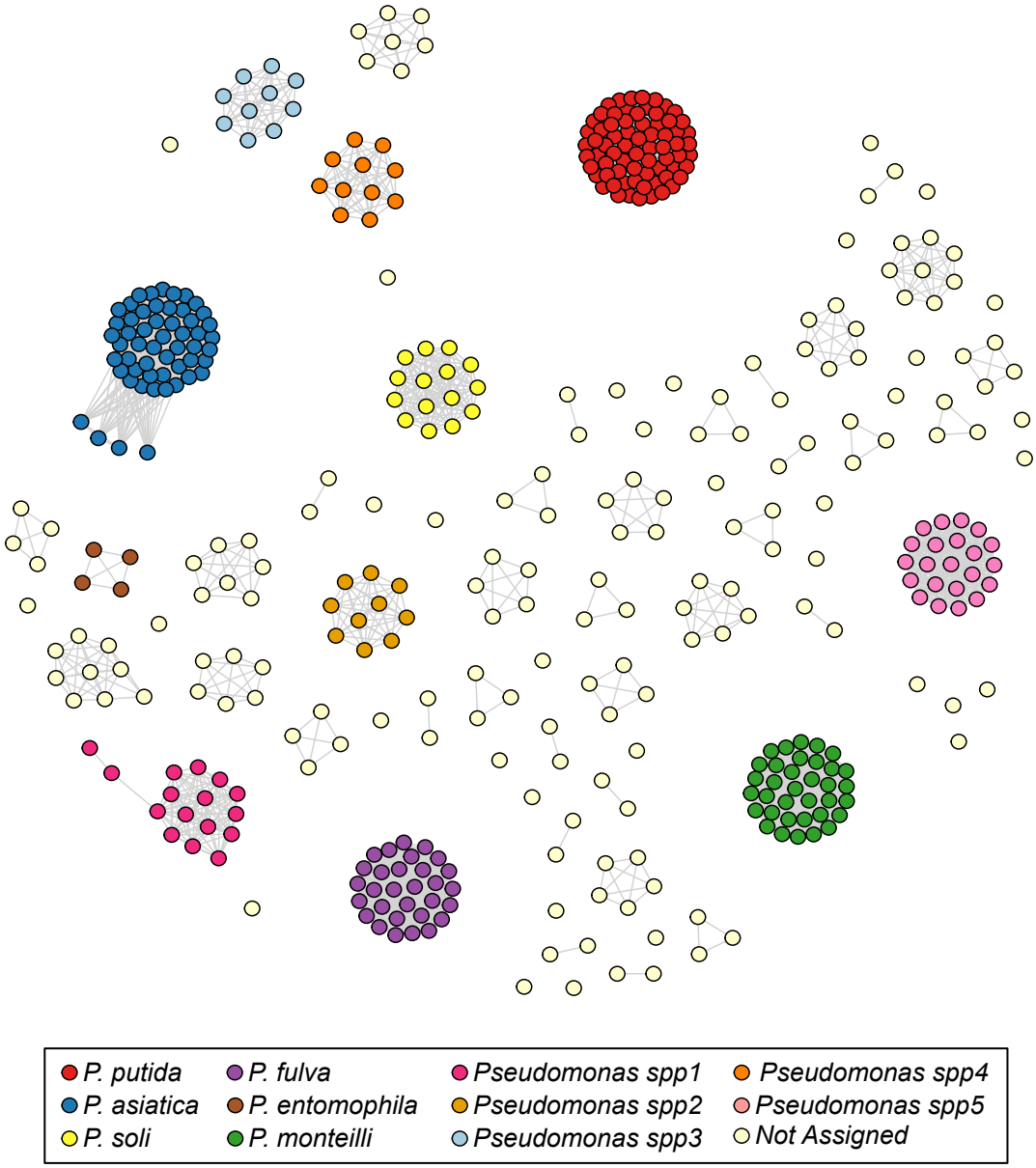
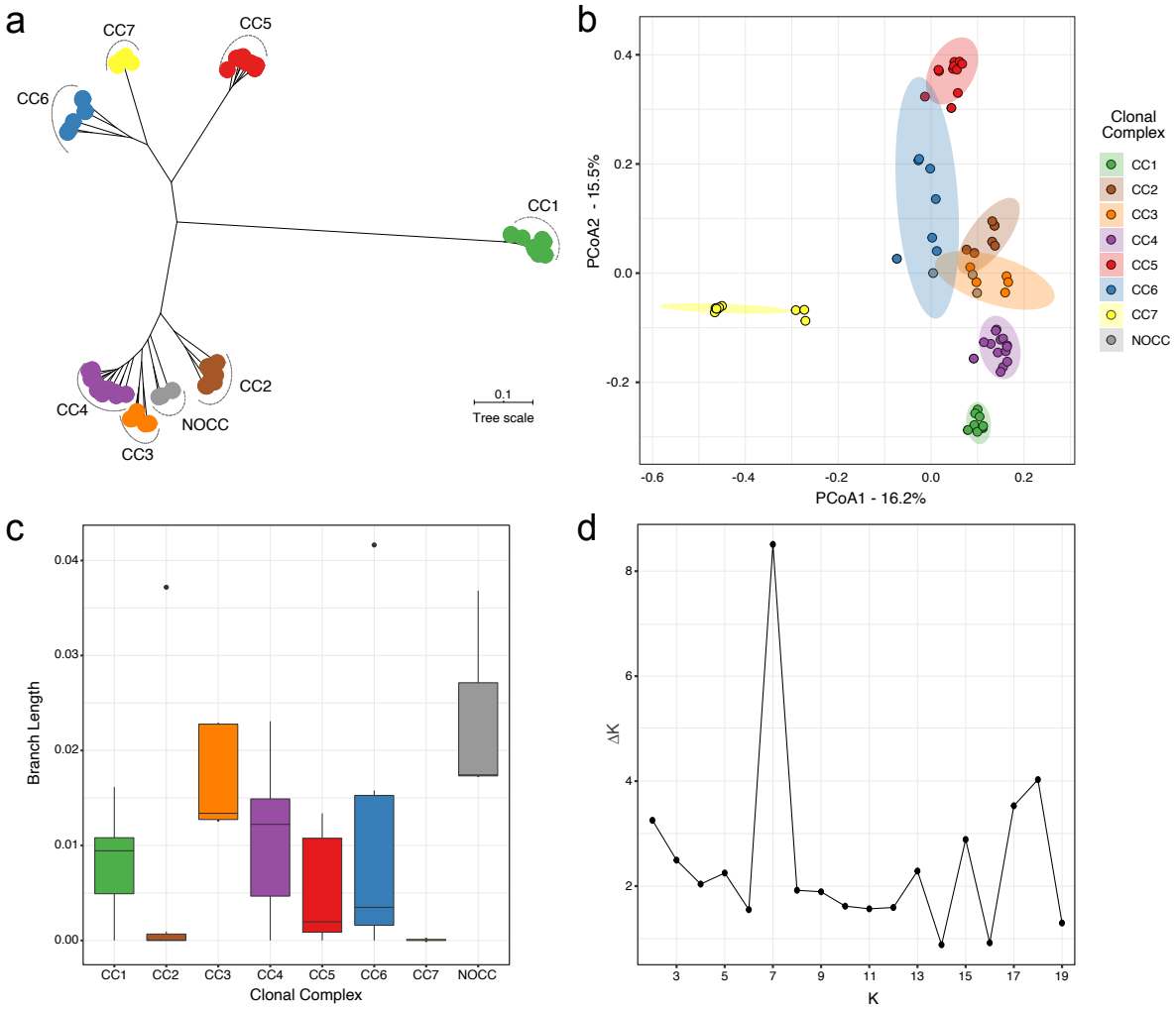


Figure 2. Network analysis of isolates from the *P. putida* group. Nodes represent isolates and edges connect isolates with at least 95% of average nucleotide identity. Clusters with type strains or at least ten isolates were highlighted, as they represent either known or potentially novel species.



705

706 **Figure 3.** Population structure of *P. putida*. **a.** phylogenetic reconstruction using SNPs
707 extracted from core genome to assign the cgMLST scheme. Colors represent distinct Clonal
708 Complexes and NOCC stands for No Clonal Complex assigned with high confidence. **b.**
709 Principal Component Analysis based on the presence/absence profile of accessory genes
710 present in 5% to 95% of the isolates. **c.** Branch lengths for each Clonal Complex to estimate
711 heterozygosity. **d.** ΔK distribution to estimate the best value for K, which supports the presence
712 of 7 *P. putida* CCs.

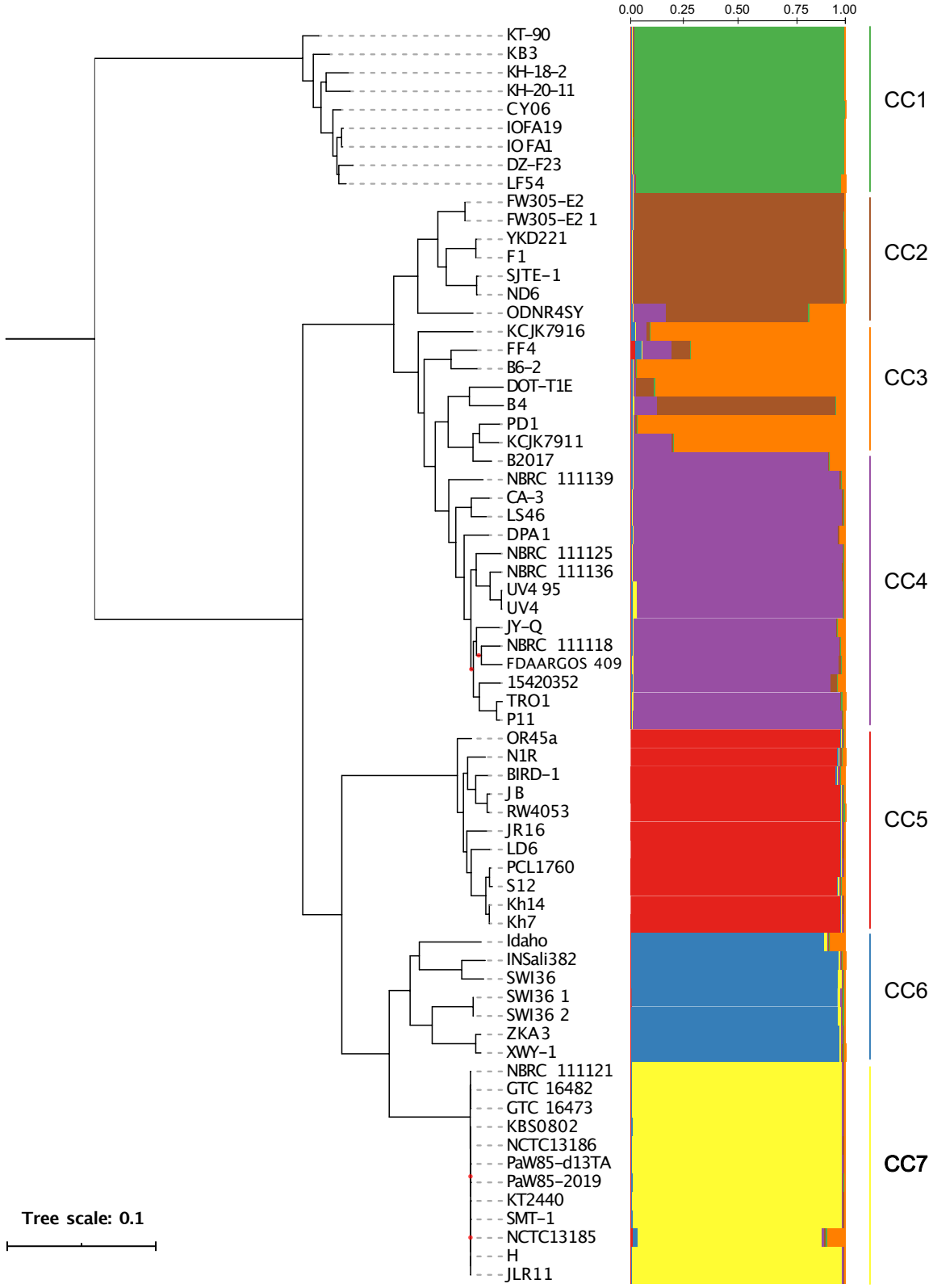


Figure 4. Phylogenetic tree and coancestry barplots, correlating the phylogenetic tree using SNPs extracted from core genome and coancestry probabilities assigned with STRUCTURE. Red dots in branches represent bootstrap values lower than 70%. Each color represents one of the seven predicted Clonal Complexes (CC).

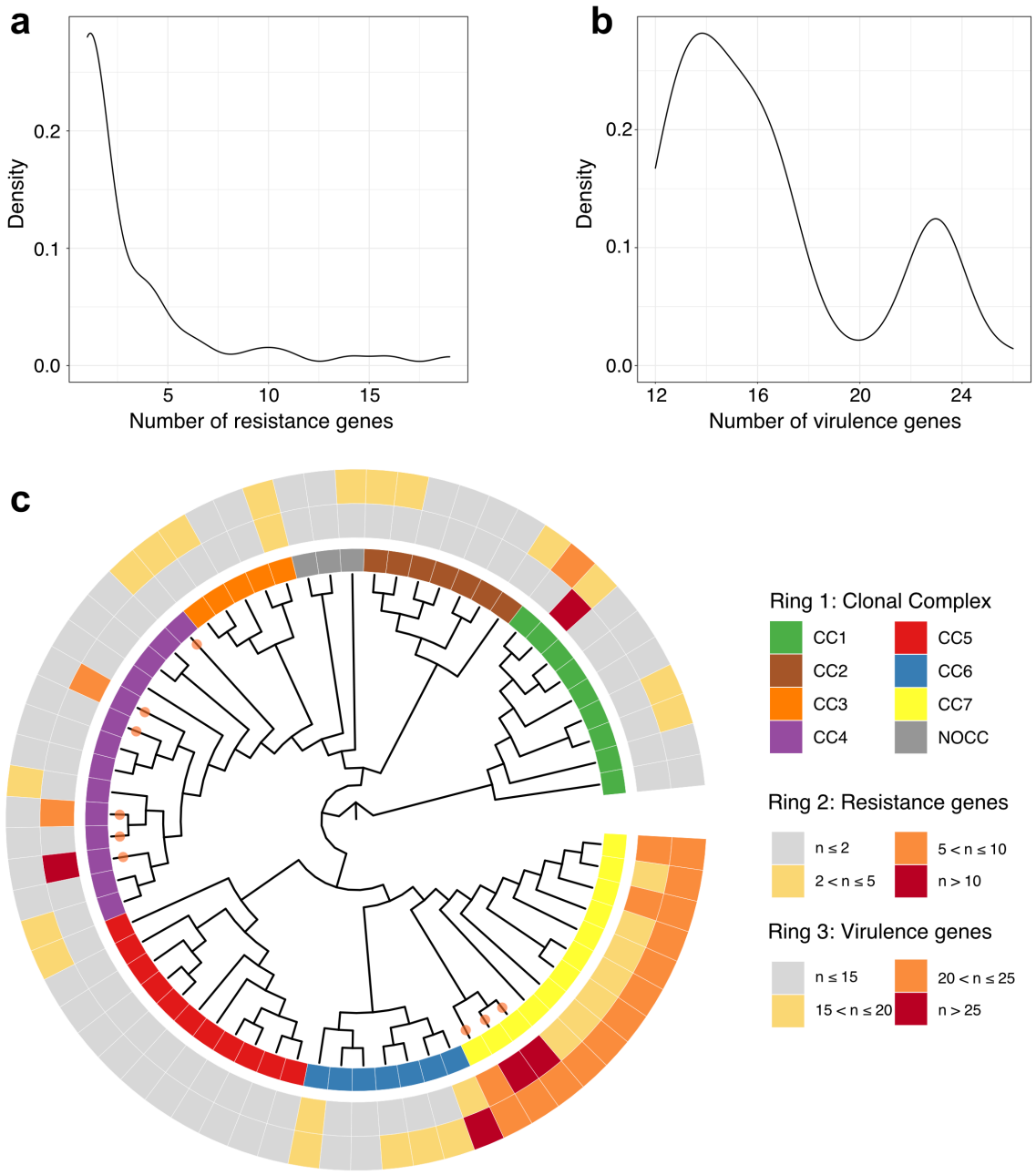


Figure 5. Population structure and distribution of virulence and resistance genes. **a, b.** distribution of acquired resistance and virulence genes. **c.** Maximum likelihood tree from SNPs present in the core genome inferred with 68 genomes used in this study. Branches from clinical isolates are marked with an orange circle. The inner ring represents the clonal complexes (NOCC: "No Clonal Complex"). The second and third rings indicate the number (n) of acquired resistance and virulence genes, respectively.

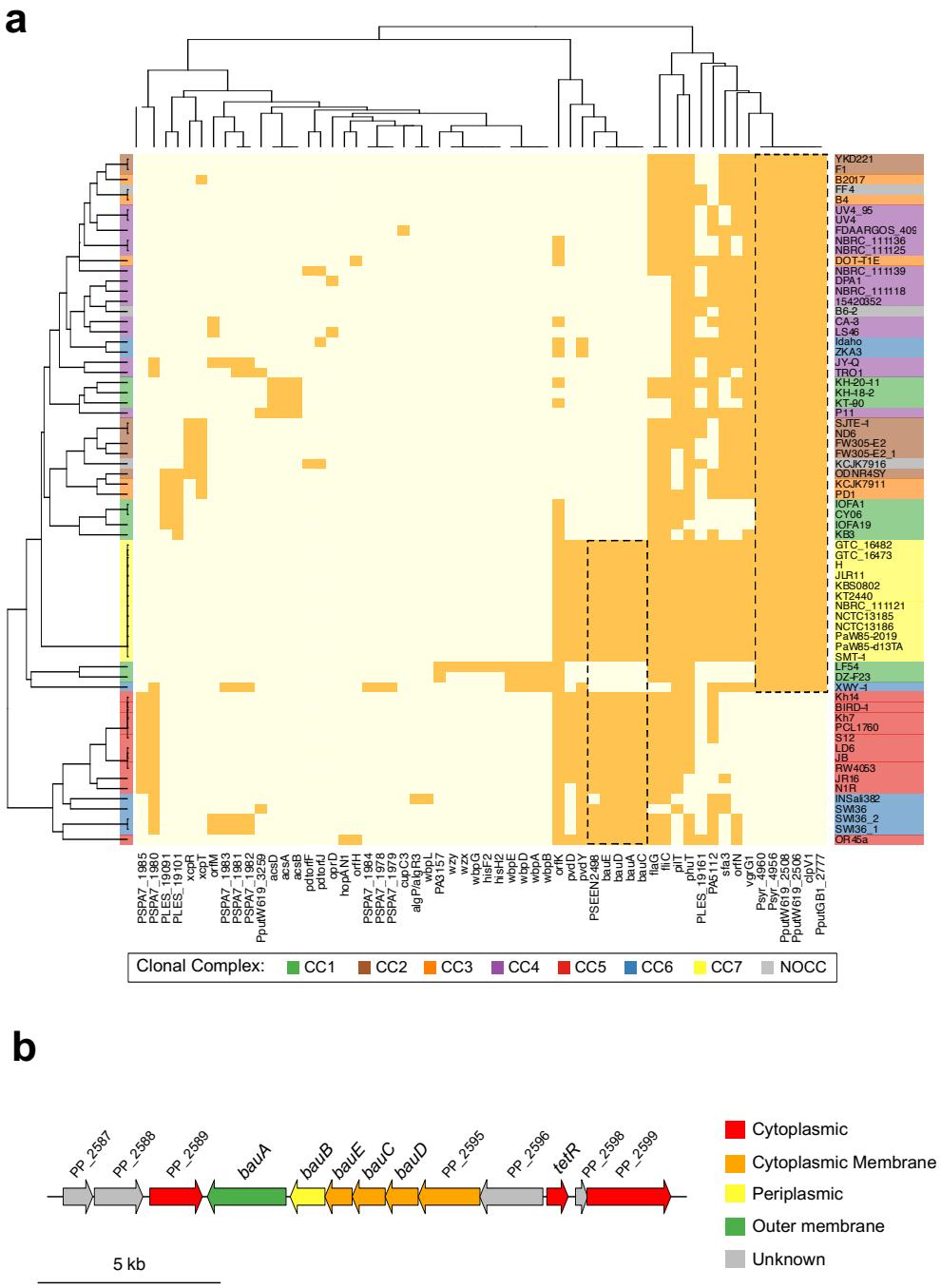


Figure 6. Acquired virulome composition. **a.** Matrix with presence (dark squares) and absence (light squares) profiles of virulence genes. Rows represent strains colored based on Clonal Complex that they belong. Columns are virulence genes identified. **b.** *P. putida* KT2440 *bauABCDE* genetic context annotated according Pseudomonas Genome Database (www.pseudomonas.com) and Song and Kim (2020).