Phylogenetic and population structure analyses uncover pervasive misclassification and help assessing the biosafety of *Pseudomonas alloputida* 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 alloputida*

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

Abstract

The *Pseudomonas putida* group comprises strains with biotechnological and clinical relevance. *P. alloputida* was proposed as a new species and highlighted the misclassification of *P. putida*. Nevertheless, the population structure of *P. alloputida* remained unexplored. We retrieved 11,025 *Pseudomonas* genomes and used *P. alloputida* Kh7^{T} to delineate the species. The *P. alloputida* population structure comprises at least 7 clonal complexes (CCs). Clinical isolates are mainly found in CC4 and acquired resistance genes are present at low frequency in plasmids. Virulence profiles support the potential of CC7 members to outcompete other plant or human pathogens through a type VI secretion system. Finally, we found that horizontal gene transfer had an important role in shaping the ability of *P. alloputida* to bioremediate aromatic compounds such as toluene. Our results provide the grounds to understand *P. alloputida* genetic diversity and safety for environmental applications.

Keywords: *Pseudomonas putida group*, *Pseudomonads*, cgMLST

1. Introduction

The genus *Pseudomonas* is diverse and composed of 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 *Pseudomonas* taxonomy was inconsistent until the division in rRNA homology groups [2]. Nevertheless, because of the limited resolution of the 16S rRNA gene for intra-specific classification, other housekeeping genes were later 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*. Species from this group are ubiquitous in soil and water, and several strains have been isolated from various environments, such as polluted soils and plant roots [5, 6]. *P. putida* species are well known to perform many functions such as plant growth promotion, bioremediation, and protection against plant pathogens [7].

Because of its high metabolic diversity, applications of *P. putida* have been proposed for environmental, industrial, and agricultural uses [5]. Among well-known *P. putida* species, *P. putida* KT2440 is a model organism to study plant-microbe interactions [8, 9]. KT2440 was isolated from garden soil in Japan and has become a workhorse for industrial biotechnology because of its high metabolic versatility, genetic accessibility, and stress-resistance [10]. In addition to the biotechnological application of *P. putida* species, some clinical isolates have been described as reservoirs of resistance genes [11, 12], raising concerns about its biosafety.

Recently, Keshavarz-Tohid et al. (2019) showed that *P. putida* KT2440 and other known *P. putida* strains (e.g. BIRD-1, F1, and DOT-T1E) are distant from the type strain *P. putida* NBRC 14164^T and hence should be classified as a member of a novel species, *Pseudomonas alloputida,* whose type strain is Kh7 (=CFBP 8484^T =LMG 29756^T) [13]. Here, we report the population structure of *P. alloputida*, which was used to estimate the diversity and distribution of bioremediation and plant growth promotion genes. Further, we 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.

2. Results and Discussion

2.1. Phylogeny and classification of *Pseudomonas 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 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 [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. alloputida* Kh7^T as an anchor-strain to evaluate the ANI values from other isolates in the *P. putida* group. The sorted distribution of ANI values from Kh7^T showed an abrupt break around 95%, supporting its effectiveness as a threshold to delineate *P. alloputida* (Figure S1b). The isolate previously classified as *P. monteilli* IOFA19 (GCA_000633915.1) had the lowest ANI value within the predicted species (95.48%), followed by a drop to 91.13% (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 containing a type strain and at least three genomes, or a cluster without a type strain, but with at least ten connected genomes. We retrieved the main species from the *P. putida* group and found clusters that most likely correspond to new species (Figure 2). Further, the species number is 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. We observed a clear separation of the clusters with type strains for *P. putida* (NBRC 14164^T) and *P. alloputida* (Kh7^T) (Figure 2). *P. putida* and *P. alloputida* comprise groups with 16 and 68 genomes, respectively (Figure 2, Table S1). The greater number of *P. alloputida* genomes might have an historical explanation. Although the phylogenetic separation of NBRC 14164^T from other main *P. putida* strains has been noticed before [18],

KT2440, a *P. alloputida* isolate [13], has also been used to categorize *P. putida* genomes over the years [6, 8, 18].

The use of NBRC 14164^T to delimit *P. putida sensu stricto* highlights that many well-known *P. putida* genomes belong to other species. Isolates that are well known for their ability to promote plant growth (W618) [19], to oxidize manganese (GB-1) [20], and to damage human tissues (HB3267) [11], are neither *P. putida* nor *P. alloputida* 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 [21]. HB3267, as well as 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 strain, confirming *P. shirazica* as an heterotypic synonym of *P. asiatica* [22]. Henceforth, we focused our analyses in the novel *P. alloputida* species because of its greater number of genomes and of the presence of key strains associated with bioremediation, plant growth promotion, and biocontrol.

2.2. 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 [23], 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; exclusive genes. By using 68 isolates, the *P. alloputida* 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). By analyzing the slope of the curve (α= 0.417), we inferred that *P. alloputida* has an open pangenome [23] (Figure S2). Our estimated α value is

much lower than the maximum threshold used to define an open pangenome ($\alpha < 1$), which is in line with a previous study [24]. Further, this low α value implies high rates of new gene families will be found if more isolates are included in the analysis.

The high number of gene families in the *P. alloputida* pangenome is explained by unique and low-frequency genes. Only 6,373 genes families (24.71%) are found between 5% to 95% of the isolates, while 15,606 (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 attributed to fragmented genomes. However, reference genomes such as DOT-T1E showed 267 unique genes, far more than the average of 160.5 unique genes found across the dataset. Although the number of low-frequency genes may be overestimated, the high prevalence of unique genes among closely-related genomes indicates a high turnover of unstable genes that might be adaptive under transient selective pressures in their environments.

We also estimated the genomic fluidity (φ) of *P. alloputida*. The φ estimator is a robust metric that represents the ratio of unique gene families to the sum of gene families, averaged over randomly chosen genomes pairs [25]. The smaller the φ, the greater the genes shared by a pair of randomly selected genomes. Analyzing φ instead of the core genome proportion provides a more realistic measure of cohesiveness within a species, particularly because low-frequency genes directly affect the pangenome size. *P. alloputida* has $\varphi = 0.20 \pm 0.04$, indicating that random pairs of *P. alloputida* genomes have an average 20% and 80% of unique and shared genes, respectively.

2.3. Population structure

Determining relationships between isolates can provide novel insights into the metabolic diversity of a given species. The Multilocus Sequence Typing (MLST) analysis is a technique to characterize genomes based on single-nucleotide polymorphisms (SNPs) within a few housekeeping genes. MLST schemes are available for several species [26]. In an MLST analysis, each combination of SNPs defines a Sequence Type (ST) that can be linked to form Clonal Complexes (CC) [26]. A variation of classical MLST is the core genome MLST (cgMLST), which provides greater resolution by using SNPs from the entire core genome [27]. Here, we used 225,009 SNPs obtained from the *P. alloputida* core genome to reconstruct the phylogenetic tree and the cgMLST profile.

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

We also performed a Discriminant Analysis of Principal Components (DAPC) [28] to recover the genes that contribute most to separate the population based on their presence/absence profiles (Figure S4). The top 50 discriminating genes separate CC5 and CC1 from the rest of the population, but fail to resolve the relationships between other CCs (Figure S5). Next, we used branch lengths from cgMLST tree as an indirect

171 estimator of diversity within the CCs (Figure 3c). CC7 is the most clonal group,
172 comprising 9 isolates, including KT2440 and three clinical isolates (GTC_16482,
173 GTC_16473, and NBRC_111121).

174 The MLST scheme for *P. putida* species comprises eight housekeeping genes
175 (*argS, gyrB, ileS, nuoC, ppsA, recA, rpoB,* and *rpoD*) [29]. We assigned isolates to STs
176 and grouped those into CCs (Table S2). This analysis revealed some incongruences
177 with previous reports [29]. For example, by using allele combinations with perfect-
178 match to predict STs, KT2440 was detected as ST69 and not as ST58 [29], indicating
179 that a revision in the public database is warranted. The predicted number of CCs was
180 also supported by the admixture model from STRUCTURE [30]. This model assumes
181 that each isolate has ancestry from one or more *K* genetically different sources, which
182 we referred to as CCs. The number of CCs corresponds to the estimated number of
183 clusters represented by parameter *K*. Instead of using the highest raw marginal
184 likelihood, we followed the protocol for estimating the best *K* value suggested by
185 Evanno et al [31]. The *ad hoc* statistics ΔK indicates that, according to our dataset, the
186 population structure of *P. alloputida* is composed of at least 7 CCs (Figure 3d).

187 Once data from eight *loci* may be insufficient to accurately describe the population
188 structure of *P. alloputida*, the populations identified by STRUCTURE were only
189 considered if they matched cgMLST results. When correlating the cgMLST tree
190 topology with ancestry proportion predicted for each isolate, we observed a clear
191 delimitation of genetic blocks for each CC (Figure 4). However, there are a few
192 inconsistencies. For example, isolate B4 has a greater ancestry proportion with CC2,
193 although cgMLST indicates its greater proximity to CC3. In addition, KCJK7916, FF4,
194 and B6-2 were not assigned to a CC because, although they have a higher proportion

of ancestry with CC3, they are paraphyletic to CC3 and CC4. Since CCs are assumed to be monophyletic, these strains were designated as No Clonal Complex (NOCC). We expect that a greater number of *P. alloputida* genomes from isolates from various sources will improve the resolution of the *P. alloputida* population structure, including the CC assignment to isolates described here as NOCC.

The cgMLST phylogenetic tree and the distance tree support CC1 as the basally branching group of *P. alloputida* (Figure 4). CC1 comprises isolates from deep-sea sediments from Indian Ocean, costal water from the Pacific Ocean, lotus field, and arthropods (Table 1). The main difference of CC1 is the lack of 229 gene families in the accessory genome, which are present in at least one isolate from all other CCs (Table S3). Among these absent gene families, there is a genomic island with nearly 46 Kbp encompassing 38 genes. Some of those genes are involved in sugar transport, as previously identified in KT2440 (CC7) [6], as well as genes encoding hypothetical proteins (coordinates 3,126,465-3,172,496 in KT2440).

2.4. Resistance profiles

We evaluated the composition of antibiotic resistance genes using the CARD database [32]. All 15 different genes in the core resistome encode MDR efflux pumps (Table 2, Table S4), including MexAB-OprM, MexEF-OprN, and MexJK, from the resistance-nodulation-cell division (RND) efflux pumps family. These efflux pumps are associated with intrinsic and acquired multidrug resistance in *P. aeruginosa* [33, 34]. However, these RND efflux pumps may play an alternative role in *P. alloputida* by pumping out toxic substances such as toluene [35]. We also identified *cpxR*, which

encodes a protein that promotes MexAB-OprM expression in the absence of the MexR repressor in *P. aeruginosa* [36], which is absent in *P. alloputida*. The presence of MexAB-OprM in the core genome, under CpxR regulation, supports its involvement with intrinsic physiology in addition to drug resistance, because this complex can be involved in both quorum-sensing and mediation of *P. aeruginosa*-host interaction [37, 38].

Regarding the acquired resistome, we found 45 different genes that confer resistance by pumping out or inactivating antibiotics, as well as by interacting with antibiotic targets (Table S4). These genes are distributed at low-frequency (Figure 5a), indicating that most of them are strain-specific or acquired through horizontal gene transfer. In general, there is no clear correlation between acquired resistome and population structure (Figure S6), although CC7 has more acquired resistance genes than other CCs (Figure 5c, Figure S6).

Our results highlight clinical strains harboring a range of resistance genes. In total, 9 out of 68 (13.2%) *P. alloputida* genomes analyzed here belong to clinical strains. Along with efflux pumps, the acquired resistome includes genes encoding antibiotic-inactivating enzymes that confer resistance to beta-lactams (e.g. bla_{CARB-3}, bla_{IMP-1}, bla_{OXA-2}, bla_{PDC-7}, bla_{TEM-1}, and bla_{VIM-2}); to aminoglycosides (e.g. *aac(6')-IIa*, *aadA*, *aph(3'')-Ia*, and *aph(6')-Id*; chloramphenicol (*cat*) and; to fosfomycin (*fosA*) (Table S2). These genes were distributed in few strains, mostly clinically relevant (Figure S6, Table S4); the top four strains with more acquired resistance genes were GTC_16473 (19 genes), GTC_16482 (16 genes), DZ-F23 (14 genes), and 15420352 (11 genes). Importantly, all of these strains (except DZ-F23) are clinical.

We evaluated the presence of acquired genes in plasmids predicted with the PLSDB database (version 2020_06_29) [39]. *P. alloputida* GTC_16473 contained the genes *aac(6')-IIa*, *aadA23*, and bla_{CARB-3} located in a scaffold with high identity with the pJR2 plasmid from *Pasteurella multocida* (NC_004772.1). We also identified bla_{OXA-2}, *aadA22*, *aac(6')-Ia*, *aac(6')-IIc*, *aph(3'')-Ib*, *aph(6')-Id*, bla_{IMP-1}, and *sul1* genes in plasmid-like sequences in *P. alloputida* GTC_16473 that have not reached the coverage thresholds to be reliably classified as plasmids, supporting an underestimation of plasmids in *P. alloputida* isolates. *P. alloputida* XWY-1 (CC6) also contains a plasmid, pXWY-1 (NZ_CP026333.1), which harbors the resistance genes *sul1*, *aadA2,* and *qacH*. This strain was isolated from rice fields in China. Finding non-clinical strains harboring plasmids with such relevant resistance genes warrants further investigation.

2.5. Virulence profiles

We used the VFDB database [40] to assess the *P. alloputida* virulence profiles. Although strains such as KT2440 have been approved as a host-vector system safety level 1 [41], the genetic proximity with animal (*P. aeruginosa*) and plant (*P. syringae*) pathogens reinforces the need to evaluate the biosafety of *P. putida* species. The core virulome of *P. alloputida* contains genes associated with twitching motility, siderophore production (pyoverdine), and alginate biosynthesis (Table S5). Importantly, *P. alloputida* lacks key virulence genes usually found in *P. aeruginosa*, such as those encoding exotoxin A, alkaline protease, elastase, rhamnolipid biosynthesis pathway components, phospholipase C, plant cell wall-degrading enzymes, and type III secretion system. Although we have detected genes associated with the type III

secretion system in both core (HopAJ2 and HopJ1) and acquired resistome (HopAN1), these genes were experimentally ruled out as virulence factors [42].

Most of the alginate regulatory and biosynthesis systems are present in the *P. alloputida* core virulome (Table S5). The overproduction of alginate plays an important role in inducing the mucoid morphotype in *P. aeruginosa* during chronic infection [43]. However, the transcriptional regulatory gene *algM* (*mucC*), associated with genotypic switching, is absent in *P. alloputida*. Since the lack of AlgM leads to the non-mucoid phenotype in *P. aeruginosa* [8], it can account for the non-mucoid phenotype in *P. alloputida* under no water stress conditions. Although the genotypic switching is unlikely to occur in *P. alloputida*, alginate production is important under water stress conditions by promoting biofilm development and protecting from desiccation [44].

The acquired virulome comprised genes for type II and VI secretions systems, adherence, and iron uptake (Table S5). While most isolates had a low frequency of resistance genes, the virulence factors displayed a bimodal distribution (Figure 5b), a pattern that has been previously observed for *Klebsiella aerogenes* [45]. Genes from the acinetobactin gene cluster and HSI-I type VI secretion system were differentially distributed across CCs (Figure 6a), although it remains unclear whether these patterns emerged mainly from gene gain or loss. In *Acinetobacter baumannii*, iron uptake is mainly performed by the siderophore acinetobactin [46], which is synthesized by the proteins encoded by the *bauABCDE* operon. We found this operon in *P. alloputida* strains within CC5, CC6, and CC7 (Figure 6a). Further, this operon was surrounded by genes coding for proteins from chemotaxis sensory transducer (PP_2599) and aminotransferases (PP_2588) families (Figure 6b). Siderophore-mediated iron

288 acquisition has been investigated in *P. alloputida* KT2440 [47] (CC7), but the role
289 played by *bauABCDE* in this species is yet to be elucidated.

290 Present in all strains from other CCs, the type VI secretion system (T6SS) HSI-I is
291 absent in CC5 and CC6 (except XWY-1) isolates (Figure 6a). In KT2440 (CC7), HSI-I
292 is a potent weapon against other bacteria (e.g. phytopathogens), increasing the
293 competitiveness of *P. alloputida* [48]. The absence of T6SS virulence genes has been
294 reported for BIRD-1 (CC5) [48], and our work generalizes this observation to all CC5
295 members. In KT2440, T6SS is crucial to kill phytopathogens such as *Xanthomonas*
296 *campestris* [49]. These results indicate that BIRD-1 and other members from CC5 are
297 likely less efficient than KT2440 as biocontrol agents. Moreover, all clinical *P.*
298 *alloputida* strains harbor T6SS, indicating their potential ability to outcompete other
299 bacteria during infections.

300

301 **2.6. Plant growth promotion and bioremediation properties**

302 The ability of *P. putida* species to promote plant growth and bioremediate toxic
303 compounds have been explored [8, 10, 50, 51]. We searched for genes involved in
304 plant growth promotion and bioremediation through a literature search of genes that
305 have already been described for *Pseudomonas* species (Table S6). We found genes
306 in the core genome, such as the pyrroloquinoline quinone-encoding operon
307 *pqqBCDEFG,* associated with mineral phosphate solubilization in *Serratia marcescens*
308 [52] and *Pseudomonas fluorescens* [53] (Table S7). Mineral solubilization has already
309 been reported experimentally for BIRD-1 [50] and KT2440 [53], indicating that all *P.*
310 *alloputida* isolates are genetically equipped to solubilize inorganic phosphate.

Another key feature that can enhance plant growth is the colonization of seeds by *P. putida* [54]. In KT2440, genes associated with surface adhesion (e.g. *lapA*, *lapBCD*), flagellum biosynthesis (e.g. *flhB*, *fliF*, *fliD*, *fliC*), and virulence regulation (e.g. *rpoN* and *gacS*) have been experimentally shown to be important for attachment to corn seeds [54]. All these genes, except *lapA* and *fliC*, belongs to the *P. alloputida* core genome (Table S7). Although often described as virulence genes, flagellum genes also play roles in the association of *P. putida* with plants. Other plant growth-promoting genes were also found in the *P. alloputida* accessory genome (Table S7), except for CC1, which lacks all putative plant growth-promoting genes.

Many plant growth-promoting bacteria can synthesize phytohormones such as indole-3-acetic acid. *P. alloputida* lacks the main genes involved in this process (e.g. *ipdC*, *iaaM*, and *iaaH*), indicating an incomplete or nonfunctional pathway. Although *P. putida* W619 (GCA_000019445.1) is considered one of the most efficient producers of IAA in comparison with other endophytic bacteria [19], this strain belongs to a different species. In addition, *P. alloputida* also lacks AcdS, an enzyme that counteract ethylene stress response, a result that has been experimentally confirmed in KT2440 [55].

Besides the ability to promote plant growth, *P. putida* can tolerate or degrade an array of compounds including heavy metals and hydrocarbons. We identified genes in the core genome that allow *P. alloputida* to resist various heavy metals such as copper (*cop* genes) and cobalt/zinc/cadmium (*czcABC*) (Table S8). The copper/silver resistance operon, *cusABC*, was present in the accessory genome. Although we found a wide range of genes associated with bioremediation, there is no clear correlation between the population structure and presence/absence profiles of such genes.

P. putida is known for its capacity to metabolize aromatic hydrocarbons such as toluene, benzene, and *p*-cymene [6, 8]. The toluene-degrading pathway includes the *todABCDE* operon and the *todST* regulator. The *p*-cymene compound can be degraded by means of the *cymAaAbBCDER* or the *cmtAaAbAcAdBCDEFGHI* operon [56]. We found a genomic island of approximately 48kb long harboring all these genes in F1, DOT-T1E, UV4, UV4/95, YKD221, and NBRC_111125 genomes (Figure S9). All these isolates, except NBRC_111125, were experimentally confirmed to degrade toluene. Further, *P. alloputida* F1 is well-known to grow on toluene [6]; YKD221 was isolated from contaminated industrial soil and degrades *cis*-dichloroethene [57]; DOT-T1E is an isolate known to grow on different carbon sources [58] and; UV4 and UV4/95 conduct important industrial biotransformation of arenes, alkenes, and phenols [59]. Interestingly, genes in this genomic island presented a very similar genetic context (Figure S7), with an upstream arm-type integrase associated with bacteriophages. We were unable to precisely define the *att* viral sites, indicating a deterioration of the original structure of the putative bacteriophage. Further, the lack of correspondence between population structure and the presence of an integrase upstream the genomic island indicates that this region was likely acquired via independent horizontal gene transfers in distinct *P. alloputida* CCs.

We also identified RND efflux pumps involved with solvent tolerance in both core (TtgABC) and accessory genomes (TtgDEF and TtgGHI). TtgABC, TtgDEF, and TtgGHI are required for DOT-T1E to efficiently tolerate toluene [35]. We observed that TtgABC is the same protein-complex predicted as MexAB-OprM, associated with antibiotic resistance in the core genome. This complex extrudes both antibiotics and solvents such as toluene in *P. alloputida* DOT-T1E [35], corroborating the additional and important function to extrude antibiotics and organic solvents in all *P. alloputida*

359 isolates. TtgDEF is located in the same genomic island of *tod* genes. This complex
360 can expel toluene, but not antibiotics [60], reinforcing the variety of molecules that can
361 be extruded by RND efflux pumps and the need to explore the structural basis of this
362 specificity, not only in *P. alloputida* isolates, but also in other bacteria.

363

364 **3. Concluding remarks**

365 Through a remarkable metabolic versatility, *P. putida* species can thrive in a wide
366 variety of niches. In this work, we explored the genetic diversity of *P. alloputida* and
367 characterized its population structure for the first time. Through a large-scale genomic
368 analysis, we identified a major problem with *P. putida species* classification, including
369 several reference strains that likely belong to new species, as also suggested
370 elsewhere [13]. *P. alloputida* has an open pangenome dominated by low-frequency
371 genes. The population structure of this species has at least 7 clonal complexes that
372 were verified by cgMLST and STRUCTURE ancestry simulations. We found that most
373 of the clinical isolates belong to CC4. Together, our results indicate that *P. alloputida*
374 clinical isolates are mainly opportunistic and do not pose considerable health concerns.

375 We analyzed genes of clinical and biotechnological interest. *P. alloputida* has
376 several RND-family efflux pumps that are important to tolerate antibiotics and other
377 toxic compounds such as toluene. The low-frequency acquired resistance genes are
378 predominant in plasmids from a few clinical strains. The detected virulence genes allow
379 *P. alloputida* to synthesize pyoverdine, attach to seed surfaces, and kill other bacteria
380 (including pathogens) through a type VI secretion system. KT2440 is a member of CC7
381 that has the T6SS along with an operon for acinetobactin biosynthesis. *P. alloputida*

lacks key genes for the production of indole-3-acetic acid. We also observed that the genes for the degradation of some aromatic compounds, including toluene, were likely horizontally acquired. Our results provide an opportunity for the development of biotechnological applications as well as insights into the genomic diversity of the novel species *P. alloputida*.

4. Methods

4.1. Datasets and genomic features

We recovered 11,025 genomes from the *Pseudomonas* genus in June 2020. To assess the quality of the genomes, we used BUSCO v4.0.6 [14] with a minimum threshold of 90% completeness. The Kh7^{T} (GCA_900291035.1) was used as a reference with mash v.2.2.2 [61] to find genomes with distances up to 0.05. We used mashtree [15] to generate the distance tree. The ANI analysis was performed with pyani 0.2.10 [62]. Network analysis was conducted in R with the igraph package (https://igraph.org). We removed S12 (GCA_000287915.1) and KT2440 (GCA_000007565.2) because they were duplicated genomes. Type strains and accession numbers used to define clusters in the network analysis are available in the Table S1. Gene prediction in all isolates was conducted with prokka v1.12 [63] to avoid bias in the identification of protein families. Plasmids were analyzed with PLSD v2020_06_29 [39].

4.2. Pangenome characterization

We inferred the *P. alloputida* pangenome using Roary 3.13.0 [64], with a minimum threshold of 85% identity to cluster proteins. Core genes were defined as those present in more than 95% of the isolates. Jaccard distances were computed by using accessory genes with prevalence between 5% and 95%. Gene content variations between *P. alloputida* ecotypes were inferred with a discriminant analysis of principal components (DAPC) using the *ade4* and *adegenet* packages [28], retaining the 30 principal components and 3 discriminant functions. Pangenome openness and fluidity were conducted with micropan [65] with 500 and 1000 permutations, respectively.

4.3. Population structure analysis

We used *in-house* scripts to extract the genes present in all isolates, which were aligned with MAFFT v7.467 [66]. SNPs were retrieved with snp-sites v2.3.3 [67] and SNP alignment was used as input to RAxML v8 [68] to reconstruct the phylogenetic tree using the general time-reversible model and gamma correction. Since we used only variable sites as input, we used ASC_GTRGAMMA to correct ascertainment bias with the Paul Lewis correction. One thousand bootstrap replicates were generated to assess the significance of internal nodes. We inferred the cgMLST scheme using the core genome SNP phylogenetic tree. The phylogenetic tree was visualized with iTOL v4 [69].

We downloaded the *P. putida* MLST scheme (on June, 2020) containing 116 different STs (https://pubmlst.org/databases/). This scheme was designed for the whole *P. putida* group, not only *P. putida* sensu stricto [29]. We used BLASTN [70] to determine the best-matching MLST allele to access STs. The allelic profile associated with each ST in our dataset was used to conduct population assignment with STRUCTURE v2.3.4 [30] with admixture model. The length of Markov chain Monte

Carlo (MCMC) was 50,000, discarding 20,000 iterations as burn-in. The simulations to calculate the parameter K ranged from 2 to 20, with 20 replicates for each K to estimate confidence intervals. Instead of using raw posterior probability to get the best K, we followed the protocol suggested by Evanno, Regnaut and Goudet [31]. Briefly, we calculated the first and second derivatives, resulting in a ΔK of 7. Therefore, we used K = 7 to analyze predicted ancestry probabilities.

4.4. Detection of genes associated with antimicrobial resistance, virulence, plant growth promotion, and bioremediation

We used the Comprehensive Antimicrobial Resistance Database (CARD) database v3.0.9 [32] to predict antibiotic resistance genes. The virulence factor database (VFDB) [40] was used to determine virulence genes. This database was downloaded on July 30 2020 and comprises 28,639 proteins associated with virulence in several pathogens. We used virulence genes previously described for the *Pseudomonas* genus. We clustered proteins based on 70% identity to build a non-redundant database using uclust v1.2.22q [71]. We built the database with plant growth promotion and bioremediation through literature searches (Table S6). All predicted proteins were globally aligned against these databases using usearch v11.0.667 [71] with 50% minimum coverage for query and subject and 60% minimum identity.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgements

449 This work was supported by Fundação Carlos Chagas Filho de Amparo à Pesquisa do
450 Estado do Rio de Janeiro (FAPERJ; grants E-26/203.309/2016 and E-
451 26/203.014/2018), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior -
452 Brasil (CAPES; Finance Code 001), and Conselho Nacional de Desenvolvimento
453 Científico e Tecnológico. The funding agencies had no role in the design of the study
454 and collection, analysis, and interpretation of data and in writing.

455

456 **REFERENCES**

457 [1] N.J. Palleroni, The Pseudomonas story, Environ Microbiol, 12 (2010) 1377-1383.
458 [2] R.Y. Stanier, N.J. Palleroni, M. Doudoroff, The aerobic pseudomonads: a taxonomic study, J Gen
459 Microbiol, 43 (1966) 159-271.
460 [3] L. Ait Tayeb, E. Ageron, F. Grimont, P.A. Grimont, Molecular phylogeny of the genus Pseudomonas
461 based on rpoB sequences and application for the identification of isolates, Res Microbiol, 156
462 (2005) 763-773.
463 [4] M. Gomila, A. Pena, M. Mulet, J. Lalucat, E. Garcia-Valdes, Phylogenomics and systematics in
464 Pseudomonas, Front Microbiol, 6 (2015) 214.
465 [5] D.C. Volke, P. Calero, P.I. Nikel, Pseudomonas putida, Trends Microbiol, 28 (2020) 512-513.
466 [6] X. Wu, S. Monchy, S. Taghavi, W. Zhu, J. Ramos, D. van der Lelie, Comparative genomics and
467 functional analysis of niche-specific adaptation in Pseudomonas putida, FEMS Microbiol Rev, 35
468 (2011) 299-323.
469 [7] K.N. Timmis, Pseudomonas putida: a cosmopolitan opportunist par excellence, Environ Microbiol,
470 4 (2002) 779-781.
471 [8] K.E. Nelson, C. Weinel, I.T. Paulsen, R.J. Dodson, H. Hilbert, V.A. Martins dos Santos, D.E. Fouts, S.R.
472 Gill, M. Pop, M. Holmes, L. Brinkac, M. Beanan, R.T. DeBoy, S. Daugherty, J. Kolonay, R. Madupu,
473 W. Nelson, O. White, J. Peterson, H. Khouri, I. Hance, P. Chris Lee, E. Holtzapple, D. Scanlan, K.
474 Tran, A. Moazzez, T. Utterback, M. Rizzo, K. Lee, D. Kosack, D. Moestl, H. Wedler, J. Lauber, D.
475 Stjepandic, J. Hoheisel, M. Straetz, S. Heim, C. Kiewitz, J.A. Eisen, K.N. Timmis, A. Dusterhoft, B.
476 Tummler, C.M. Fraser, Complete genome sequence and comparative analysis of the
477 metabolically versatile Pseudomonas putida KT2440, Environ Microbiol, 4 (2002) 799-808.

478 [9] E. Belda, R.G. van Heck, M. Jose Lopez-Sanchez, S. Cruveiller, V. Barbe, C. Fraser, H.P. Klenk, J.
479 Petersen, A. Morgat, P.I. Nikel, D. Vallenet, Z. Rouy, A. Sekowska, V.A. Martins Dos Santos, V. de
480 Lorenzo, A. Danchin, C. Medigue, The revisited genome of Pseudomonas putida KT2440
481 enlightens its value as a robust metabolic chassis, Environ Microbiol, 18 (2016) 3403-3424.
482 [10] P.I. Nikel, V. de Lorenzo, Pseudomonas putida as a functional chassis for industrial biocatalysis:
483 From native biochemistry to trans-metabolism, Metab Eng, 50 (2018) 142-155.
484 [11] M. Fernandez, M. Porcel, J. de la Torre, M.A. Molina-Henares, A. Daddaoua, M.A. Llamas, A. Roca,
485 V. Carriel, I. Garzon, J.L. Ramos, M. Alaminos, E. Duque, Analysis of the pathogenic potential of
486 nosocomial Pseudomonas putida strains, Front Microbiol, 6 (2015) 871.
487 [12] L. Molina, Z. Udaondo, E. Duque, M. Fernandez, P. Bernal, A. Roca, J. de la Torre, J.L. Ramos,
488 Specific Gene Loci of Clinical Pseudomonas putida Isolates, PLoS One, 11 (2016) e0147478.
489 [13] V. Keshavarz-Tohid, J. Vacheron, A. Dubost, C. Prigent-Combaret, P. Taheri, S. Tarighi, S.M.
490 Taghavi, Y. Moenne-Loccoz, D. Muller, Genomic, phylogenetic and catabolic re-assessment of
491 the Pseudomonas putida clade supports the delineation of Pseudomonas alloputida sp. nov.,
492 Pseudomonas inefficax sp. nov., Pseudomonas persica sp. nov., and Pseudomonas shirazica sp.
493 nov, Syst Appl Microbiol, 42 (2019) 468-480.
494 [14] M. Seppey, M. Manni, E.M. Zdobnov, BUSCO: Assessing Genome Assembly and Annotation
495 Completeness, Methods Mol Biol, 1962 (2019) 227-245.
496 [15] L.S. Katz, T. Griswold, S.S. Morrison, J.A. Caravas, S. Zhang, H.C. den Bakker, X. Deng, H.A. Carleton,
497 Mashtree: a rapid comparison of whole genome sequence files, Journal of Open Source
498 Software, 4 (2019).
499 [16] M. Richter, R. Rossello-Mora, Shifting the genomic gold standard for the prokaryotic species
500 definition, Proc Natl Acad Sci U S A, 106 (2009) 19126-19131.
501 [17] L.M. Bobay, The Prokaryotic Species Concept and Challenges, in: H. Tettelin, D. Medini (Eds.) The
502 Pangenome: Diversity, Dynamics and Evolution of Genomes, Cham (CH), 2020, pp. 21-49.
503 [18] K. Yonezuka, J. Shimodaira, M. Tabata, S. Ohji, A. Hosoyama, D. Kasai, A. Yamazoe, N. Fujita, T.
504 Ezaki, M. Fukuda, Phylogenetic analysis reveals the taxonomically diverse distribution of the
505 Pseudomonas putida group, J Gen Appl Microbiol, 63 (2017) 1-10.
506 [19] S. Taghavi, C. Garafola, S. Monchy, L. Newman, A. Hoffman, N. Weyens, T. Barac, J. Vangronsveld,
507 D. van der Lelie, Genome survey and characterization of endophytic bacteria exhibiting a
508 beneficial effect on growth and development of poplar trees, Appl Environ Microbiol, 75 (2009)
509 748-757.
510 [20] R.A. Rosson, K.H. Nealson, Manganese binding and oxidation by spores of a marine bacillus, J
511 Bacteriol, 151 (1982) 1027-1034.

[21] L. Molina, Z. Udaondo, E. Duque, M. Fernandez, C. Molina-Santiago, A. Roca, M. Porcel, J. de la Torre, A. Segura, P. Plesiat, K. Jeannot, J.L. Ramos, Antibiotic resistance determinants in a Pseudomonas putida strain isolated from a hospital, PLoS One, 9 (2014) e81604.

[22] M. Tohya, S. Watanabe, T. Tada, H.H. Tin, T. Kirikae, Genome analysis-based reclassification of Pseudomonas fuscovaginae and Pseudomonas shirazica as later heterotypic synonyms of Pseudomonas asplenii and Pseudomonas asiatica, respectively, Int J Syst Evol Microbiol, 70 (2020) 3547-3552.

[23] H. Tettelin, V. Masignani, M.J. Cieslewicz, C. Donati, D. Medini, N.L. Ward, S.V. Angiuoli, J. Crabtree, A.L. Jones, A.S. Durkin, R.T. Deboy, T.M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J.D. Peterson, C.R. Hauser, J.P. Sundaram, W.C. Nelson, R. Madupu, L.M. Brinkac, R.J. Dodson, M.J. Rosovitz, S.A. Sullivan, S.C. Daugherty, D.H. Haft, J. Selengut, M.L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K.J. O'Connor, S. Smith, T.R. Utterback, O. White, C.E. Rubens, G. Grandi, L.C. Madoff, D.L. Kasper, J.L. Telford, M.R. Wessels, R. Rappuoli, C.M. Fraser, Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome", Proc Natl Acad Sci U S A, 102 (2005) 13950-13955.

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

[25] A.O. Kislyuk, B. Haegeman, N.H. Bergman, J.S. Weitz, Genomic fluidity: an integrative view of gene diversity within microbial populations, BMC Genomics, 12 (2011) 32.

[26] K.A. Jolley, J.E. Bray, M.C.J. Maiden, Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications, Wellcome Open Res, 3 (2018) 124.

[27] M. de Been, M. Pinholt, J. Top, S. Bletz, A. Mellmann, W. van Schaik, E. Brouwer, M. Rogers, Y. Kraat, M. Bonten, J. Corander, H. Westh, D. Harmsen, R.J. Willems, Core Genome Multilocus Sequence Typing Scheme for High- Resolution Typing of Enterococcus faecium, J Clin Microbiol, 53 (2015) 3788-3797.

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

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

[30] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics, 155 (2000) 945-959.

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

[32] B.P. Alcock, A.R. Raphenya, T.T.Y. Lau, K.K. Tsang, M. Bouchard, A. Edalatmand, W. Huynh, A.V. Nguyen, A.A. Cheng, S. Liu, S.Y. Min, A. Miroshnichenko, H.K. Tran, R.E. Werfalli, J.A. Nasir, M.

547 Oloni, D.J. Speicher, A. Florescu, B. Singh, M. Faltyn, A. Hernandez-Koutoucheva, A.N. Sharma,
548 E. Bordeleau, A.C. Pawlowski, H.L. Zubyk, D. Dooley, E. Griffiths, F. Maguire, G.L. Winsor, R.G.
549 Beiko, F.S.L. Brinkman, W.W.L. Hsiao, G.V. Domselaar, A.G. McArthur, CARD 2020: antibiotic
550 resistome surveillance with the comprehensive antibiotic resistance database, Nucleic Acids
551 Res, 48 (2020) D517-D525.
552 [33] P.D. Lister, D.J. Wolter, N.D. Hanson, Antibacterial-resistant Pseudomonas aeruginosa: clinical
553 impact and complex regulation of chromosomally encoded resistance mechanisms, Clin
554 Microbiol Rev, 22 (2009) 582-610.
555 [34] H.P. Schweizer, Efflux as a mechanism of resistance to antimicrobials in Pseudomonas aeruginosa
556 and related bacteria: unanswered questions, Genet Mol Res, 2 (2003) 48-62.
557 [35] A. Rojas, E. Duque, G. Mosqueda, G. Golden, A. Hurtado, J.L. Ramos, A. Segura, Three efflux pumps
558 are required to provide efficient tolerance to toluene in Pseudomonas putida DOT-T1E, J
559 Bacteriol, 183 (2001) 3967-3973.
560 [36] Z.X. Tian, X.X. Yi, A. Cho, F. O'Gara, Y.P. Wang, CpxR Activates MexAB-OprM Efflux Pump
561 Expression and Enhances Antibiotic Resistance in Both Laboratory and Clinical nalB-Type Isolates
562 of Pseudomonas aeruginosa, PLoS Pathog, 12 (2016) e1005932.
563 [37] K. Evans, L. Passador, R. Srikumar, E. Tsang, J. Nezezon, K. Poole, Influence of the MexAB-OprM
564 multidrug efflux system on quorum sensing in Pseudomonas aeruginosa, J Bacteriol, 180 (1998)
565 5443-5447.
566 [38] P. Sanchez, J.F. Linares, B. Ruiz-Diez, E. Campanario, A. Navas, F. Baquero, J.L. Martinez, Fitness of
567 in vitro selected Pseudomonas aeruginosa nalB and nfxB multidrug resistant mutants, J
568 Antimicrob Chemother, 50 (2002) 657-664.
569 [39] V. Galata, T. Fehlmann, C. Backes, A. Keller, PLSDB: a resource of complete bacterial plasmids,
570 Nucleic Acids Res, 47 (2019) D195-D202.
571 [40] B. Liu, D. Zheng, Q. Jin, L. Chen, J. Yang, VFDB 2019: a comparative pathogenomic platform with
572 an interactive web interface, Nucleic Acids Res, 47 (2019) D687-D692.
573 [41] L.F.C. Kampers, R.J.M. Volkers, V.A.P. Martins Dos Santos, Pseudomonas putida KT2440 is HV1
574 certified, not GRAS, Microb Biotechnol, 12 (2019) 845-848.
575 [42] A.P. Macho, J. Ruiz-Albert, P. Tornero, C.R. Beuzon, Identification of new type III effectors and
576 analysis of the plant response by competitive index, Mol Plant Pathol, 10 (2009) 69-80.
577 [43] R.L. Henry, C.M. Mellis, L. Petrovic, Mucoid Pseudomonas aeruginosa is a marker of poor survival
578 in cystic fibrosis, Pediatr Pulmonol, 12 (1992) 158-161.
579 [44] W.S. Chang, M. van de Mortel, L. Nielsen, G. Nino de Guzman, X. Li, L.J. Halverson, Alginate
580 production by Pseudomonas putida creates a hydrated microenvironment and contributes to

biofilm architecture and stress tolerance under water-limiting conditions, J Bacteriol, 189 (2007) 8290-8299.

[45] H. Passarelli-Araujo, J.K. Palmeiro, K.C. Moharana, F. Pedrosa-Silva, L.M. Dalla-Costa, T.M. Venancio, Genomic analysis unveils important aspects of population structure, virulence, and antimicrobial resistance in Klebsiella aerogenes, FEBS J, 286 (2019) 3797-3810.

[46] W.Y. Song, H.J. Kim, 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 (2020) 477-487.

[47] S. Matthijs, G. Laus, J.M. Meyer, K. Abbaspour-Tehrani, M. Schafer, H. Budzikiewicz, P. Cornelis, Siderophore-mediated iron acquisition in the entomopathogenic bacterium Pseudomonas entomophila L48 and its close relative Pseudomonas putida KT2440, Biometals, 22 (2009) 951-964.

[48] P. Bernal, L.P. Allsopp, A. Filloux, M.A. Llamas, The Pseudomonas putida T6SS is a plant warden against phytopathogens, ISME J, 11 (2017) 972-987.

[49] S. Validov, F. Kamilova, S. Qi, D. Stephan, J.J. Wang, N. Makarova, B. Lugtenberg, Selection of bacteria able to control Fusarium oxysporum f. sp. radicis-lycopersici in stonewool substrate, J Appl Microbiol, 102 (2007) 461-471.

[50] A. Roca, P. Pizarro-Tobias, Z. Udaondo, M. Fernandez, M.A. Matilla, M.A. Molina-Henares, L. Molina, A. Segura, E. Duque, J.L. Ramos, Analysis of the plant growth-promoting properties encoded by the genome of the rhizobacterium Pseudomonas putida BIRD-1, Environ Microbiol, 15 (2013) 780-794.

[51] T. Gong, R. Liu, Y. Che, X. Xu, F. Zhao, H. Yu, C. Song, Y. Liu, C. Yang, Engineering Pseudomonas putida KT2440 for simultaneous degradation of carbofuran and chlorpyrifos, Microb Biotechnol, 9 (2016) 792-800.

[52] F.P. Matteoli, H. Passarelli-Araujo, R.J.A. Reis, L.O. da Rocha, E.M. de Souza, L. Aravind, F.L. Olivares, T.M. Venancio, Genome sequencing and assessment of plant growth-promoting properties of a Serratia marcescens strain isolated from vermicompost, BMC Genomics, 19 (2018) 750.

[53] S.H. Miller, P. Browne, C. Prigent-Combaret, E. Combes-Meynet, J.P. Morrissey, F. O'Gara, Biochemical and genomic comparison of inorganic phosphate solubilization in Pseudomonas species, Environ Microbiol Rep, 2 (2010) 403-411.

[54] E. Duque, J. de la Torre, P. Bernal, M.A. Molina-Henares, M. Alaminos, M. Espinosa-Urgel, A. Roca, M. Fernandez, S. de Bentzmann, J.L. Ramos, Identification of reciprocal adhesion genes in pathogenic and non-pathogenic Pseudomonas, Environ Microbiol, 15 (2013) 36-48.

616 [55] B.R. Glick, B. Todorovic, J. Czarny, Z. Cheng, J. Duan, B. McConkey, Promotion of Plant Growth by
617 Bacterial ACC Deaminase, Critical Reviews in Plant Sciences, 26 (2007).
618 [56] R.W. Eaton, p-Cymene catabolic pathway in Pseudomonas putida F1: cloning and characterization
619 of DNA encoding conversion of p-cymene to p-cumate, J Bacteriol, 179 (1997) 3171-3180.
620 [57] K. Yonezuka, N. Araki, J. Shimodaira, S. Ohji, A. Hosoyama, M. Numata, A. Yamazoe, D. Kasai, E.
621 Masai, N. Fujita, T. Ezaki, M. Fukuda, Isolation and characterization of a bacterial strain that
622 degrades cis-dichloroethenein the absence of aromatic inducers, J Gen Appl Microbiol, 62 (2016)
623 118-125.
624 [58] C. Daniels, P. Godoy, E. Duque, M.A. Molina-Henares, J. de la Torre, J.M. Del Arco, C. Herrera, A.
625 Segura, M.E. Guazzaroni, M. Ferrer, J.L. Ramos, Global regulation of food supply by
626 Pseudomonas putida DOT-T1E, J Bacteriol, 192 (2010) 2169-2181.
627 [59] T. Skvortsov, P. Hoering, K. Arkhipova, R.C. Whitehead, D.R. Boyd, C.C.R. Allen, Draft Genome
628 Sequences of Pseudomonas putida UV4 and UV4/95, Toluene Dioxygenase-Expressing
629 Producers of cis-1,2-Dihydrodiols, Genome Announc, 6 (2018).
630 [60] G. Mosqueda, J.L. Ramos, A set of genes encoding a second toluene efflux system in Pseudomonas
631 putida DOT-T1E is linked to the tod genes for toluene metabolism, J Bacteriol, 182 (2000) 937-
632 943.
633 [61] B.D. Ondov, T.J. Treangen, P. Melsted, A.B. Mallonee, N.H. Bergman, S. Koren, A.M. Phillippy,
634 Mash: fast genome and metagenome distance estimation using MinHash, Genome Biol, 17
635 (2016) 132.
636 [62] L. Pritchard, H.R. Glover, S. Humphris, J.G. Elphinstone, I.K. Toth, Genomics and taxonomy in
637 diagnostics for food security: soft-rotting enterobacterial plant pathogens, Anal. Methods, 8
638 (2016) 10-24.
639 [63] T. Seemann, Prokka: rapid prokaryotic genome annotation, Bioinformatics, 30 (2014) 2068-2069.
640 [64] A.J. Page, C.A. Cummins, M. Hunt, V.K. Wong, S. Reuter, M.T. Holden, M. Fookes, D. Falush, J.A.
641 Keane, J. Parkhill, Roary: rapid large-scale prokaryote pan genome analysis, Bioinformatics, 31
642 (2015) 3691-3693.
643 [65] L. Snipen, K.H. Liland, micropan: an R-package for microbial pan-genomics, BMC Bioinformatics,
644 16 (2015) 79.
645 [66] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements
646 in performance and usability, Mol Biol Evol, 30 (2013) 772-780.
647 [67] A.J. Page, B. Taylor, A.J. Delaney, J. Soares, T. Seemann, J.A. Keane, S.R. Harris, SNP-sites: rapid
648 efficient extraction of SNPs from multi-FASTA alignments, Microb Genom, 2 (2016) e000056.
649 [68] A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
650 phylogenies, Bioinformatics, 30 (2014) 1312-1313.

651 [69] I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v4: recent updates and new developments,
652 Nucleic Acids Res, 47 (2019) W256-W259.
653 [70] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST
654 and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res, 25
655 (1997) 3389-3402.
656 [71] R.C. Edgar, Search and clustering orders of magnitude faster than BLAST, Bioinformatics, 26 (2010)
657 2460-2461.
658
659

Tables

Table 1. *Pseudomonas alloputida* isolates used in this study with average nucleotide identity (ANI) from Kh7^T.

Strain	Classified as	ANI	CC	Location	Source	Accession
15420352	*P. putida*	0.97	CC4	China	urine	GCA_013305625.1
B2017	*P. putida*	0.97	CC3	Spain	root	GCA_007279645.1
B4	*P. putida*	0.97	CC3	China	soil	GCA_003671955.1
B6-2	*P. putida*	0.97	NOCC	-	soil	GCA_000226035.3
BIRD-1	*P. putida*	0.99	CC5	Spain	rhizosphere	GCA_000183645.1
CA-3	*P. putida*	0.97	CC4	Ireland	waste material	GCA_002810225.1
CY06	*P. monteilii*	0.96	CC1	China	shrimp	GCA_002835905.1
DOT-T1E	*P. putida*	0.97	CC3	Spain	wastewater	GCA_000281215.1
DPA1	*P. putida*	0.97	CC4	Greece	soil	GCA_002891885.1
DZ-F23	*P. putida*	0.95	CC1	China	fly	GCA_002094775.1
F1	*P. putida*	0.97	CC2	USA	soil	GCA_000016865.1
FDAARGOS_409	*P. putida*	0.97	CC4	USA	blood	GCA_002554535.1
FF4	*Pseudomonas sp.*	0.97	NOCC	Chile	wastewater	GCA_007049805.1
FW305-E2	*P. putida*	0.97	CC2	USA	groundwater	GCA_900095365.1
FW305-E2_1	*Pseudomonas sp.*	0.97	CC2	USA	groundwater	GCA_002901725.1
GTC_16473	*Pseudomonas sp.*	0.97	CC7	Japan	*Homo sapiens*	GCA_001753855.1
GTC_16482	*Pseudomonas sp.*	0.97	CC7	Japan	*Homo sapiens*	GCA_001319995.1
H	*P. putida*	0.97	CC7	Germany	soil	GCA_001077495.1
Idaho	*P. putida*	0.98	CC6	China	-	GCA_000226475.2
INSali382	*P. putida*	0.98	CC6	Portugal	vegetable	GCA_001653615.1
IOFA1	*P. putida*	0.96	CC1	Indian Ocean	sediment	GCA_001293025.1
IOFA19	*P. monteilii*	0.95	CC1	Indian Ocean	sediment	GCA_000633915.1
JB	*P. putida*	0.99	CC5	Czech Republic	soil	GCA_001767335.1
JLR11	*P. putida*	0.97	CC7	Spain	wastewater	GCA_001183585.1
JR16	*P. putida*	0.99	CC5	India	soil	GCA_004519745.1
JY-Q	*Pseudomonas sp.*	0.97	CC4	China	tabaco extract	GCA_001655295.1
KB3	*P. putida*	0.96	CC1	Poland	soil	GCA_004614175.1
KBS0802	*Pseudomonas sp.*	0.97	CC7	USA	soil	GCA_005937845.2
KCJK7911	*P. putida*	0.97	CC3	USA	water	GCA_003053335.1
KCJK7916	*P. putida*	0.97	NOCC	USA	water	GCA_003053385.1
KH-18-2	*P. putida*	0.96	CC1	Pacific Ocean	water	GCA_002906815.1
KH-20-11	*P. putida*	0.95	CC1	Pacific Ocean	water	GCA_002906795.1
Kh14	*Pseudomonas sp.*	1.00	CC5	Iran	rhizosphere	GCA_900291005.1
Kh7	*Pseudomonas sp.*	1.00	CC5	Iran	rhizosphere	GCA_900291035.1
KT-90	*P. putida*	0.96	CC1	Pacific Ocean	coastal water	GCA_002906755.1
KT2440	*P. putida*	0.97	CC7	Japan	rhizosphere	GCA_900167985.1
LD6	*P. putida*	1.00	CC5	China	rhizosphere	GCA_003586135.1
LF54	*P. putida*	0.96	CC1	Japan	lotus field	GCA_000390005.2
LS46	*P. putida*	0.97	CC4	Canada	water	GCA_000294445.2
N1R	*P. putida*	0.99	CC5	USA	soil	GCA_900156185.1
NBRC_111118	*Pseudomonas sp.*	0.97	CC4	Japan	*Homo sapiens*	GCA_001320085.1
NBRC_111121	*Pseudomonas sp.*	0.97	CC7	Japan	sputum	GCA_001320165.1

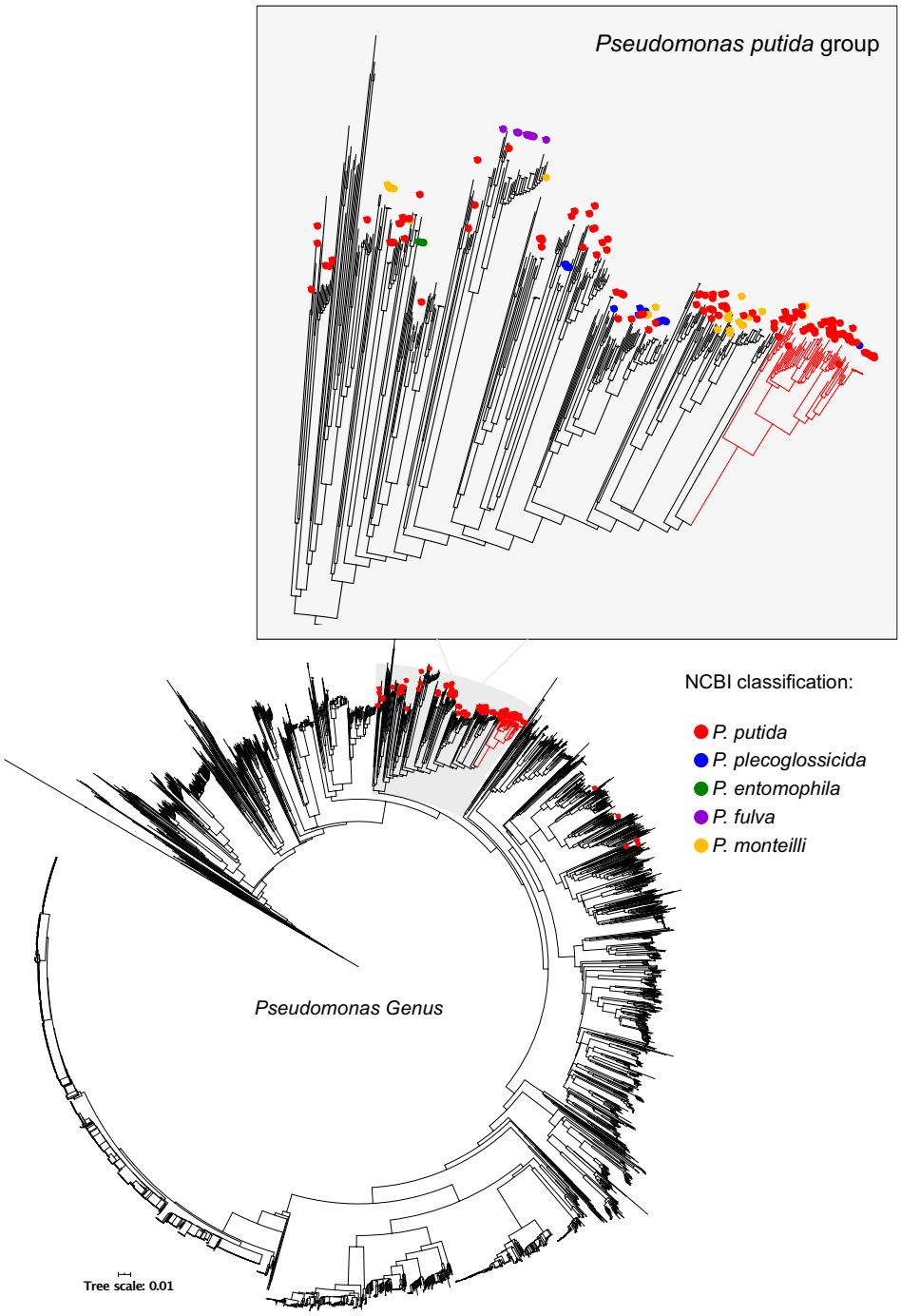
NBRC_111125	*Pseudomonas sp.*	0.97	CC4	Japan	urine	GCA_001320295.1
NBRC_111136	*Pseudomonas sp.*	0.97	CC4	Japan	urine	GCA_001320745.1
NBRC_111139	*Pseudomonas sp.*	0.97	CC4	Japan	eye discharge	GCA_001753955.1
NCTC13185	*P. putida*	0.97	CC7	-	-	GCA_901482375.1
NCTC13186	*P. putida*	0.97	CC7	-	-	GCA_900636645.1
ND6	*P. putida*	0.97	CC2	China	wastewater	GCA_000264665.2
ODNR4SY	*P. putida*	0.97	CC2	USA	water	GCA_009905395.1
OR45a	*P. putida*	0.99	CC5	Poland	activated sludge	GCA_004614155.1
P11	*P. hunanensis*	0.97	CC4	China	high-arsenic soil	GCA_002910975.1
PaW85-2019	*P. putida*	0.97	CC7	Estonia	-	GCA_011750655.1
PaW85-d13TA	*P. putida*	0.97	CC7	Estonia	-	GCA_011750675.1
PCL1760	*P. putida*	1.00	CC5	Spain	rhizosphere	GCA_001282125.1
PD1	*P. putida*	0.97	CC3	USA	root	GCA_000799625.1
RW4053	*Pseudomonas sp.*	0.99	CC5	Germany	river sediments	GCA_003184135.1
S12	*P. putida*	1.00	CC5	Netherlands	soil	GCA_000495455.2
SJTE-1	*P. putida*	0.97	CC2	China	soil	GCA_000271965.2
SMT-1	*Pseudomonas sp.*	0.97	CC7	China	soil	GCA_003204195.1
SWI36	*Pseudomonas sp.*	0.97	CC6	USA	soil	GCA_002948105.1
SWI36_1	*Pseudomonas sp.*	0.98	CC6	USA	soil	GCA_004153505.1
SWI36_2	*Pseudomonas sp.*	0.98	CC6	USA	soil	GCA_004153435.1
TRO1	*P. putida*	0.97	CC4	Denmark	activated sludge	GCA_000367825.1
UV4	*P. putida*	0.97	CC4	UK	laboratory strain	GCA_002165695.1
UV4_95	*P. putida*	0.97	CC4	UK	laboratory strain	GCA_002165665.1
XWY-1	*Pseudomonas sp.*	0.97	CC6	China	rice fields	GCA_002953115.1
YKD221	*P. putida*	0.97	CC2	Japan	soil	GCA_000787655.1
ZKA3	*P. plecoglossicida*	0.98	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%)

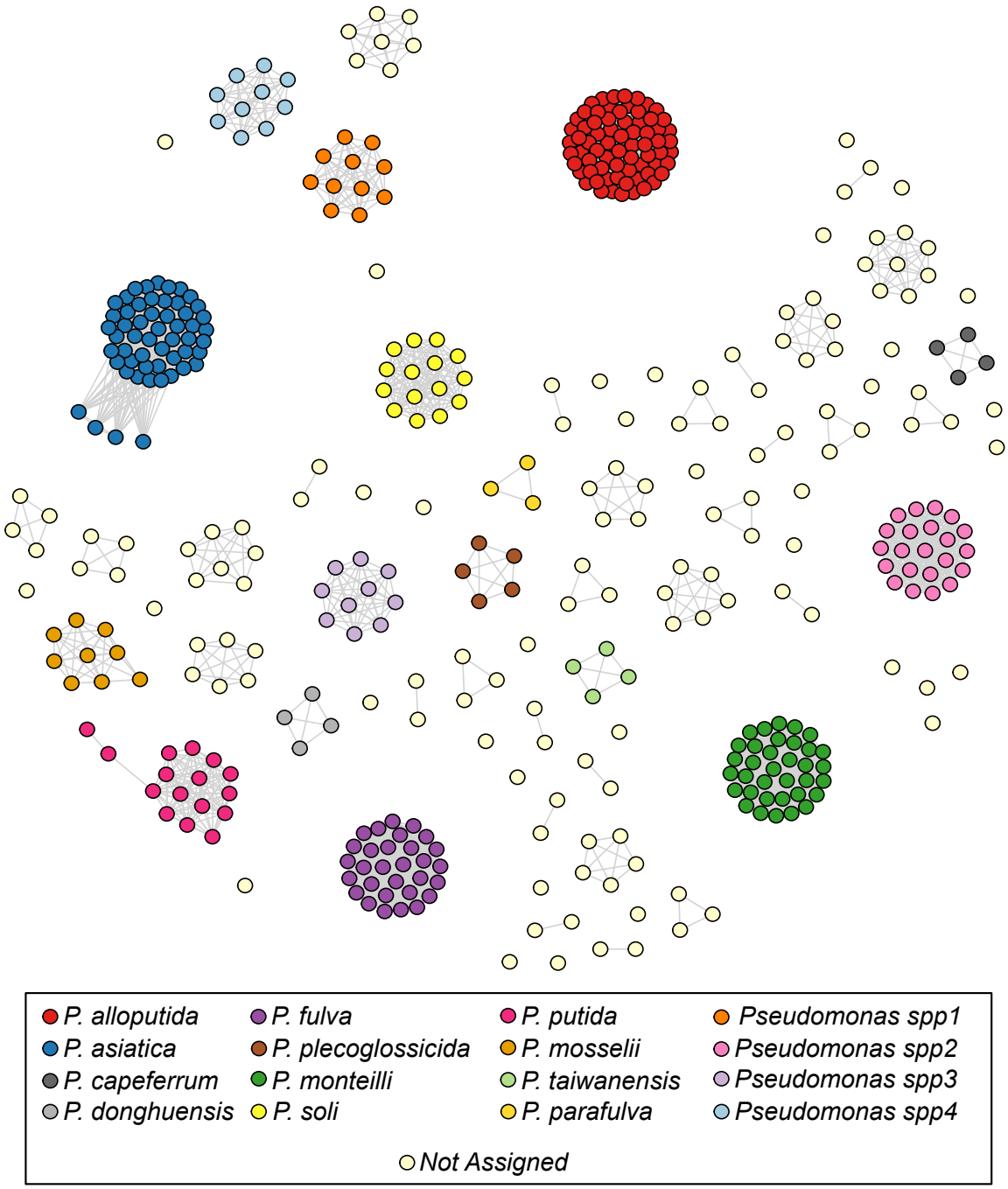
*Number of different genes.

667 **Figures**



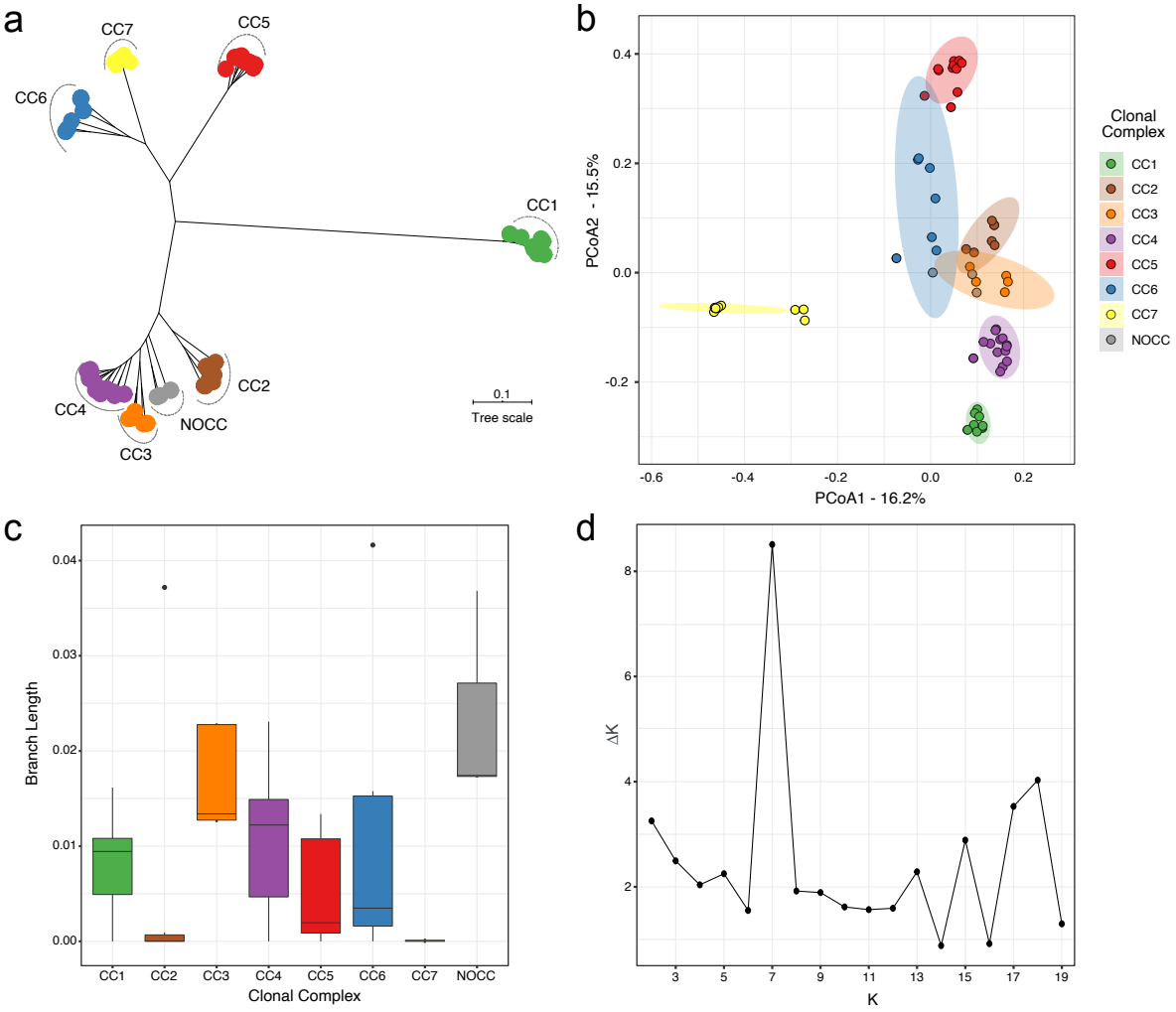
668

669 **Figure 1.** Distance tree of 10,457 *Pseudomonas* genomes from *Pseudomonas* genus.
670 Genomes classified as *P. putida* according to NCBI are marked as red circle. *P. putida*
671 group was highlighted to assess the distribution of misclassified genomes. *P. alloputida*
672 branches are colored in red.



673

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

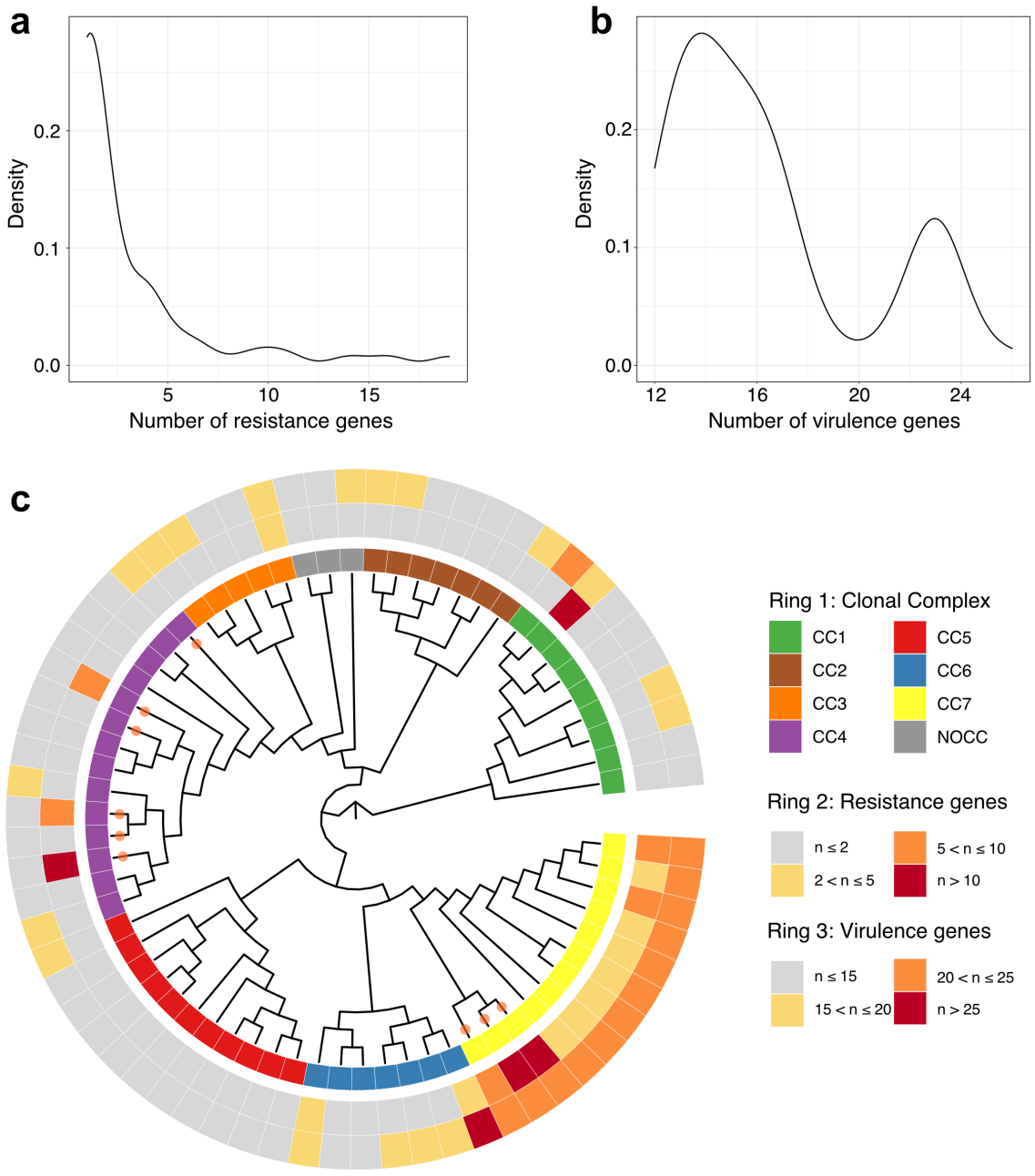


678

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

686

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



691

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

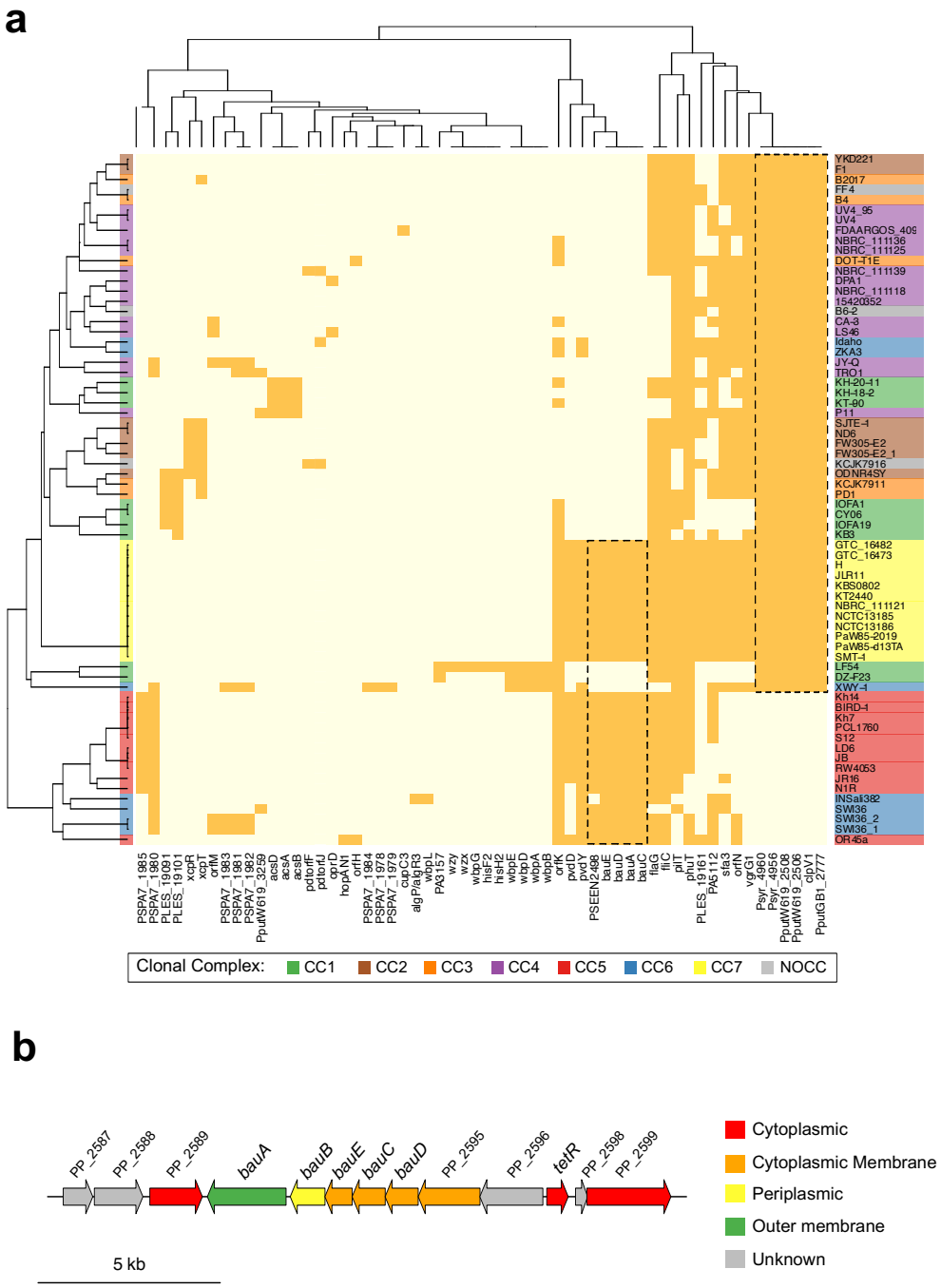


Figure 6. Acquired virulome composition. **a.** Matrix with presence (dark squares) and absence (light squares) profiles of virulence genes. Rows represent strains colored based on Clonal Complex that they belong. Columns are virulence genes identified. **b.** *P. alloputida* KT2440 *bauABCDE* genetic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