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

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

Members of the *Pseudomonas* genus inhabit a wide variety of environments, which is correlated with their versatile genome and metabolic capacity. *Pseudomonas putida* is a saprophytic soil bacterium, known to bioremediate toxic compounds and to promote plant growth. Nevertheless, some clinical *P. putida* strains have been reported, raising concerns about the safety of this species for biotechnology applications. Despite the ability to occupy different niches, the population structure of *P. putida* remains unclear. Here, we report the population structure of *P. putida* for the first time, which is essential to better investigate its biology. We retrieved 11,025 *Pseudomonas* genomes and used *P. putida* KT2440 as a type strain to define *P. putida sensu stricto*, as well as to estimate its population structure and pangenome. By using network and average nucleotide identity analyses, we found that 90 of the 138 deposited *P. putida* genomes were misclassified, including some reference genomes. Further, we found 23 additional *P. putida* genomes that were not designated as such. *P. putida* has an open pangenome dominated by low-frequency genes. The population structure of *P. putida* is composed by at least 7 clonal complexes (CCs). Clinical isolates are not restricted to a specific CC, although CC4 has 6 out the 9 clinical isolates. Acquired resistance genes are present at low frequency, particularly in plasmids from clinical isolates. Virulence profiles support the higher 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, which is key to assess its safety for environmental and biotechnology applications.

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

INTRODUCTION

The genus *Pseudomonas* is diverse and composed by Gram-negative bacteria[1, 2]. Since the first description of *Pseudomonas*, in 1894, the taxonomy of the genus has evolved in parallel with new methodologies to discriminate species[1]. The taxonomy was inconsistent until the division of *Pseudomonas* into rRNA homology groups[2]. However, because of the low-resolution of 16S rRNA gene for intra-specific classification, other housekeeping genes were used to improve taxonomic resolution[3, 4].

The *Pseudomonas putida* group includes other species such as *P. monteilii*, *P. fulva*, *P. plecoglossicida*, and *P. putida sensu stricto* (*P. alloputida*). *P. putida* belongs to a wide group of fluorescent species from the *Pseudomonas* genus. It is ubiquitous in soil and water and several strains have been isolated from various environments, such as polluted soils and plant roots[5]. *P. putida* is well known to perform many functions such as plant growth promotion, bioremediation, and protection from plant pathogens[6].

Because of its high metabolic diversity, applications of *P. putida* have been proposed for environmental, industrial, and agricultural uses[5]. *P. putida* KT2440 was the first strain of this species to have its genome sequenced, and it is considered a model organism to study plant-microbe interactions[7, 8]. Therefore, KT2440 can be thought as a type strain for *P. putida*. KT2440 was isolated from garden soil in Japan and it is certified as a host-vector system safety level 1[9]. This strain has become a workhorse for industrial biotechnology because of its high metabolic versatility, genetic accessibility, and stress-resistance capability[10]. In addition to the biotechnological application of *P. putida*, some clinical isolates have been described as reservoirs of resistance genes [11, 12], raising concerns about its biosafety.

Although *P. putida* species 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[13]), 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[14]. We computed the pairwise distances between each isolate using mashtree[15] to reconstruct the phylogeny of the genus, which is highly diverse (Figure 1). We mapped each genome deposited in the NCBI 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[16]. 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[16, 17]. 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[13]. Further, these values are likely 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)[18], to oxidize manganese (GB-1)[19], and to damage human tissues (HB3267)[11], are not *P. putida* strains, as they belong to different groups in the network. For example, HB3267 was classified as *P. putida* because of its close phylogenetic relationship with the nicotine degrader S16[20]. HB3267, along with DLL-E4, SF1, and S11, were proposed as members of a new species, *Pseudomonas shirazica*[13]. However, we found that these strains, along with S16, grouped with *Pseudomonas asiatica* type strains. Importantly, both *P. shirazica* and *P. asiatica* were proposed around the same time[13, 21] as attempts to describe the same species. Therefore, the nomenclature of the group containing HB3267 warrants a more careful curation. Keshavarz-Tohid et al. suggested to rename *P. putida sensu stricto* to *P. alloputida* based on the proposed type-strain Kh7[13]. Nevertheless, Kh7 is also a member from *P. putida sensu stricto* group containing the type strain KT2440.

The precise taxonomy of the *P. putida* group still requires a deeper analysis. Although the strain NBRC 14164 has been proposed as a *P. putida* type strain[22], it falls in a group containing 16 other genomes, named in our network as *Pseudomonas spp1*. Moreover, the phylogenetic analysis of the *P. putida* group also revealed the phylogenetic isolation of NBRC 14164 from the other main *P. putida* strains[23]. We considered the group *Pseudomonas spp1* as a potentially novel species and kept KT2440 as the type strain because of its pioneering isolation, description, and genome sequencing[7]. We excluded strains from potentially different species that were classified as *P. putida* (Table S2) and focused on *P. putida sensu stricto* (*P. alloputida*), represented as the red group in the network with 68 isolates (Figure 2).

Pangenome analysis

An effective way to investigate the evolution of a given population is through pangenome analysis. A pangenome is defined as the total set of genes in a given species[24], which is subdivided into core genes, when present in all isolates; accessory genes, when present in at least two (but not in all) isolates or; unique genes, when exclusive to an isolate. The pangenome of *P. putida* has been previously estimated with only 9 isolates[25]. However, among those isolates, GB1, HB3267, S16, and W619 are likely from other species and were removed from our analysis (Table S2). Hence, we re-assessed the pangenome of *P. putida* using 68 isolates with well-supported taxonomic delineation.

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

Regarding the pangenome openness, the slope of the curve (α= 0.417) supports that *P. putida* has an open pangenome[24] (Figure S2). Our estimated α value is much lower than the maximum threshold used to define an open pangenome ($\alpha < 1$), corroborating the previous prediction[25]. Further, this low α value implies that, if more isolates are included in the analysis, new gene families will be detected at a high rate. The high number of gene families in the pangenome of *P. putida* is dominated by unique and low-frequency genes. Only 6,373 genes (24.71%) are found between 5% to 95% of the isolates, while 15,606 genes (60.53%) are found in less than 5% of the isolates, including 10,917 unique genes. The high number of low-frequency genes could be partially associated with fragmented genomes. However, reference genomes such as DOT-T1E showed 267 unique genes, far more than the average of 160.5

157 unique genes found across the dataset. Although the number of low-frequency genes
158 may be overestimated, the high prevalence of unique genes among closely-related
159 genomes indicates a high turnover of unstable genes that might be adaptive under
160 transient selective pressures in their environments.

161 Once determining the exact number of pan and core genomes remains intractable,
162 we also estimated the genomic fluidity (φ) of *P. putida*. The φ estimator is a robust
163 metric that represents the ratio of unique gene families to the sum of gene families in
164 genomes pairs[26]. The smaller the φ, the greater the genes shared by a pair of randomly
165 selected genomes. Analyzing φ instead of the core genome proportion provides a more
166 realistic measure of cohesiveness within a species because low-frequency genes
167 directly affect the pangenome size. *P. putida* has $\varphi = 0.20 \pm 0.04$, indicating that
168 random pairs of *P. putida* genomes have an average 20% and 80% of unique and
169 shared genes, respectively.

170

171 **Population structure**

172 Determining relationships between isolates can provide novel insights into the
173 metabolic diversity of a given species. The Multilocus Sequence Typing (MLST)
174 analysis is a technique based on single-nucleotide polymorphisms (SNPs) within a few
175 housekeeping genes. MLST schemes are available for several species[27]. In an MLST
176 analysis, each combination of SNPs defines a Sequence Type (ST) that can be linked
177 to form Clonal Complexes (CC)[27]. A variation of classical MLST is the core genome
178 MLST (cgMLST), which provides greater resolution by using SNPs from the entire core
179 genome[28]. Here, we used 225,009 SNPs obtained from the *P. putida* core genome to
180 reconstruct the phylogenetic tree and the cgMLST profile.

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

189 We also performed a Discriminant Analysis of Principal Components (DAPC)[29] to
190 recover the genes that contribute most to separate the population based on
191 presence/absence profiles (Figure S4). The top 50 discriminating genes separate CC5
192 and CC7 from the rest of the population, but fail to resolve the relationships between
193 other CCs (Figure S5). Next, we used branch lengths from cgMLST tree as an indirect
194 estimator of diversity within the CCs (Figure 3c). CC7 is the most clonal group,

195 comprising 9 isolates including KT2440 and three clinical isolates (GTC_16482,
196 GTC_16473, and NBRC_111121).

197 The previously published MLST scheme for *P. putida* comprises eight housekeeping
198 genes (*argS, gyrB, ileS, nuoC, ppsA, recA, rpoB,* and *rpoD*)[30]. We assigned isolates
199 to STs and grouped those into CCs (Table S3). This analysis revealed some
200 incongruences with the previous reports[30]. For example, based on allele combinations
201 with perfect-match to predict the STs, KT2440 was detected as ST69 and not as
202 ST58[30], indicating that a revision in the public database is necessary. The predicted
203 number of CCs was also supported by the admixture model from STRUCTURE[31]. This
204 model assumes that each isolate has ancestry from one or more *K* genetically different
205 sources, which we referred to as CCs. The number of CCs corresponds to the
206 estimated number of clusters represented by parameter *K*. Instead of using the highest
207 raw marginal likelihood, we followed the protocol for estimating the best *K* value
208 suggested by Evanno et al[32]. The *ad hoc* statistics ΔK indicates that, according to our
209 dataset, the population structure of *P. putida* is composed by at least 7 CCs (Figure
210 3d).

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

224 The cgMLST phylogenetic tree and the distance tree support CC1 as the basally
225 branching group of *P. putida* (Figure 4). CC1 comprises isolates from deep-sea
226 sediments from Indian Ocean, costal water from the Pacific Ocean, lotus field, and
227 arthropods (Table 1). The main difference of CC1 is the lack of 229 gene families in
228 the accessory genome, which are present in at least one isolate from all other CCs
229 (Table S4). Among these absent gene families, there is a genomic island with nearly
230 46 Kbp and harboring 38 genes. Some of those genes are involved in sugar transport,
231 as previously identified in KT2440 (CC7)[33], as well as hypothetical proteins
232 (coordinates 3,126,465-3,172,496 in KT2440).

233

234 **Resistance profiles**

235 We evaluated the composition of antibiotic resistance genes using the CARD
236 database[34]. All 15 different genes in the core resistome encode MDR efflux pumps
237 (Table 2, Table S5). We found MDR efflux complexes such as MexAB-OprM, MexEF-
238 OprN, and MexJK, from the resistance-nodulation-cell division (RND) efflux pumps
239 family. These efflux pumps are associated with intrinsic and acquired multidrug
240 resistance in *P. aeruginosa*[35, 36]. However, these RND efflux pumps may play an
241 alternative role in *P. putida* by pumping out toxic substrates such as toluene[37]. We also
242 identified *cpxR*, which encodes a protein that promotes MexAB-OprM expression in *P.*
243 *aeruginosa* in the absence of the MexR repressor[38], which is absent in *P. putida*. The
244 presence of MexAB-OprM in the core genome, under CpxR regulation, supports its
245 involvement with intrinsic physiology in addition to drug resistance, because this
246 complex can be involved in both quorum-sensing and host interaction mediation in *P.*
247 *aeruginosa*[39, 40].

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

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

263 We evaluated the presence of acquired genes in plasmids predicted with the PLSDB
264 database (version 2020_06_29)[41]. *P. putida* GTC_16473 contained the genes *aac(6')-*
265 *IIa*, *aadA23*, and bla_{CARB-3} located in a scaffold with high identity with the pJR2 plasmid
266 from *Pasteurella multocida* (NC_004772.1). We also identified bla_{OXA-2}, *aadA22*,
267 *aac(6')-Ia*, *aac(6')-IIc*, *aph(3'')-Ib*, *aph(6')-Id*, bla_{IMP-1}, and *sul1* genes in plasmid-like
268 sequences in *P. putida* GTC_16473, although they are in contigs that have not reached
269 the coverage thresholds to be reliably classified as plasmids, supporting an
270 underestimation of plasmids in *P. putida* isolates. *P. putida* XWY-1 (CC6) also contains
271 a plasmid, pXWY-1 (NZ_CP026333.1), which harbors the resistance genes *sul1*,
272 *aadA2,* and *qacH*. This strain was isolated from rice fields in China, and has

273 environmental importance. Finding non-clinical strains harboring plasmids with
274 resistance genes of this relevance warrants further investigation.

275 Our results also highlight clinical strains harboring a range of resistance genes[11, 42].
276 However, although these strains fall in the *P. putida* group (Figure 1), they do not meet
277 the ANI criteria to be classified as *P. putida sensu stricto (P. alloputida)*. The only
278 clinical strains that were isolated and properly classified as *P. putida* were
279 FDAARGOS_409 and 15420352. In addition, other seven clinical strains included in
280 our work were previously classified as *Pseudomonas sp.* and were hitherto
281 unrecognized as *P. putida sensu stricto* (Table 1). Therefore, while we found clinical
282 strains misclassified as *P. putida*, there are also genuine *P. putida* isolates (e.g.
283 GTC_16473) with multiple resistance genes that could be considered outbreak
284 candidates. In total, 9 out of 68 (13.2%) *P. putida* genomes analyzed here belong to
285 clinical strains.

286

287 **Virulence profiles**

288 We used the VFDB database[43] to assess the *P. putida* virulence profiles. Although
289 strains such as KT2440 have been approved as a host-vector safety strain[9], the
290 genetic proximity with animal (*P. aeruginosa*) and plant (*P. syringae*) pathogens
291 reinforces the need to evaluate the biosafety of *P. putida* species as a whole. The core
292 virulome of *P. putida* contains genes associated with twitching motility, siderophore
293 production (pyoverdine), and alginate biosynthesis (Table S6). Importantly, *P. putida*
294 lacks key virulence genes usually found in *P. aeruginosa*, such as those encoding
295 exotoxin A, alkaline protease, elastase, rhamnolipid biosynthesis components,
296 phospholipase C, plant cell wall-degrading enzymes, and type III secretion system.
297 Although we have detected genes associated with the type III secretion system in both
298 core (HopAJ2 and HopJ1) and acquired resistome (HopAN1), these genes were
299 experimentally ruled out[44].

300 Most of the alginate regulatory and biosynthesis systems are present in the core
301 virulome of *P. putida* (Table S7). The overproduction of alginate plays an important
302 role in inducing the mucoid morphotype in *P. aeruginosa* during chronic infection[45].
303 However, the transcriptional regulatory gene *algM* (*mucC*), associated with genotypic
304 switching, is absent in *P. putida*. Since the lack of AlgM leads to the non-mucoid
305 phenotype in *P. aeruginosa*[7], it can account for the non-mucoid phenotype in *P. putida*
306 under no water stress conditions. Although the genotypic switching unlikely occurs in
307 *P. putida*, the alginate production is important under water stress conditions by
308 promoting biofilm development and protecting from desiccation[46].

309 The acquired virulome comprised genes for type II and VI secretions systems,
310 adherence, and iron uptake (Table S6). While most isolates had a low frequency of

311 resistance genes, the virulence factors displayed a bimodal distribution (Figure 5b), a
312 pattern that has been previously observed for *Klebsiella aerogenes*[47]. Genes from
313 acinetobactin gene cluster and HSI-I type VI secretion system were differentially
314 distributed across CCs (Figure S7a), although it remains unclear whether these
315 patterns emerged mainly from gene gain or loss. *Acinetobacter baumannii* has
316 acinetobactin as the major siderophore associated with iron uptake[48], which is
317 synthesized through the *bauABCDE* operon. We found this operon in *P. putida* strains
318 within CC5, CC6, and CC7 (Figure S7a). Further, this operon was surrounded by
319 genes coding for proteins from chemotaxis sensory transducer (PP_2599) and
320 aminotransferases (PP_2588) families (Figure S7b). Siderophore-mediated iron
321 acquisition has been explored in *P. putida* KT2440[49] (CC7), but the role played by
322 *bauABCDE* in this species warrants further experimental investigations.

323 Present in all other strains from other CCs, the type VI secretion system (T6SS)
324 HSI-I is absent in both CC5 and CC6 (except XWY-1) isolates (Figure S7a). In KT2440
325 (CC7), HSI-I is a potent weapon against other bacteria such as phytopathogens,
326 increasing the competitiveness of *P. putida*[50]. The absence of T6SS virulence genes
327 was observed in BIRD-1 (CC5)[50], and our work generalizes this observation for all CC5
328 members. In KT2440, T6SS is crucial to kill phytopathogens such as *Xanthomonas*
329 *campestris*[51]. These results indicate that BIRD-1 and other members from CC5 are
330 likely less efficient than KT2440 as biocontrol agents. Moreover, all clinical *P. putida*
331 strains harbor T6SS, indicating their potential ability to outcompete other bacteria
332 during infections.

333

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

335 The ability of *P. putida* to promote plant-growth and bioremediate toxic compounds
336 have been explored[7, 10, 52, 53]. We searched for genes involved in plant growth
337 promotion and bioremediation through a literature search of genes that have already
338 been described for Pseudomonas species (Table S7). We found genes in the core
339 genome such as the pyrroloquinoline quinone-encoding operon *pqqBCDEFG,*
340 associated with mineral phosphate solubilization in *Serratia marcescens*[54] and
341 *Pseudomonas fluorescens*[55] (Table S8). Mineral solubilization has already been
342 reported experimentally for BIRD-1[52] and KT2440[55], indicating that all *P. putida* isolates
343 are genetically equipped to solubilize inorganic phosphate.

344 Another key feature that can enhance plant growth is the colonization of seeds by
345 *P. putida*[56]. In KT2440, genes associated with surface adhesion (e.g. *lapA*, *lapBCD*),
346 flagellum biosynthesis (e.g. *flhB*, *fliF*, *fliD*, *fliC*), and virulence regulation (e.g. *rpoN* and
347 *gacS*) have been experimentally shown to be important for attachment to corn seeds[56].
348 All these genes, except *lapA* and *fliC*, belongs to the *P. putida* core genome (Table
349 S8). Although often described as virulence genes, flagellum genes also play roles in

350 the association of *P. putida* with plants. Other plant growth-promoting genes were also
351 found in the *P. putida* accessory genome (Table S8), except for CC1, which lacks all
352 putative plant growth-promoting genes.

353 Many plant growth-promoting bacteria can synthesize phytohormones such as
354 indole-3-acetic acid. *P. putida* lacks the main genes involved in this process (e.g. *ipdC*,
355 *iaaM*, and *iaaH*), indicating an incomplete or nonfunctional pathway. Although *P. putida*
356 W619 (GCA_000019445.1) is considered one of the most efficient producers of IAA in
357 comparison with other endophytic bacteria[18], this strain was excluded from our analysis
358 (Table S2) because it is likely misclassified as *P. putida*. In addition, *P. putida* also
359 lacks AcdS, an enzyme that counteract the plant's ethylene stress response, a result
360 that has been experimentally confirmed in KT2440[57].

361 Besides the ability to promote plant-growth, *P. putida* can tolerate or degrade an
362 array of compounds including heavy metals and hydrocarbons. We identified genes in
363 the core genome that allow *P. putida* to resist various heavy metals such as copper
364 (*cop* genes) and cobalt/zinc/cadmium (*czcABC*) (Table S9). The copper/silver
365 resistance operon *cusABC* was present in the accessory genome. Although the wide
366 range of genes associated with bioremediation, we did not find a clear correlation
367 between population structure and gene presence/absence profile.

368 *P. putida* is known for its capacity to metabolize aromatic hydrocarbons such as
369 toluene, benzene, and *p*-cymene[7, 33]. The toluene-degrading pathway includes the
370 *todABCDE* operon and the *todST* regulator. The *p*-cymene compound can be
371 degraded by means of the *cymAaAbBCDER* or the *cmtAaAbAcAdBCDEFGHI*
372 operon[58]. We found a genomic island of approximately 48kb long harboring all these
373 genes in F1, DOT-T1E, UV4, UV4/95, YKD221, and NBRC_111125 genomes (Figure
374 S9). All these isolates were experimentally confirmed to degrade toluene (except
375 NBRC_111125). *P. putida* F1 is best known to grow on toluene[33]; YKD221 was isolated
376 from contaminated industrial soil and degrades *cis*-dichloroethene[59]; DOT-T1E is an
377 isolate known to grow on different carbon sources[60] and; UV4 and UV4/95 conduct
378 important industrial biotransformation of arenes, alkenes, and phenols[61]. Interestingly,
379 genes in this genomic island presented a very similar genetic context (Figure S8), with
380 an upstream arm-type integrase associated with bacteriophages. We were unable to
381 precisely define the *att* viral sites, indicating a deterioration of the original structure of
382 the putative bacteriophage. The lack of correspondence between population structure
383 and the presence of an integrase upstream the genomic island indicates that this
384 region was likely acquired via independent horizontal gene transfers in distinct *P.*
385 *putida* CCs.

386 We also identified RND efflux pumps involved with solvent tolerance in both core
387 (TtgABC) and accessory genomes (TtgDEF and TtgGHI). TtgABC, TtgDEF, and
388 TtgGHI are required for DOT-T1E to efficiently tolerate toluene[37]. We observed that
389 TtgABC is the same protein-complex predicted as MexAB-OprM, associated with

390 antibiotic resistance in the core genome. This complex extrudes both antibiotics and
391 solvents such as toluene in *P. putida* DOT-T1E[37], corroborating the additional and
392 important function to extrude antibiotics and organic solvents in all *P. putida* isolates.
393 TtgDEF is located in the genomic island co-occurring with *tod* genes. This complex
394 can expel toluene, but not antibiotics[62], reinforcing the broad different molecules that
395 can be used as substrates for RND efflux pumps, and the need to explore what
396 determines this specificity in *P. putida* isolates and other bacteria.

397

398 **CONCLUDING REMARKS**

399 Through a remarkable metabolic versatility, *P. putida* can thrive in a wide variety of
400 niches. In this work, we explored the genetic diversity of *P. putida sensu stricto* (*P.*
401 *alloputida*) and characterized its population structure for the first time. Through a large-
402 scale genomic analysis, we identified a major problem with *P. putida* classification,
403 including with several reference strains that likely belong to new species, as suggested
404 elsewhere[13]. *P. putida* has an open pangenome dominated by low-frequency genes.
405 The population structure of this species has at least 7 clonal complexes that were
406 verified by cgMLST and STRUCTURE ancestry simulations. We found no association
407 of clinical isolates with a specific CC. Together, our results indicate that *P. putida*
408 clinical isolates are mainly opportunistic and do not pose considerable health concerns.

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

420

421 **MATERIALS AND METHODS**

422 **Dataset and genomic features**

423 We recovered 11,025 genomes from the *Pseudomonas* genus in June 2020. To
424 assess the quality of the genomes, we used BUSCO v4.0.6[14] with a threshold of 90%
425 completeness. The KT2440 (GCA_900167985.1) was used as a reference with mash
426 v.2.2.2[63] to obtain genomes with a distance lower than 0.05. We used mashtree[15] to

427 generate the distance tree. The ANI analysis was performed with pyani 0.2.10[64].
428 Network analysis was conducted in R with the igraph package (https://igraph.org). We
429 removed both S12 (GCA_000287915.1) and KT2440 (GCA_000007565.2) because
430 they were duplicated genomes. Gene prediction in all isolates was conducted with
431 prokka v1.12[65] to avoid bias in the identification of protein families. Plasmids were
432 analyzed PLSD v2020_06_29[41].

433 **Pangenome characterization**

434 We inferred the *P. putida* pangenome using Roary 3.13.0[66], with a minimum
435 threshold of 85% identity to cluster proteins. Core genes were defined as those present
436 in more than 95% of the isolates. Jaccard distances were computed by using
437 accessory genes with prevalence between 5% and 95%. Multivariate analysis was
438 performed to asses gene content variations between ecotypes from *P. putida* through
439 DAPC analysis using the *ade4* and *adegenet* packages[29], retaining the 30 principal
440 components and 3 discriminant functions. Pangenome openness and fluidity were
441 conducted with micropan[67] with 500 and 1000 permutations, respectively.

442 **Population structure analysis**

443 We used *in-house* scripts to extract the genes present in all isolates, which were
444 aligned with MAFFT v7.467[68]. SNPs were retrieved with snp-sites v2.3.3[69] and SNP
445 alignment was used as input to RAxML v8[70] to reconstruct the phylogenetic tree using
446 the general time-reversible model and gamma correction. Since we used only variable
447 sites as input, we used ASC_GTRGAMMA to correct ascertainment bias with the Paul
448 Lewis correction. One thousand bootstrap replicates were generated to assess the
449 significance of internal nodes. We inferred the cgMLST scheme using the core genome
450 SNP phylogenetic tree. The phylogenetic tree was visualized with iTOL v4[71].

451 We downloaded the *P. putida* MLST scheme (on June, 2020) containing 116
452 different STs (https://pubmlst.org/databases/). This scheme was designed for the
453 whole *P. putida* group, not only *P. putida* sensu stricto[30]. We used BLASTN[72] to
454 determine the best-matching MLST allele to access STs. The allelic profile associated
455 with each ST in our dataset was used to conduct population assignment with
456 STRUCTURE v2.3.4[31] with admixture model. The length of Markov chain Monte Carlo
457 (MCMC) was 50,000, discarding 20,000 iterations as burn-in. The simulations to
458 calculate the parameter K ranged from 2 to 20, with 20 replicates for each K to estimate
459 confidence intervals. Instead of using raw posterior probability to get the best K, we
460 followed the protocol suggested by Evanno et al[32]. Briefly, we calculated the first and
461 second derivatives to estimate a ΔK of 7. Therefore, we used K = 7 to analyze predicted
462 ancestry probabilities.

463

464 **Detection of genes associated with resistance, virulence, plant-growth**
465 **promotion, and bioremediation**

466 We used the Comprehensive Antimicrobial Resistance Database (CARD) database
467 v3.0.9[34] to predict antibiotic resistance genes. The virulence factor database (VFDB)[43]
468 was used to determine virulence genes. This database was downloaded on July 30
469 2020 and comprises 28,639 proteins associated with virulence in several pathogens.
470 We used virulence genes previously described for Pseudomonas genus. We clustered
471 proteins based on 70% identity to retrieve a non-redundant database using uclust
472 v1.2.22q[73]. We built the database with plant-growth promotion and bioremediation
473 through literature searches (Table S7). All predicted proteins were globally aligned
474 against these databases using usearch v11.0.667[73] with 50% coverage for both query
475 and subject and minimum 60% identity.

476

477 **ACKNOWLEDGEMENTS**

478 This work was supported by Fundação Carlos Chagas Filho de Amparo à Pesquisa do
479 Estado do Rio de Janeiro (FAPERJ; grants E-26/203.309/2016 and E-
480 26/203.014/2018), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior -
481 Brasil (CAPES; Finance Code 001), and Conselho Nacional de Desenvolvimento
482 Científico e Tecnológico. The funding agencies had no role in the design of the study
483 and collection, analysis, and interpretation of data and in writing.

484

485 **REFERENCES**

486 1. Palleroni NJ. The Pseudomonas story. *Environ Microbiol* **12**, 1377-1383 (2010).
487
488 2. Stanier RY, Palleroni NJ, Doudoroff M. The aerobic pseudomonads: a taxonomic study. *J Gen*
489 *Microbiol* **43**, 159-271 (1966).
490
491 3. Ait Tayeb L, Ageron E, Grimont F, Grimont PA. Molecular phylogeny of the genus Pseudomonas
492 based on rpoB sequences and application for the identification of isolates. *Res Microbiol* **156**,
493 763-773 (2005).
494
495 4. Gomila M, Pena A, Mulet M, Lalucat J, Garcia-Valdes E. Phylogenomics and systematics in
496 Pseudomonas. *Front Microbiol* **6**, 214 (2015).
497
498 5. Volke DC, Calero P, Nikel PI. Pseudomonas putida. *Trends Microbiol* **28**, 512-513 (2020).
499
500 6. Timmis KN. Pseudomonas putida: a cosmopolitan opportunist par excellence. *Environ*
501 *Microbiol* **4**, 779-781 (2002).
502

503 7. Nelson KE*, et al.* Complete genome sequence and comparative analysis of the metabolically
504 versatile Pseudomonas putida KT2440. *Environ Microbiol* **4**, 799-808 (2002).
505
506 8. Belda E*, et al.* The revisited genome of Pseudomonas putida KT2440 enlightens its value as a
507 robust metabolic chassis. *Environ Microbiol* **18**, 3403-3424 (2016).
508
509 9. Kampers LFC, Volkers RJM, Martins Dos Santos VAP. Pseudomonas putida KT2440 is HV1
510 certified, not GRAS. *Microb Biotechnol* **12**, 845-848 (2019).
511
512 10. Nikel PI, de Lorenzo V. Pseudomonas putida as a functional chassis for industrial biocatalysis:
513 From native biochemistry to trans-metabolism. *Metab Eng* **50**, 142-155 (2018).
514
515 11. Fernandez M*, et al.* Analysis of the pathogenic potential of nosocomial Pseudomonas putida
516 strains. *Front Microbiol* **6**, 871 (2015).
517
518 12. Molina L*, et al.* Specific Gene Loci of Clinical Pseudomonas putida Isolates. *PLoS One* **11**,
519 e0147478 (2016).
520
521 13. Keshavarz-Tohid V*, et al.* Genomic, phylogenetic and catabolic re-assessment of the
522 Pseudomonas putida clade supports the delineation of Pseudomonas alloputida sp. nov.,
523 Pseudomonas inefficax sp. nov., Pseudomonas persica sp. nov., and Pseudomonas shirazica
524 sp. nov. *Syst Appl Microbiol* **42**, 468-480 (2019).
525
526 14. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation
527 Completeness. *Methods Mol Biol* **1962**, 227-245 (2019).
528
529 15. Katz LS*, et al.* Mashtree: a rapid comparison of whole genome sequence files. *Journal of Open*
530 *Source Software* **4**, (2019).
531
532 16. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species
533 definition. *Proc Natl Acad Sci U S A* **106**, 19126-19131 (2009).
534
535 17. Bobay LM. The Prokaryotic Species Concept and Challenges. In: *The Pangenome: Diversity,*
536 *Dynamics and Evolution of Genomes* (eds Tettelin H, Medini D) (2020).
537
538 18. Taghavi S*, et al.* Genome survey and characterization of endophytic bacteria exhibiting a
539 beneficial effect on growth and development of poplar trees. *Appl Environ Microbiol* **75**, 748-
540 757 (2009).
541
542 19. Rosson RA, Nealson KH. Manganese binding and oxidation by spores of a marine bacillus. *J*
543 *Bacteriol* **151**, 1027-1034 (1982).
544
545 20. Molina L*, et al.* Antibiotic resistance determinants in a Pseudomonas putida strain isolated
546 from a hospital. *PLoS One* **9**, e81604 (2014).
547
548 21. Tohya M*, et al.* Pseudomonas asiatica sp. nov., isolated from hospitalized patients in Japan and
549 Myanmar. *Int J Syst Evol Microbiol* **69**, 1361-1368 (2019).
550
551 22. Ohji S, Yamazoe A, Hosoyama A, Tsuchikane K, Ezaki T, Fujita N. The Complete Genome
552 Sequence of Pseudomonas putida NBRC 14164T Confirms High Intraspecies Variation. *Genome*
553 *Announc* **2**, (2014).
554

23. Yonezuka K*, et al.* Phylogenetic analysis reveals the taxonomically diverse distribution of the Pseudomonas putida group. *J Gen Appl Microbiol* **63**, 1-10 (2017).

24. Tettelin H*, et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* **102**, 13950-13955 (2005).

25. Udaondo Z, Molina L, Segura A, Duque E, Ramos JL. Analysis of the core genome and pangenome of Pseudomonas putida. *Environ Microbiol* **18**, 3268-3283 (2016).

26. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* **12**, 32 (2011).

27. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* **3**, 124 (2018).

28. de Been M*, et al.* Core Genome Multilocus Sequence Typing Scheme for High- Resolution Typing of Enterococcus faecium. *J Clin Microbiol* **53**, 3788-3797 (2015).

29. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet* **11**, 94 (2010).

30. Ogura K, Shimada K, Miyoshi-Akiyama T. A multilocus sequence typing scheme of Pseudomonas putida for clinical and environmental isolates. *Sci Rep* **9**, 13980 (2019).

31. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000).

32. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* **14**, 2611-2620 (2005).

33. Wu X, Monchy S, Taghavi S, Zhu W, Ramos J, van der Lelie D. Comparative genomics and functional analysis of niche-specific adaptation in Pseudomonas putida. *FEMS Microbiol Rev* **35**, 299-323 (2011).

34. Alcock BP*, et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* **48**, D517-D525 (2020).

35. Lister PD, Wolter DJ, Hanson ND. Antibacterial-resistant Pseudomonas aeruginosa: clinical impact and complex regulation of chromosomally encoded resistance mechanisms. *Clin Microbiol Rev* **22**, 582-610 (2009).

36. Schweizer HP. Efflux as a mechanism of resistance to antimicrobials in Pseudomonas aeruginosa and related bacteria: unanswered questions. *Genet Mol Res* **2**, 48-62 (2003).

37. Rojas A*, et al.* Three efflux pumps are required to provide efficient tolerance to toluene in Pseudomonas putida DOT-T1E. *J Bacteriol* **183**, 3967-3973 (2001).

38. Tian ZX, Yi XX, Cho A, O'Gara F, Wang YP. CpxR Activates MexAB-OprM Efflux Pump Expression and Enhances Antibiotic Resistance in Both Laboratory and Clinical nalB-Type Isolates of Pseudomonas aeruginosa. *PLoS Pathog* **12**, e1005932 (2016).

39. Evans K, Passador L, Srikumar R, Tsang E, Nezezon J, Poole K. Influence of the MexAB-OprM multidrug efflux system on quorum sensing in Pseudomonas aeruginosa. *J Bacteriol* **180**, 5443-5447 (1998).

40. Sanchez P*, et al.* Fitness of in vitro selected Pseudomonas aeruginosa nalB and nfxB multidrug resistant mutants. *J Antimicrob Chemother* **50**, 657-664 (2002).

41. Galata V, Fehlmann T, Backes C, Keller A. PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res* **47**, D195-D202 (2019).

42. Mustapha MM*, et al.* Draft Genome Sequences of Four Hospital-Associated Pseudomonas putida Isolates. *Genome Announc* **4**, (2016).

43. Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res* **47**, D687-D692 (2019).

44. Macho AP, Ruiz-Albert J, Tornero P, Beuzon CR. Identification of new type III effectors and analysis of the plant response by competitive index. *Mol Plant Pathol* **10**, 69-80 (2009).

45. Henry RL, Mellis CM, Petrovic L. Mucoid Pseudomonas aeruginosa is a marker of poor survival in cystic fibrosis. *Pediatr Pulmonol* **12**, 158-161 (1992).

46. Chang WS, van de Mortel M, Nielsen L, Nino de Guzman G, Li X, Halverson LJ. Alginate production by Pseudomonas putida creates a hydrated microenvironment and contributes to biofilm architecture and stress tolerance under water-limiting conditions. *J Bacteriol* **189**, 8290-8299 (2007).

47. Passarelli-Araujo H, Palmeiro JK, Moharana KC, Pedrosa-Silva F, Dalla-Costa LM, Venancio TM. Genomic analysis unveils important aspects of population structure, virulence, and antimicrobial resistance in Klebsiella aerogenes. *FEBS J* **286**, 3797-3810 (2019).

48. Song WY, Kim HJ. Current biochemical understanding regarding the metabolism of acinetobactin, the major siderophore of the human pathogen Acinetobacter baumannii, and outlook for discovery of novel anti-infectious agents based thereon. *Nat Prod Rep* **37**, 477-487 (2020).

49. Matthijs S*, et al.* Siderophore-mediated iron acquisition in the entomopathogenic bacterium Pseudomonas entomophila L48 and its close relative Pseudomonas putida KT2440. *Biometals* **22**, 951-964 (2009).

50. Bernal P, Allsopp LP, Filloux A, Llamas MA. The Pseudomonas putida T6SS is a plant warden against phytopathogens. *ISME J* **11**, 972-987 (2017).

51. Validov S*, et al.* Selection of bacteria able to control Fusarium oxysporum f. sp. radicis-lycopersici in stonewool substrate. *J Appl Microbiol* **102**, 461-471 (2007).

52. Roca A*, et al.* Analysis of the plant growth-promoting properties encoded by the genome of the rhizobacterium Pseudomonas putida BIRD-1. *Environ Microbiol* **15**, 780-794 (2013).

53. Gong T*, et al.* Engineering Pseudomonas putida KT2440 for simultaneous degradation of carbofuran and chlorpyrifos. *Microb Biotechnol* **9**, 792-800 (2016).

54. Matteoli FP*, et al.* Genome sequencing and assessment of plant growth-promoting properties of a Serratia marcescens strain isolated from vermicompost. *BMC Genomics* **19**, 750 (2018).

55. Miller SH, Browne P, Prigent-Combaret C, Combes-Meynet E, Morrissey JP, O'Gara F. Biochemical and genomic comparison of inorganic phosphate solubilization in Pseudomonas species. *Environ Microbiol Rep* **2**, 403-411 (2010).

56. Duque E*, et al.* Identification of reciprocal adhesion genes in pathogenic and non-pathogenic Pseudomonas. *Environ Microbiol* **15**, 36-48 (2013).

57. Glick BR, Todorovic B, Czarny J, Cheng Z, Duan J, McConkey B. Promotion of Plant Growth by Bacterial ACC Deaminase. *Critical Reviews in Plant Sciences* **26**, (2007).

58. Eaton RW. p-Cymene catabolic pathway in Pseudomonas putida F1: cloning and characterization of DNA encoding conversion of p-cymene to p-cumate. *J Bacteriol* **179**, 3171-3180 (1997).

59. Yonezuka K*, et al.* Isolation and characterization of a bacterial strain that degrades cis-dichloroethenein the absence of aromatic inducers. *J Gen Appl Microbiol* **62**, 118-125 (2016).

60. Daniels C*, et al.* Global regulation of food supply by Pseudomonas putida DOT-T1E. *J Bacteriol* **192**, 2169-2181 (2010).

61. Skvortsov T, Hoering P, Arkhipova K, Whitehead RC, Boyd DR, Allen CCR. Draft Genome Sequences of Pseudomonas putida UV4 and UV4/95, Toluene Dioxygenase-Expressing Producers of cis-1,2-Dihydrodiols. *Genome Announc* **6**, (2018).

62. Mosqueda G, Ramos JL. A set of genes encoding a second toluene efflux system in Pseudomonas putida DOT-T1E is linked to the tod genes for toluene metabolism. *J Bacteriol* **182**, 937-943 (2000).

63. Ondov BD*, et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132 (2016).

64. Pritchard L, Glover HR, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods* **8**, 10-24 (2016).

65. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).

66. Page AJ*, et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3693 (2015).

67. Snipen L, Liland KH. micropan: an R-package for microbial pan-genomics. *BMC Bioinformatics* **16**, 79 (2015).

68. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013).

69. Page AJ*, et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* **2**, e000056 (2016).

711
712 70. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
713 phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
714
715 71. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments.
716 *Nucleic Acids Res* **47**, W256-W259 (2019).
717
718 72. Altschul SF*, et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search
719 programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
720
721 73. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**,
722 2460-2461 (2010).
723

724 **Tables**

725 **Table 1.** *Pseudomonas putida* isolates used in this study with average nucleotide
726 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

727

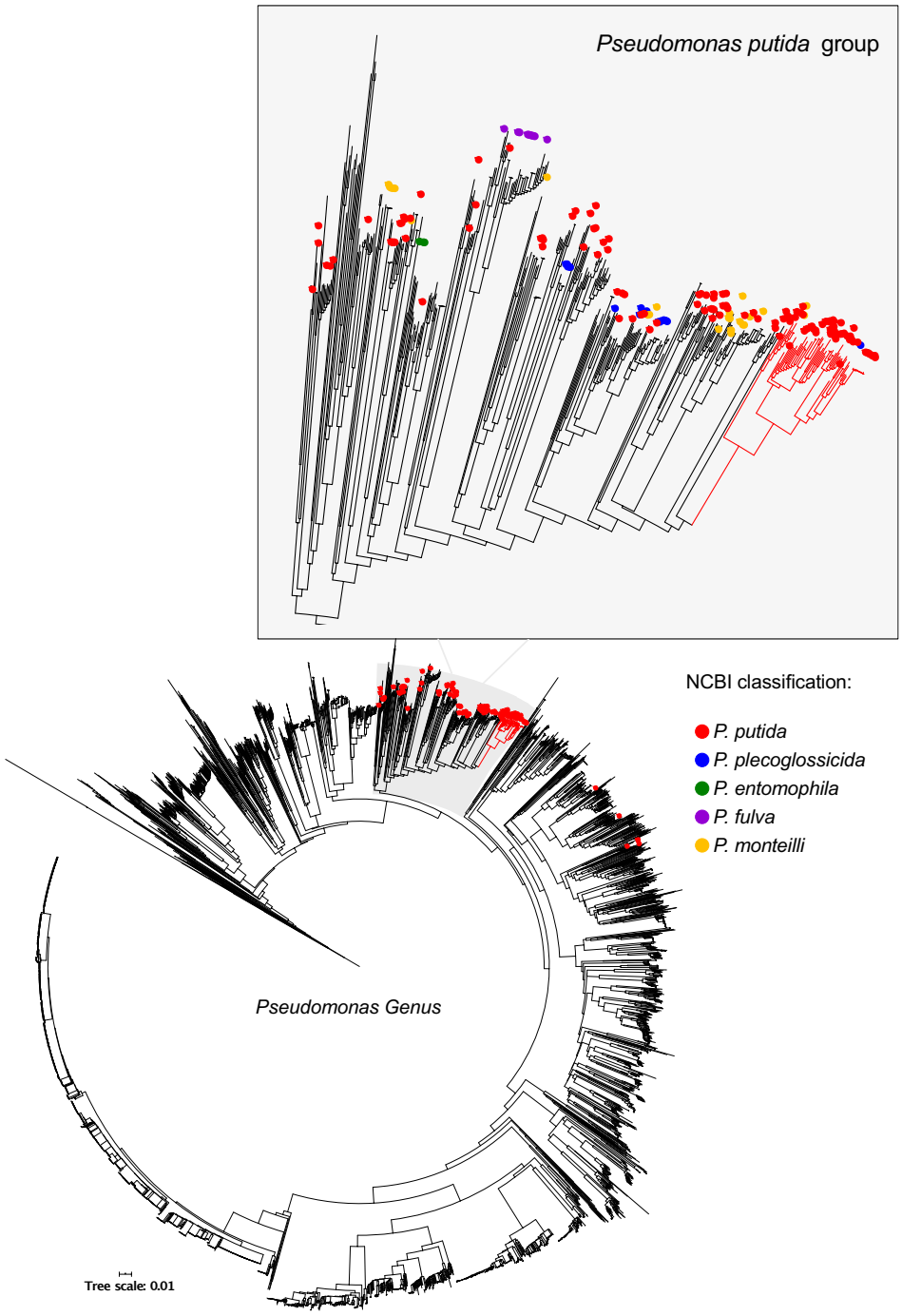
728

729 Table 2. Frequency of resistance mechanisms categories in both core and accessory
730 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%)

731 *Number of different genes.

732 **Figures**



733

734 **Figure 1.** Phylogenetic reconstruction of 10,457 *Pseudomonas* genomes from
735 *Pseudomonas* genus. Genomes classified as *P. putida* according to NCBI are marked
736 as red circle. *P. putida* group was highlighted to assess the distribution of misclassified
737 genomes. *P. putida* (*P. alloputida*) branches are colored in red.

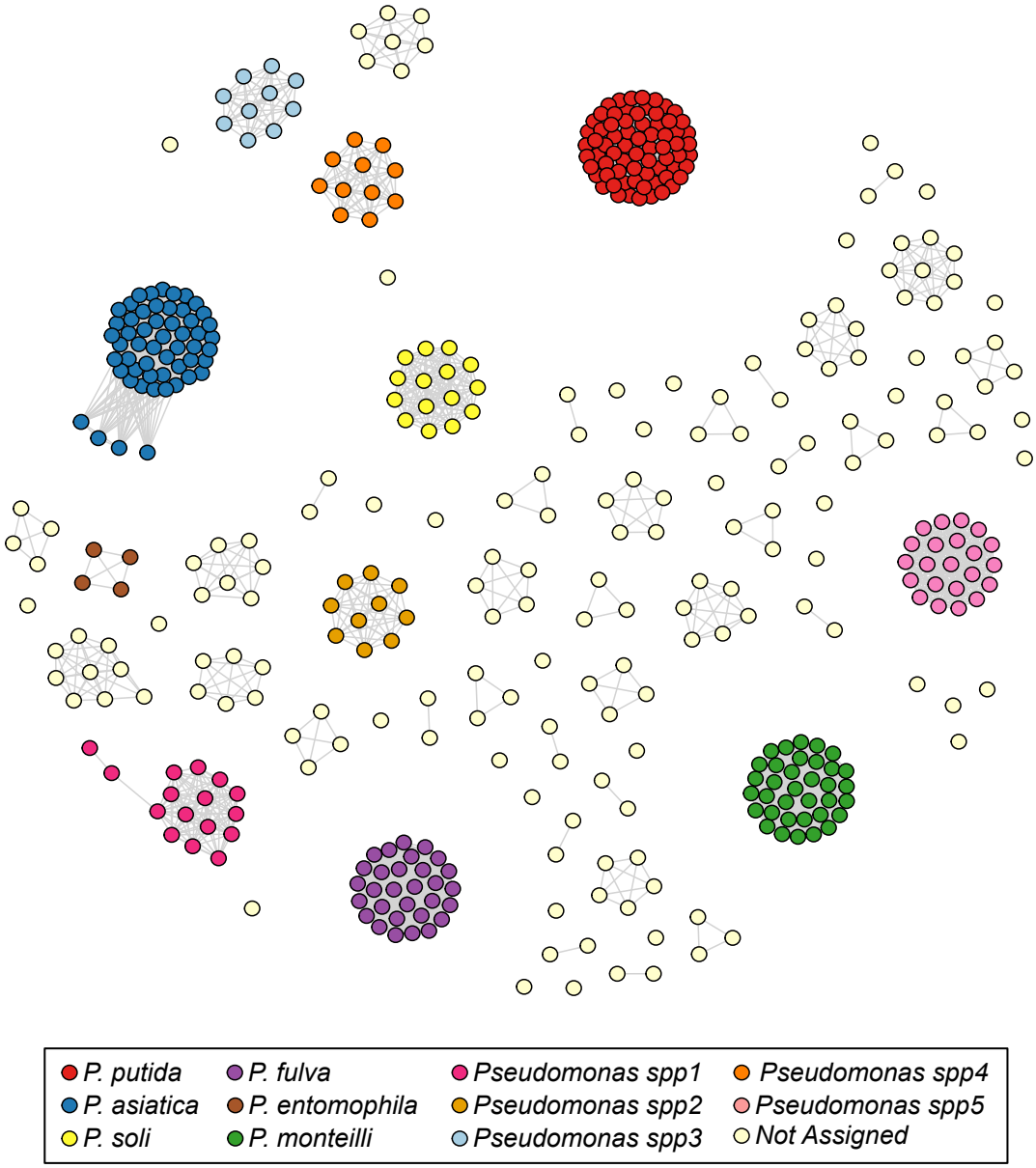
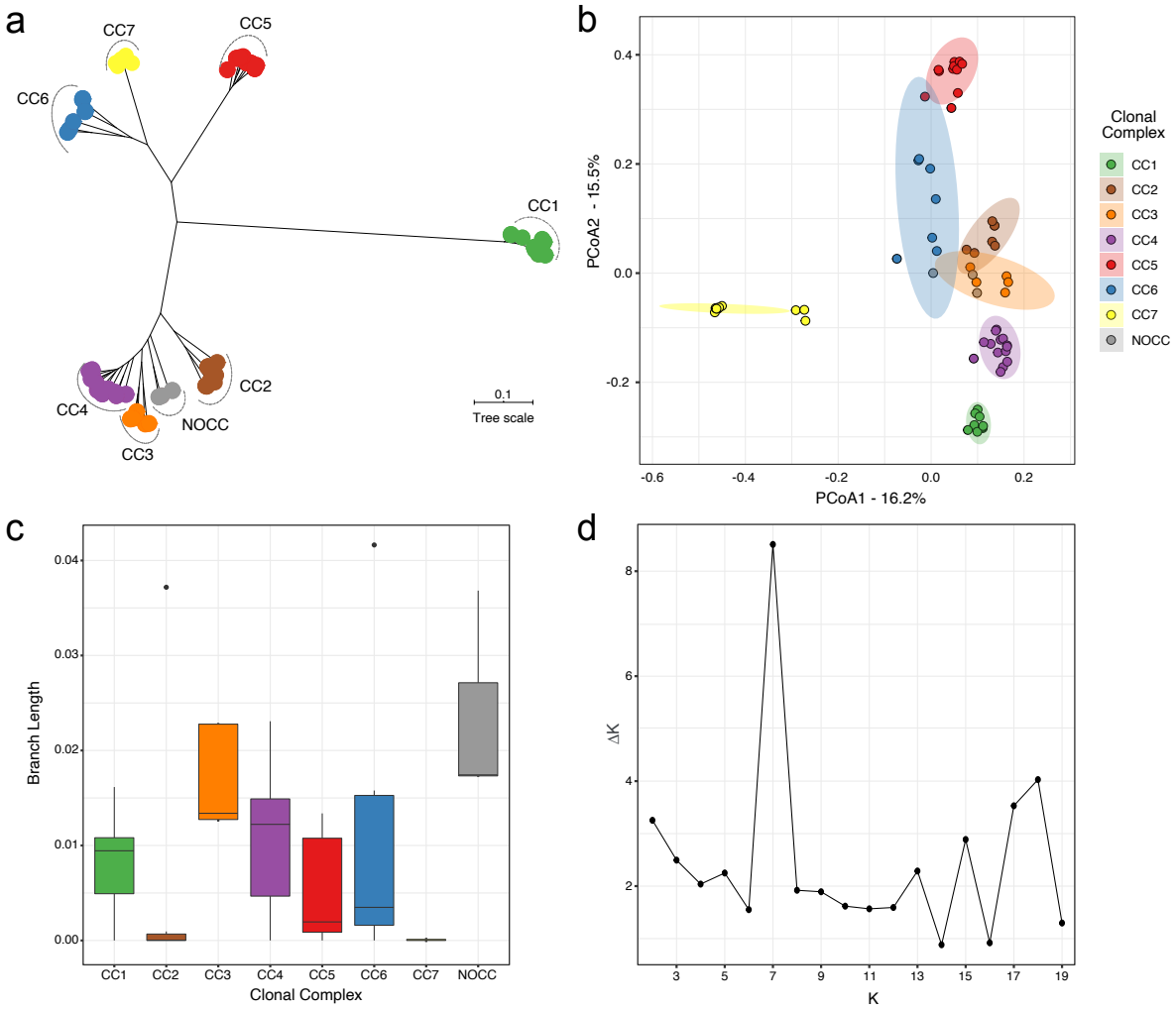


Figure 2. Network analysis of isolates from the *P. putida* group. Nodes represent isolates and edges connect isolates with at least 95% of average nucleotide identity. Clusters with type strains or at least ten isolates were highlighted, as they represent either known or potentially novel species.



743

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

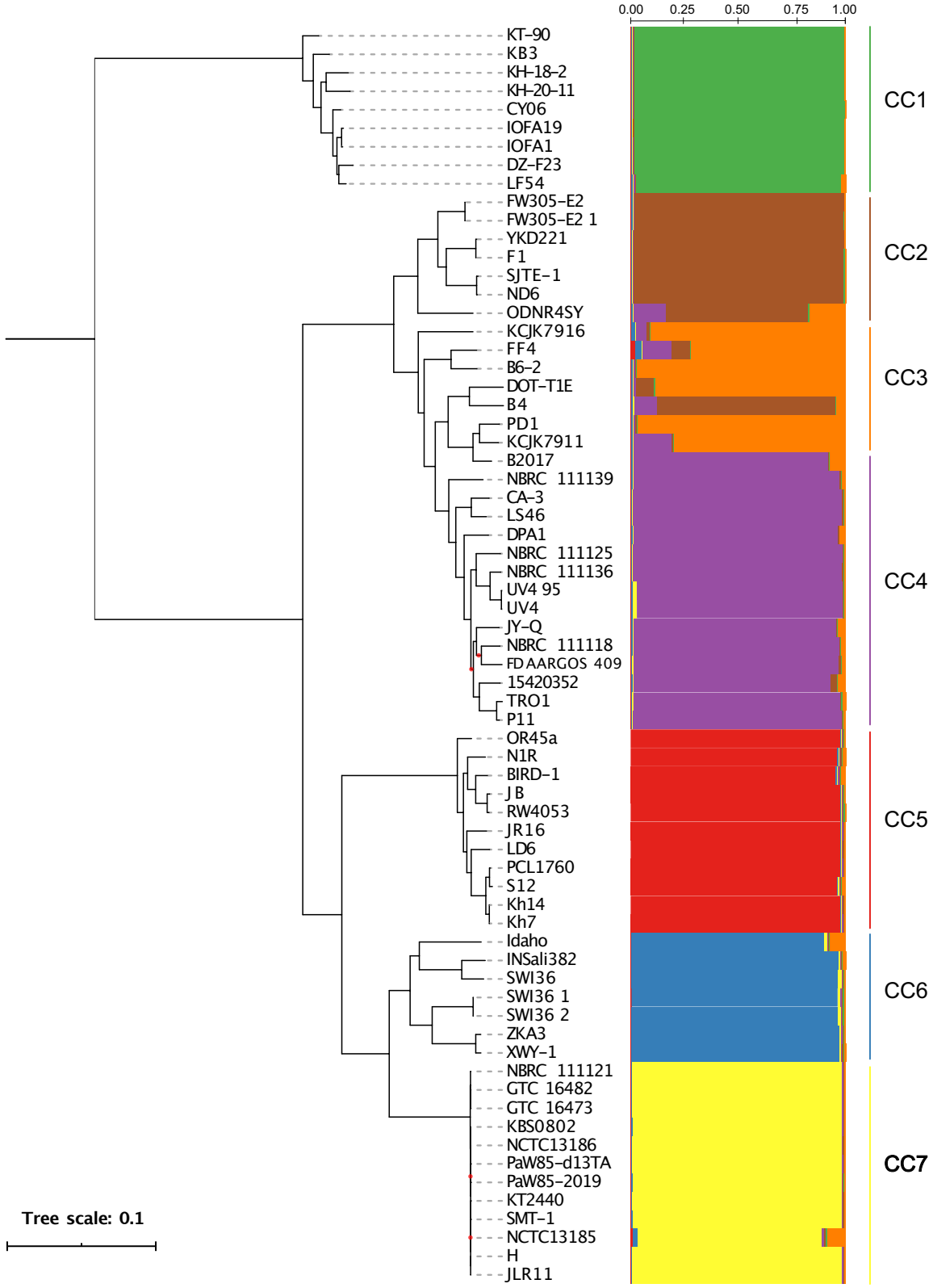


Figure 4. Phylogenetic tree and coancestry barplots, correlating the phylogenetic tree using SNPs extracted from core genome and coancestry probabilities assigned with STRUCTURE. Red dots in branches represent bootstrap values lower than 70%. Each color represents one of the seven predicted Clonal Complexes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

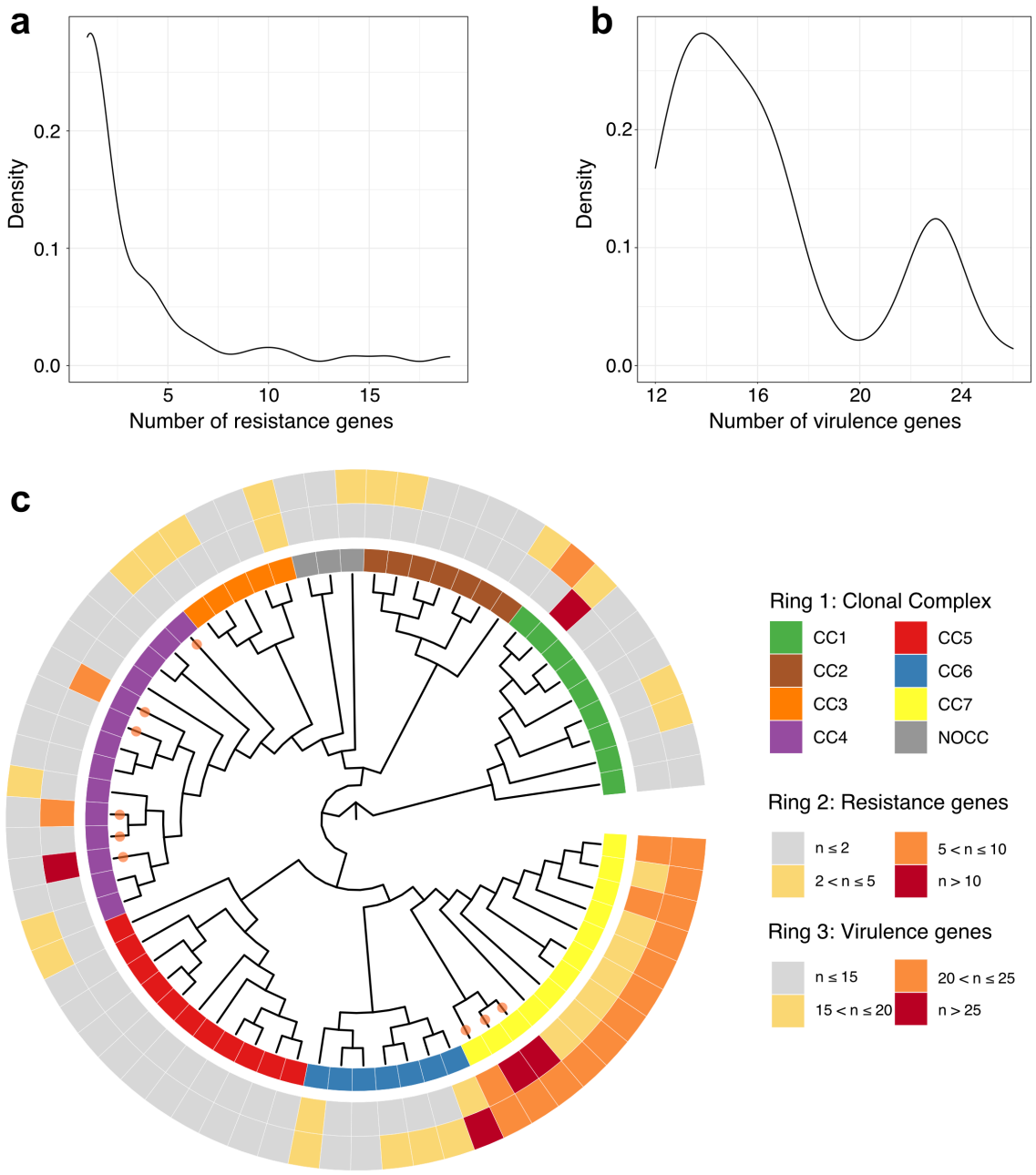


Figure 5. Population structure and distribution of virulence and resistance genes. **a, b.** distribution of acquired resistance and virulence genes. **c.** Maximum likelihood tree from SNPs present in the core genome inferred with 68 genomes used in this study. Branches from clinical isolates are marked 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.