FEDERAL UNIVERSITY OF MINAS GERAIS
INSTITUTE OF BIOLOGICAL SCIENCES
INTERUNIT POST-GRADUATE PROGRAM IN BIOINFORMATICS

HEMANOEL PASSARELLI ARAUJO

UNVEILING THE ECOLOGICAL ROLE OF BACTERIAL GENETIC DISCONTINUITY

Belo Horizonte - MG
September, 2023

HEMANOEL PASSARELLI ARAUJO

UNVEILING THE ECOLOGICAL ROLE OF BACTERIAL GENETIC DISCONTINUITY

Thesis presented as a partial requirement for obtaining a Ph. D. degree in Bioinformatics by the Interunit Post-Graduate Program in Bioinformatics at the Institute of Biological Sciences at Federal University of Minas Gerais

Supervisor: Dr. Thiago Motta Venancio
Co-supervisor: Glória Regina Franco

Belo Horizonte
September, 2023

043 Araujo, Hemanoel Passarelli.
Unveiling the ecological role of bacterial genetic discontinuity [manuscrito] / Hemanoel Passarelli Araujo. – 2023.
70 f. : il. ; 29,5 cm.

Orientador: Dr. Thiago Motta Venancio. Coorientadora: Glória Regina Franco.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Genoma Bacteriano. 3. Variação Genética. 4. Taxonomia. 5. Pseudomonas. I. Venâncio, Thiago Motta. II. Franco, Glória Regina. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

ATA DE DEFESA DE TESE

HEMANOEL PASSARELLI ARAUJO

Às quatorze horas do dia **26 de setembro de 2023**, reuniu-se, no aplicativo Zoom, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Unveiling the ecological role of bacterial genetic discontinuity"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Thiago Motta Venancio**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Após a apresentação da tese, seguida de arguição e discussão, foram atribuídas as seguintes indicações:

Professor(a)/Pesquisador(a)	Instituição	Indicação
Dr. Thiago Motta Venancio	UENF	**Aprovado**
Dra. Glória Regina Franco	UFMG	**Aprovado**
Dr. Aristóteles Góes Neto	UFMG	**Aprovado**
Dr. Diogo Antonio Tschoeke	UFRJ	**Aprovado**
Dr. Fabrício Rodrigues dos Santos	UFMG	**Aprovado**
Dr. Luiz Eduardo Vieira Del Bem	UFMG	**Aprovado**
Dr. Robson Francisco de Souza	USP	**Aprovado**

Resultado Final: Aprovado com louvor

A banca examinadora deliberou, por unanimidade, pela aprovação da tese com distinção, conferindo-lhe a menção de "**Aprovado com Louvor**". Os membros da banca ressaltaram a relevância e a contribuição significativa do trabalho, bem como a qualidade excepcional da apresentação e da argumentação.

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 26 de setembro de 2023.

Documento assinado eletronicamente por **Robson Francisco de Souza**, **Usuário Externo**, em 28/09/2023, às 16:13, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Diogo Antonio Tschoeke**, **Usuário Externo**, em 29/09/2023, às 07:33, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Thiago Venancio**, **Usuário Externo**, em 29/09/2023, às 08:07, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Aristoteles Goes Neto**, **Professor do Magistério Superior**, em 29/09/2023, às 10:00, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Gloria Regina Franco**, **Professora do Magistério Superior**, em 29/09/2023, às 10:07, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fabricio Rodrigues dos Santos**, **Membro de comissão**, em 29/09/2023, às 10:18, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Luiz Eduardo Vieira Del Bem**, **Professor do Magistério Superior**, em 02/10/2023, às 14:49, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2656664** e o código CRC **77A204A8**.

Referência: Processo nº 23072.258491/2023-32 SEI nº 2656664

Acknowledgement

Aos meus pais, Adelson e Erilza, agradeço profundamente pelo apoio incondicional ao longo de minha trajetória acadêmica. Mãe, o seu incansável esforço não apenas moldou meu futuro, mas também me inspirou a minha seguir meus sonhos. Aos meus avós, Lúcia e Aleir, e aos meus irmãos, Graziela e Hisrael, minha gratidão por estarem sempre presentes em minha vida.

Em especial, não consigo expressar o quão grato eu sou por ter tido o privilégio de compartilhar até o meu primeiro respirar com você, Hisrael. Você é muito mais do que meu irmão gêmeo, é o meu melhor amigo da vida. Agradeço também a você, Matheus, por tornar este período do doutorado mais leve. Fico feliz em ter tido a sua amizade ao longo de todos estes anos. Por fim, quero te agradecer, Alyson, pelo companheirismo nestas etapas finais.

Durante meu doutorado, tive o privilégio de colaborar com diversos grupos de pesquisa, e cada pessoa contribuiu significativamente para minha formação como pesquisador. À equipe da UENF, agradeço pela parceria que começou durante minha graduação. Aos "Gloriosos", sou grato pela amizade e pelo excelente ambiente de trabalho que compartilhamos. Ao grupo de Harvard, agradeço pelos momentos inesquecíveis e ideias enriquecedoras. Vocês todos se tornaram minha segunda família ao longo deste doutorado.

Ao meu orientador, Thiago, expresso minha profunda gratidão. Desde o momento em que me apresentou a bioinformática, você foi mais do que um orientador; foi um amigo constante com quem pude compartilhar todos os meus planos de forma leve. Sou sortudo por ter tido você ao longo de todos esses anos da minha trajetória acadêmica. Também agradeço à Glória pelo carinho e o cuidado especial durante cada segundo meu na UFMG. Ao Bill, que supervisionou meu período em Harvard, sou grato por ter tido a oportunidade de me inspirar na sua forma de fazer ciência.

Por fim, sou imensamente grato à Fulbright Brasil pelo financiamento que tornou possível minha estadia na Universidade de Harvard (2022-2023). Além disso, este trabalho recebeu apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Resumo

Descontinuidade genética bacteriana refere-se a transições bruscas na identidade genômica entre espécies e é um conceito fundamental na elucidação da diversidade e evolução microbiana. O presente trabalho explora como a diversidade genética pode ser observada como grupos discretos dentro e entre espécies. A investigação começa ao nível de espécie, utilizando *Pseudomonas alloputida* como modelo – uma espécie com relevância biotecnológica e clínica. A análise da estrutura populacional desta espécie revela a existência de pelo menos sete complexos clonais, destacando a diversidade genética e oferecendo novas perspectivas sobre o seu potencial biotecnológico. Como próximo passo deste estudo, o gênero diverso de *Pseudomonas* é investigado, elucidando problemas de classificação taxonômica a partir da análise de redes de dados genômicos. Este trabalho também revelou o quão estruturadas e estáveis são as redes sob limiares próximos ao utilizado para delimitar espécies. Finalmente, o conceito de descontinuidade genética bacteriana é examinado numa escala mais ampla, abrangendo diversas espécies bacterianas. A partir de um conjunto de mais de 210.000 genomas, quebras nas distribuições de identidade genômica são reveladas, indicando a existência de fronteiras genéticas dentro das populações. O impacto de métricas obtidas a partir do pangenoma para inferir descontinuidade genética também fornece informações sobre a relevância ecológica deste fenômeno. Em conclusão, esta tese explora a descontinuidade genética bacteriana a partir de uma perspectiva holística, fornecendo bases para entendermos quais são as implicações ecológicas e taxonômicas de tais quebras. Ela também destaca a necessidade de reavaliarmos as classificações tradicionais de espécies em um era onde os dados genômicos são abundantes.

Palavras-chave: Conceito de espécies bacterianas, diversidade microbiana, taxonomia, pangenoma, redes.

Abstract

Bacterial genetic discontinuity is characterized by sharp transitions in genomic identity among species. It stands as a cornerstone concept in elucidating microbial diversity and evolution. The present work explores how genetic diversity can be observed as discrete groups within and between species. The investigation commences at a species level, using *Pseudomonas alloputida* as a model – a species with biotechnological and clinical significance. The population structure analysis unveils at least seven clonal complexes, highlighting the genetic diversity within this species and offering insights into its biotechnological potential. Expanding the scope, this study delves into the diverse *Pseudomonas* genus, challenging traditional species classifications using network analyses of genomic data. Taxonomic inconsistencies and the existence of distinct *Pseudomonas* groups question the current taxonomic framework. This work also revealed how structured and stable are the networks under thresholds close to those used to delimit species. Finally, the concept of bacterial genetic discontinuity is examined at a broader scale, encompassing diverse bacterial species. By harnessing a dataset comprising over 210,000 genomes, clear breakpoints in genomic identity distributions are revealed, shedding light on the existence of genetic boundaries within populations. The impact of pangenome features on estimating genetic discontinuity provides insights into the ecological relevance of this phenomenon. In conclusion, this thesis explores the intricate landscape of bacterial genetic discontinuity, offering a holistic perspective on its ecological and taxonomic implications. It also highlights the need for reevaluating traditional species classifications in the Genomics Era.

Keywords: Bacterial species concept, microbial diversity, taxonomy, pangenome, networks.

Table of Contents

1 Research Background *9*

1.1 Bacterial Diversity and Evolution 9

1.2 The Bacterial Species Concept 10

1.3 Network Analysis Using Genomic Data 10

1.4 Bacterial Genetic Discontinuity 11

2 Research Aim *12*

2.1 Specific aims 12

3 Resulting Articles *13*

3.1 Phylogenetic analysis and population structure of *Pseudomonas alloputida* 14

3.2 Network analysis of ten thousand genomes shed light on *Pseudomonas* diversity and classification 27

3.3 Unveiling bacterial genetic discontinuity across different species provide insights into genetic and ecological diversity 35

4 Integrative Discussion *55*

5 Research Perspectives and Conclusion *57*

6 References *58*

1 Research Background

1.1 Bacterial Diversity and Evolution

From the earliest bacterial genome-scale comparisons, it became clear that the genomes of numerous species do not adhere strictly to vertical descent[1]. In bacteria, substantial variations in gene content can exist among genomes from the same species, with only a portion of genes present across all genomes[2].

Bacterial evolution is driven by various mechanisms, with horizontal gene transfer (HGT) being a key contributor. HGT allows bacteria to acquire genes from unrelated organisms that enables rapid adaptation, including responses to challenges like antibiotics. This dynamic process, facilitated by mechanisms like transformation, conjugation, and transduction, can blur species boundaries and shape bacterial diversity and adaptability[3].

The pangenome concept has emerged as a powerful tool for deciphering bacterial diversity and evolution[4]. A pangenome refers to the set of non-redundant genes in a given species[2]. It comprises the core genome (genes shared by almost all individuals), the accessory genome (variable genes present in a group of genomes), and singletons (unique genes found in only one or a few individuals).

Pangenome metrics, such as openness, provide a quantitative measure of the HGT dynamics within bacterial populations. Pangenome openness refers to how the pangenome size changes as more genomes are sequenced within a species or group. In bacterial ecology, this metric is particularly insightful as it reflects the adaptability of bacterial communities to diverse environments. For instance, niche specialists such as symbionts that are more likely to exist in stable environments with very low diversity have more closed pangenomes[4].

The integration of genomics to understand bacterial diversity and evolution has far-reaching implications in unraveling the complexities of prokaryotes. Within this genomic landscape, bacteria are now recognized as highly dynamic entities[3]. However, this fluidity poses a challenge to the conventional concept of species, as genetic boundaries blur, and the species definition becomes less clear.

1.2 The Bacterial Species Concept

The notion of a species is a fundamental concept in biology. However, its application to bacteria has been a longstanding and often an inconclusive debate, further complicated by the prevalence of HGT as an essential source of evolutionary innovation for bacteria[5]. Defining bacterial species extends beyond a human need for categorization; it holds implications for medicine, industry, and diversity studies[6].

Advancements in bioinformatics and genomics have shed light on questions about whether bacteria and other microbes are characterized by discrete clusters (species) or a genetic continuum due to frequent horizontal gene transfer[7]. Some studies based on a limited number of closely related genomes suggest a genetic continuum[8]. Conversely, others argue that horizontal gene transfer may not be frequent enough to blur species boundaries, and closely-related organisms exchange DNA more frequently, maintaining distinct clusters[9].

The idea of reverse ecology employs horizontal gene transfer (HGT) patterns to define bacterial species[10]. By assessing the frequency of genetic exchange between bacterial strains, researchers can identify groups more likely to belong to the same species. This approach offers a dynamic and ecologically informed perspective on microbial taxonomy, reflecting the complex genetic exchange in natural environments.

Recent integration of Next-Generation Sequencing (NGS) and bioinformatics tools has enhanced the resolution of genome comparisons. Massive pairwise genomic comparisons have revealed clear breaks in bacterial genetic distribution[7, 11, 12], with whole-genome average nucleotide identity (ANI) emerging as a robust method to define species. Typically, organisms within the same species exhibit ≥95% ANI among themselves, offering a valuable tool for species delineation[13].

1.3 Network Analysis Using Genomic Data

Biological networks have been an essential analytical tool to better understand microbial diversity and ecology[14, 15]. A network is a mathematical representation of interconnected nodes and edges, where nodes represent entities (e.g., genes,

individuals, or species) and edges denote relationships between them. In the context of microbial genomics, network structures can be harnessed to infer communities as genomic species[16].

Consider a network where nodes represent genomes and edges the genomic identity between them. One of the key advantages of network analysis lies in its flexibility: different thresholds for defining species can lead to distinct network structures that can be measured. For example, our recent study employing ten thousand genomes of *Pseudomonas* highlighted how the network structure tends to be stable around 95% identity and how other network metrics can be employed to enhance our knowledge on bacterial diversity[16].

1.4 Bacterial Genetic Discontinuity

This thesis centers on bacterial genetic discontinuity, challenging the traditional view of species as genetic mosaics. Genetic discontinuity represents abrupt transitions in genomic identity among bacterial populations. While genomic sequencing has empowered researchers to track and characterize genetic discontinuity patterns systematically[7, 12, 17], three questions remain: (i) does bacterial diversity exist as a continuum or as discrete species groups? (ii) how can we measure genetic discontinuity; (iii) what are the ecological implications of this phenomenon?

By shedding light on the existence of clear breakpoints in genomic identity distributions, this research aims to quantify and uncover the ecological relevance of genetic discontinuity. The concept of genetic discontinuity offers a paradigm shift in understanding microbial diversity, challenging traditional species boundaries and providing insights into the dynamic relationships that shape the microbial world.

2 Research Aim

This research aims to investigate bacterial genetic discontinuity and its ecological implications across different taxonomic levels, from species-specific patterns to the broader context of bacterial diversity.

2.1 Specific aims

1- Investigate the genetic discontinuity patterns within *Pseudomonas putida* species complex;
2- Characterize the population structure and genetic diversity of *Pseudomonas alloputida*;
3- Examine how pathogenic and bioremediation traits relate with intra-species genetic groups (clonal complexes);
4- Assess the genetic boundaries across various *Pseudomonas* species through identity network analyses;
5- Quantify genetic discontinuity patterns in a large dataset comprising diverse bacterial species;
6- Examine the ecological implications of genetic discontinuity across different bacterial lifestyles;
7- Identify key features that may influence genetic discontinuity predictions.

3 Resulting Articles

This thesis is structured around three articles, each with a unique focus on bacterial genetic discontinuity. The first article delves into patterns of genetic discontinuity at the species level, utilizing *Pseudomonas alloputida* as a model organism. The second article broadens the scope to explore the diversity of the *Pseudomonas* genus, addressing taxonomic inconsistencies and proposing a more accurate representation of its genetic diversity. In the final article, we transcend taxonomic boundaries to investigate whether the observed patterns of genetic discontinuity are consistent across diverse bacterial species. We also quantify and inspect about the ecological role of genetic discontinuity in bacterial species.

3.1 Phylogenetic analysis and population structure of *Pseudomonas alloputida*

Pseudomonas is a bacterial genus housing over 250 characterized and validated species. It is organized into three primary phylogenetic lineages based on genetic markers like 16S rRNA and essential housekeeping genes: *Pseudomonas aeruginosa*, *Pseudomonas pertucinogena*, and *Pseudomonas fluorescens*. The *P. fluorescens* lineage, encompasses six distinctive phylogenetic groups, one of which is represented by *P. putida*.

The *P. putida* group encompasses various species, with *P. putida* sensu stricto serving as the group's representative name. These species thrive in diverse ecological niches, often inhabiting soil and water environments. They are recognized for their versatile functionalities, including promoting plant growth, bioremediating environmental pollutants, and defending against plant pathogens.

The first article in this thesis addresses the genetic makeup and population structure of *Pseudomonas alloputida*. By using an identity network, we observed a very structured network with *P. alloputida* clearly detectable as a community. This study unveils the existence of at least seven clonal complexes within *P. alloputida*, with clinical isolates predominantly found in CC4. Moreover, the article examines the presence of resistance genes in plasmids and assesses virulence profiles, shedding light on the pathogenic potential of *P. alloputida* strains. Additionally, we also explored the role of horizontal gene transfer in shaping the ability of this species in bioremediating aromatic compounds is explored, offering insights into its biotechnological potential.

Genomics 113 (2021) 3762–3773

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Phylogenetic analysis and population structure of *Pseudomonas alloputida*

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

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

ARTICLE INFO

Keywords:
Pseudomonas putida group
Pseudomonads
cgMLST

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 its potential for biotechnological applications.

1. Introduction

Pseudomonas is a diverse and complex bacterial genus that contains more than 250 species characterized and validated [1]. The genus is further divided into three phylogenetic lineages (*Pseudomonas aeruginosa*, *Pseudomonas pertucinogena*, and *Pseudomonas fluorescens*) based on 16S rRNA and other housekeeping genes [2]. The *P. fluorescens* lineage contains six phylogenetic groups; one of them is represented by *P. putida*.

The *P. 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 polluted soils and plant roots [3,4]. *P. putida* species are well known to perform many functions such as plant growth promotion, bioremediation, and protection against plant pathogens [5].

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 members of a novel species, *Pseudomonas alloputida,* whose type strain is Kh7 (=CFBP 8484^T =LMG 29756^T) [6]. Here, we report the population structure of *P. alloputida*, which was used to estimate the diversity and prevalence of resistance and virulence genes. Further, we used the inferred population structure to better understand the distribution of bioremediation and plant growth promotion genes and to assess the biotechnological potential of 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 [7]. We computed the pairwise distances between each isolate using mashtree [8] to compute the distance tree of the genus, which is highly diverse (Fig. 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% (Fig. S1a); this clade corresponds to the *P. putida* group (Fig. 1) and comprises other species such as *P. plecoglossicida*, *P. monteilli,* and *P. fulva*.

ANI analysis provides a raw estimate of bacterial species [9]. 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 [9,10]. 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* (Fig. S1b). The isolate previously classified as *P. monteilli* IOFA19 (GCA_000633915.1)

* Corresponding authors at: Av. Alberto Lamego 2000, P5 sala 217, Parque Califórnia, Campos dos Goytacazes CEP: 28013-602, RJ, Brazil.
E-mail addresses: hemanuel.passarelli@gmail.com (H. Passarelli-Araujo), thiago.venancio@gmail.com (T.M. Venancio).

https://doi.org/10.1016/j.ygeno.2021.09.008
Received 22 March 2021; Received in revised form 16 July 2021; Accepted 11 September 2021
Available online 13 September 2021

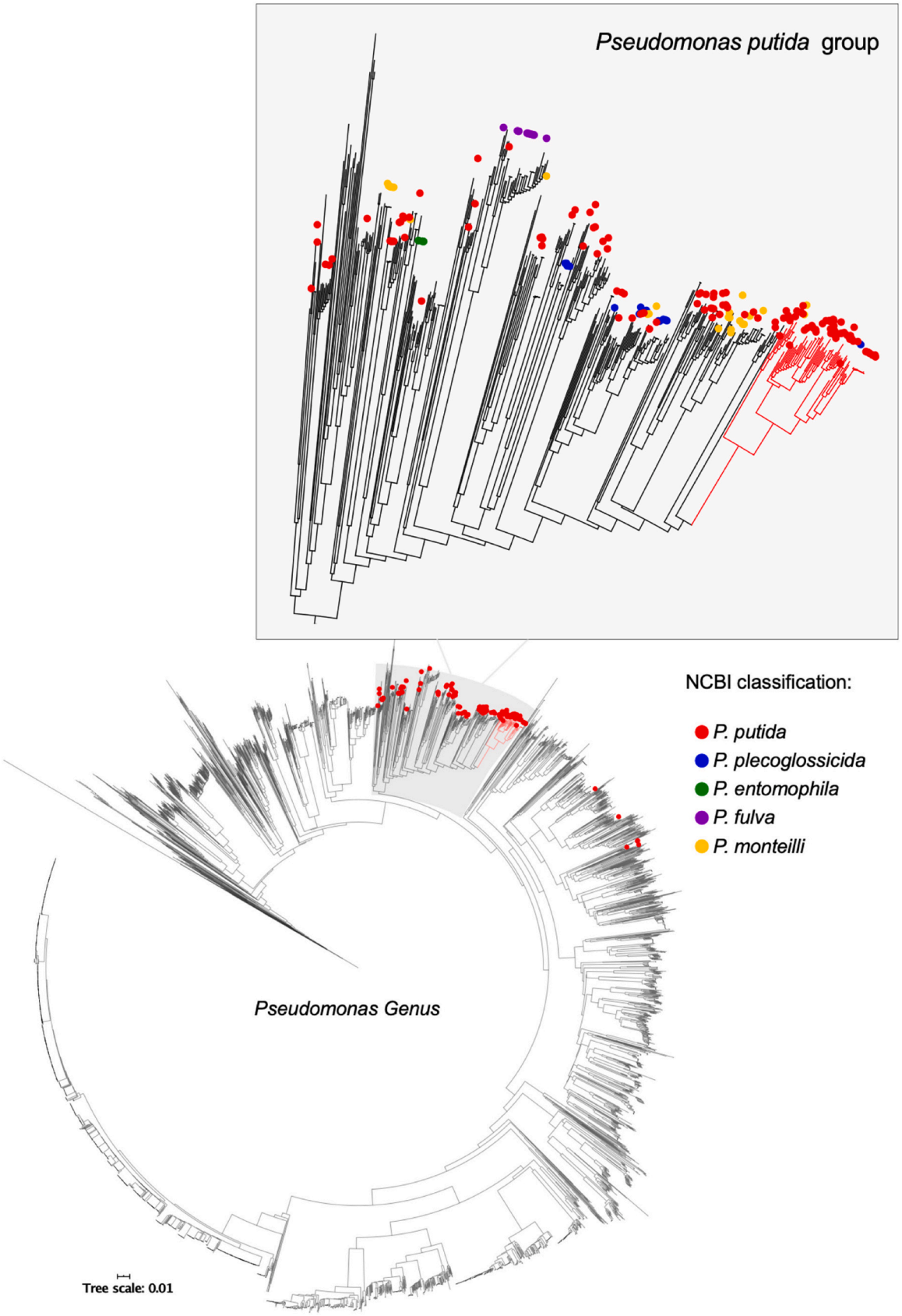

Fig. 1. Distance tree of 10,457 *Pseudomonas* genomes. Genomes classified as *P. putida* according to NCBI are marked as red circle. *P. putida* group was highlighted to assess the distribution of misclassified genomes. *P. alloputida* branches are colored in red.

had the lowest ANI value within the predicted species (95.48%), followed by a drop to 91.13% (Fig. 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 (Fig. 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 may correspond to new species (Fig. 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) (Fig. 2). *P. putida* and *P. alloputida* comprise groups with 16 and 68 genomes, respectively (Fig. 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 [11], KT2440, a *P. alloputida* isolate [6], has also been used to categorize *P. putida* genomes over the years [4,11,12].

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) [13], to oxidize manganese (GB-1) [14], and to damage human tissues (HB3267) [15], 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 [16]. HB3267, as well as DLL-E4, SF1, and S11, were proposed as members of a new species, *Pseudomonas shirazica* [6]. 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* [17]. 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 [18], 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

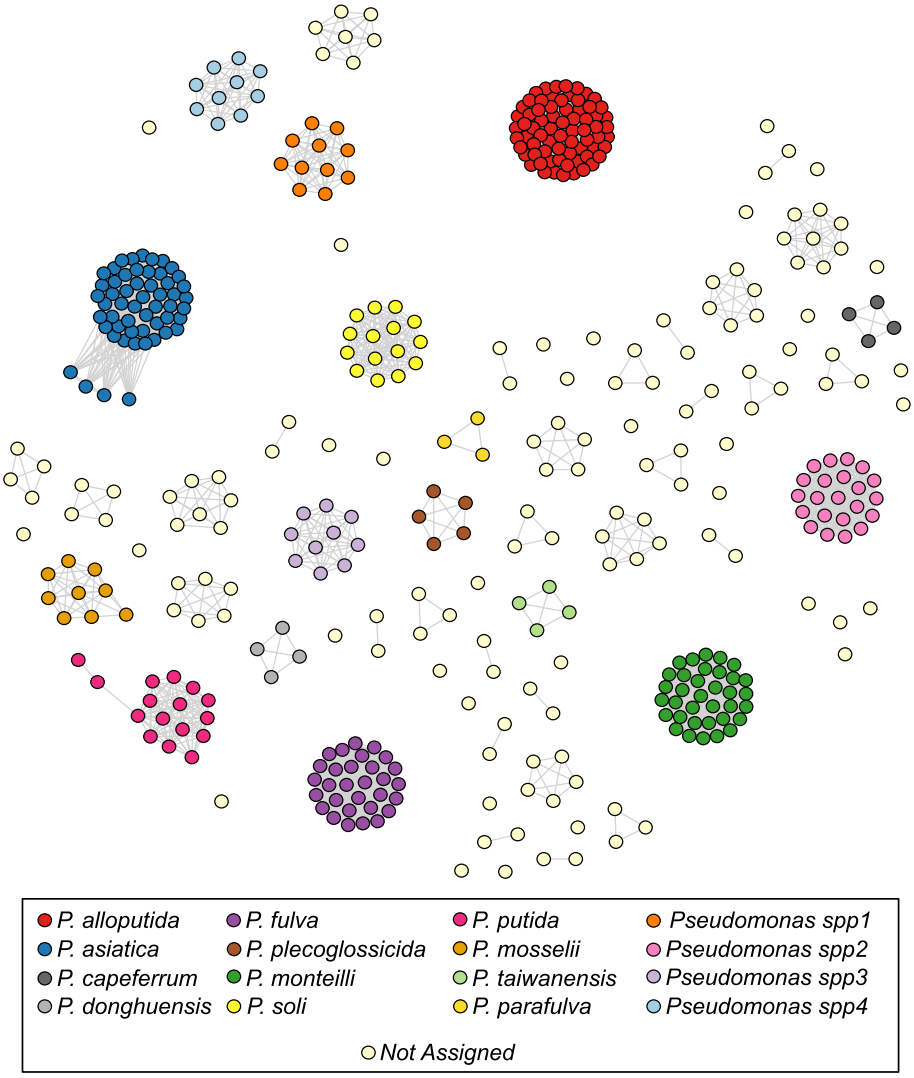

Fig. 2. Network analysis of isolates from the *P. putida* group. Nodes represent isolates and edges connect isolates with at least 95% of average nucleotide identity. Clusters with type strains or at least ten genomes were highlighted, as they represent either known or potentially novel species.

P. alloputida pangenome comprises 25,782 gene families, of which 3803 (14.75%) are present in at least 95% of the isolates. By analyzing the slope of the curve ($\alpha = 0.417$), we inferred that *P. alloputida* has an open pangenome [18] (Fig. 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 [19]. 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 6373 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, the reference genome DOT-T1E has 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 [20]. 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 [21]. In an MLST analysis, each combination of SNPs defines a Sequence Type (ST) that can be linked to form Clonal Complexes (CC) [21]. A variation of classical MLST is the core genome MLST (cgMLST), which provides greater resolution by using SNPs from the entire core genome [22]. Here, we used 225,009 SNPs obtained from the *P. alloputida* core genome to reconstruct the phylogenetic tree and the cgMLST profile.

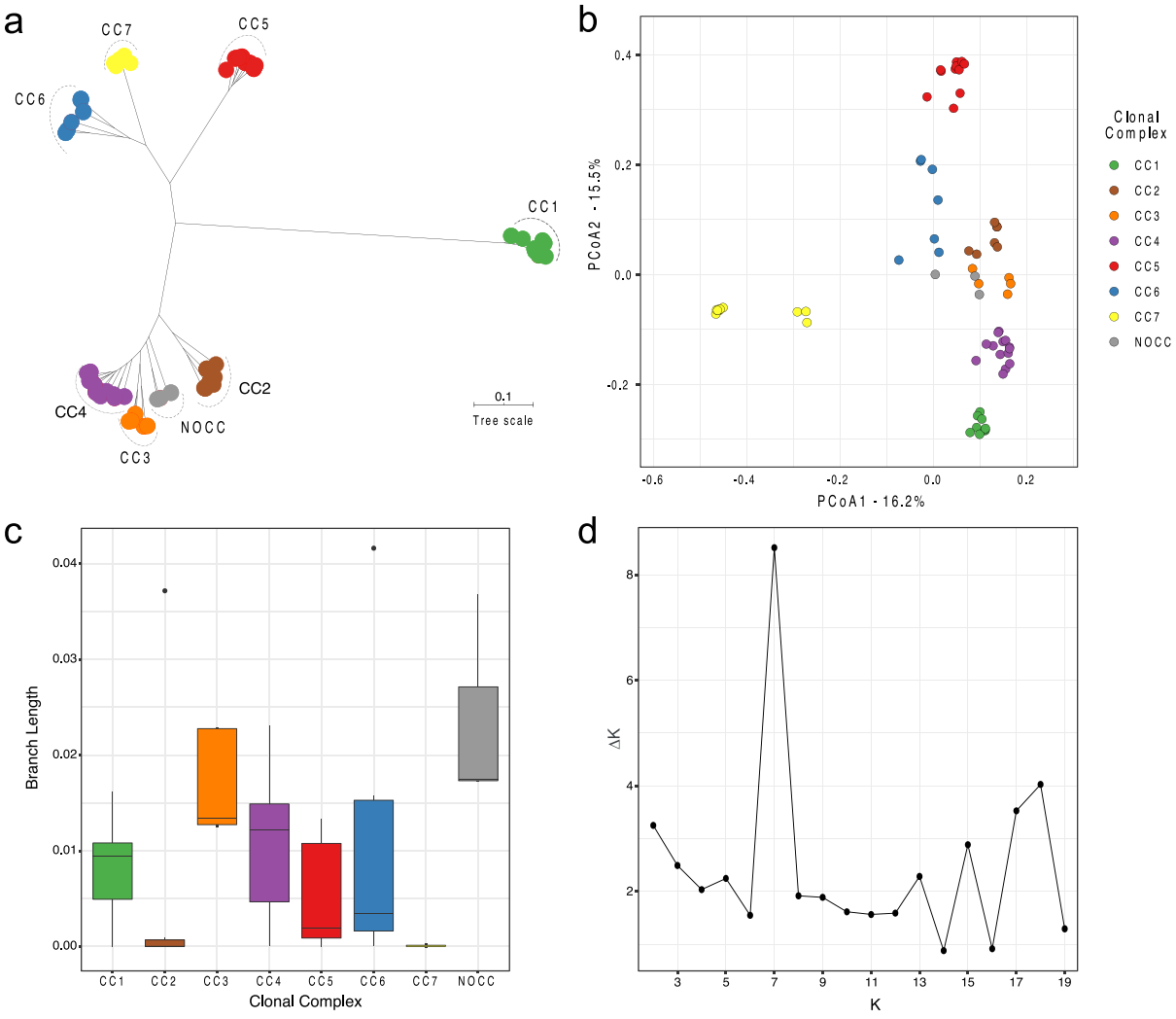

Fig. 3. Population structure of *P. alloputida*. a. phylogenetic reconstruction using SNPs extracted from core genome to assign the cgMLST scheme. Colors represent distinct Clonal Complexes and NOCC stands for No Clonal Complex assigned with high confidence. b. Principal Coordinate Analysis based on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

The cgMLST tree unveiled 7 CCs (Fig. 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 (Fig. 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 (Fig. 3b).

We also performed a Discriminant Analysis of Principal Components (DAPC) [23] to recover the genes that contribute most to separate the population based on their presence/absence profiles (Fig. 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 (Fig. S5). Next, we used branch lengths from cgMLST tree as an indirect estimator of diversity within the CCs (Fig. 3c). CC7 is the most clonal group, comprising 9 isolates, including KT2440 and three clinical isolates (GTC_16482, GTC_16473, and NBRC_111121).

The MLST scheme for *P. putida* species comprises eight housekeeping genes (*argS, gyrB, ileS, nuoC, ppsA, recA, rpoB,* and *rpoD*) [24]. We assigned isolates to STs and grouped those into CCs (Table S2). This analysis revealed some incongruences with previous reports [24]. For example, by using allele combinations with perfect-match to predict STs, KT2440 was detected as ST69 and not as ST58 [24], indicating that a revision in the public database is warranted. The predicted number of CCs was also supported by the admixture model from STRUCTURE [25]. This model assumes that each isolate has ancestry from one or more *K* genetically different sources, which we referred to as CCs. The number of CCs corresponds to the estimated number of clusters represented by parameter *K*. Instead of using the highest raw marginal likelihood, we followed the protocol for estimating the best *K* value suggested by Evanno et al. [26]. The ad hoc statistics ΔK indicates that, according to our dataset, the population structure of *P. alloputida* is composed of at least 7 CCs (Fig. 3d).

Once data from eight *loci* may be insufficient to accurately describe the population structure of *P. alloputida*, the populations identified by STRUCTURE were only considered if they matched cgMLST results. When correlating the cgMLST tree topology with ancestry proportion predicted for each isolate, we observed a clear delimitation of genetic blocks for each CC (Fig. 4). However, there are a few inconsistencies. For example, isolate B4 has a greater ancestry proportion with CC2, although cgMLST indicates its greater proximity to CC3. In addition, KCJK7916, FF4, 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* (Fig. 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) [4], 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 [27]. 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* [28,29]. However, these RND efflux pumps may play an alternative role in *P. alloputida* by pumping out toxic substances such as toluene [30]. We also identified *cpxR*, which encodes a protein that promotes MexAB-OprM expression in the absence of the MexR repressor in *P. aeruginosa* [31], 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 [32,33].

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 (Fig. 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 (Fig. S6), although CC7 has more acquired resistance genes than other CCs (Fig. 5c, Fig. 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 (Fig. 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) [34]. *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 [35] to assess the *P. alloputida* virulence profiles. 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, and plant cell wall-degrading enzymes.

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 (Fig. 5b), a pattern that has been previously observed for *Klebsiella aerogenes* [36]. Genes from the acinetobactin gene cluster and HSI-I type VI secretion system were differentially distributed across CCs (Fig. 6a), although it remains unclear whether these patterns emerged mainly from gene gain or loss. In *A. baumannii*, iron uptake is

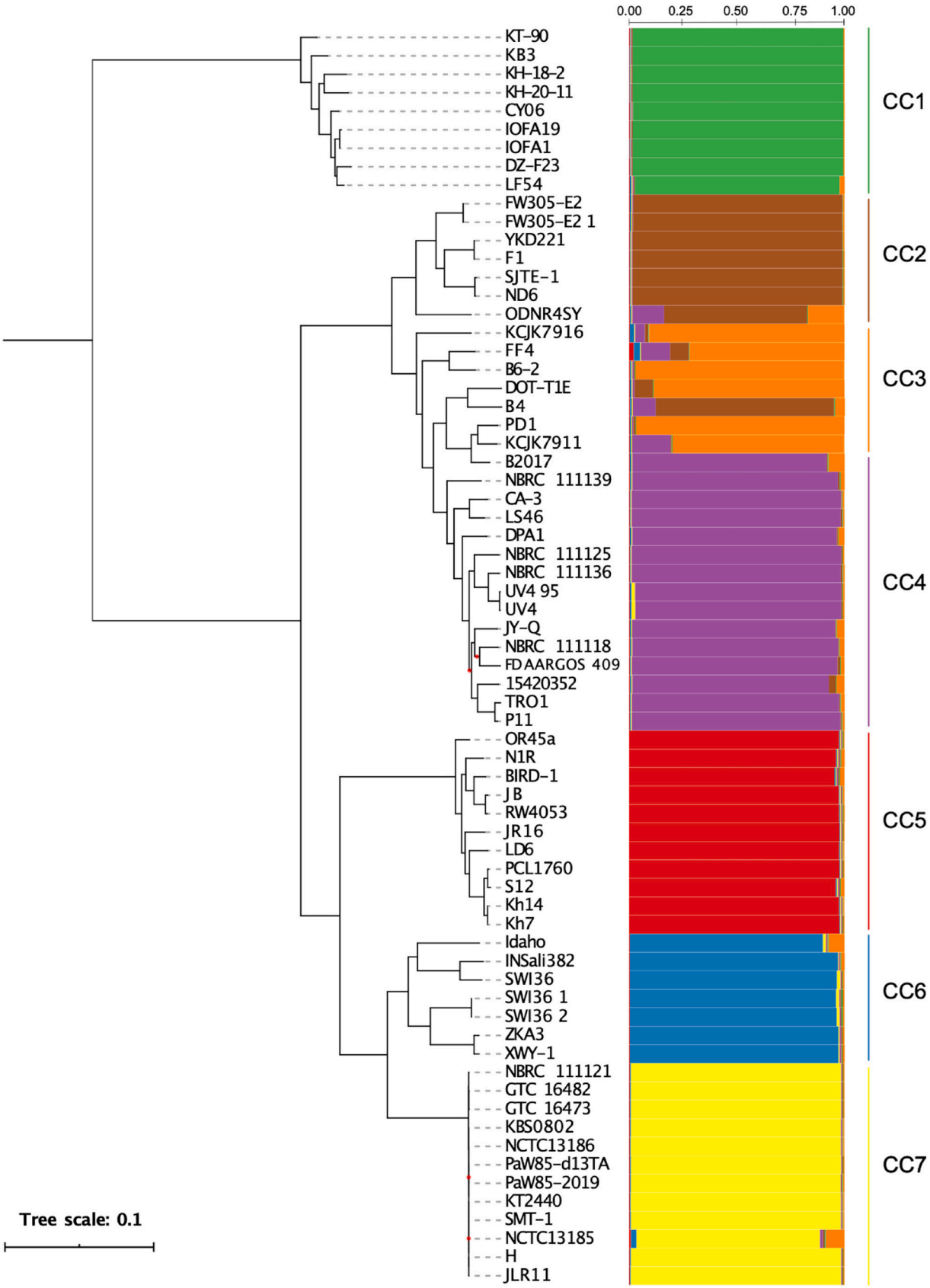

Fig. 4. Phylogenetic tree and coancestry barplots correlating the phylogenetic tree using SNPs extracted from core genome and coancestry probabilities assigned with STRUCTURE. Red dots in branches represent bootstrap values lower than 70%. Each colour represents one of the seven predicted Clonal Complexes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

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

mainly performed by the siderophore acinetobactin [37], which is synthesized by the proteins encoded by the *bauABCDE* operon. We found this operon in *P. alloputida* strains within CC5, CC6, and CC7 (Fig. 6a). Further, this operon was surrounded by genes coding for proteins from chemotaxis sensory transducer (PP_2599) and aminotransferases (PP_2588) families (Fig. 6b). Siderophore-mediated iron acquisition has been investigated in *P. alloputida* KT2440 [38] (CC7), but the role played by *bauABCDE* in this species is yet to be elucidated.

Present in all strains from other CCs, the type VI secretion system (T6SS) HSI-I is absent in CC5 and CC6 (except XWY-1) isolates (Fig. 6a).

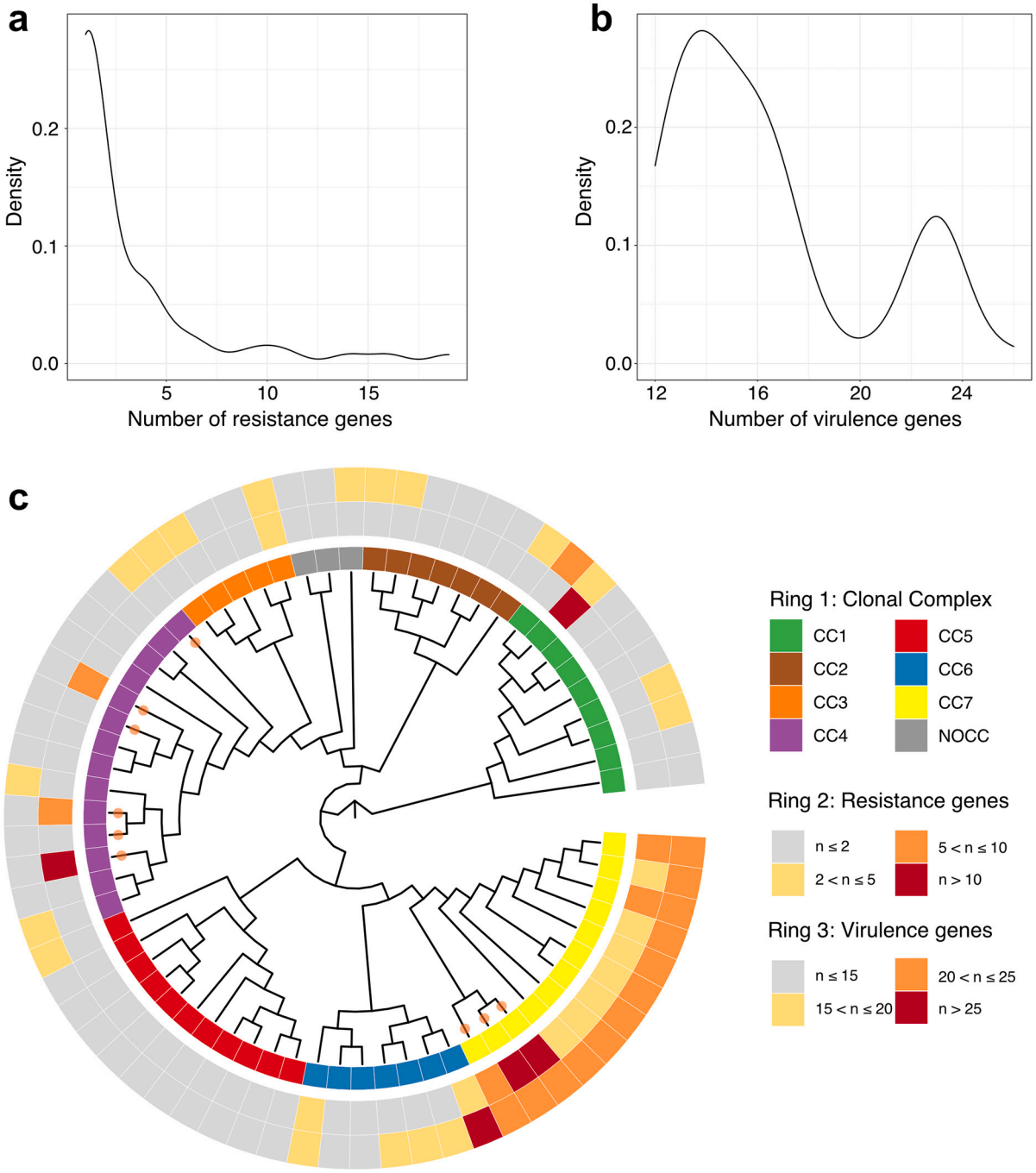

Fig. 5. Population structure and distribution of virulence and resistance genes. a, b. distribution of acquired resistance and virulence genes. c. Maximum likelihood tree from SNPs present in the core genome inferred with 68 genomes used in this study. Branches from clinical isolates are marked with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

In KT2440 (CC7), HSI-I is a potent weapon against other bacteria (e.g. phytopathogens), increasing the competitiveness of *P. alloputida* [39]. The absence of T6SS virulence genes has been reported for BIRD-1 (CC5) [39], and our work generalizes this observation to all CC5 members. In KT2440, T6SS is crucial to kill phytopathogens such as *Xanthomonas campestris* [40]. These results indicate that BIRD-1 and other members from CC5 are likely less efficient than KT2440 as biocontrol agents. Moreover, all clinical *P. alloputida* strains harbor T6SS, indicating their potential ability to outcompete other bacteria during infections.

2.6. Plant growth promotion and bioremediation properties

The ability of *P. putida* species to promote plant growth and bioremediate toxic compounds have been explored [12,41–43]. We searched for genes involved in plant growth promotion and bioremediation through a literature search of genes that have already been described for *Pseudomonas* species (Table S6). We found genes in the core genome, such as the pyrroloquinoline quinone-encoding operon *pqqBCDEFG,* associated with mineral phosphate solubilization in *Serratia marcescens* [44] and *Pseudomonas fluorescens* [45] (Table S7). Mineral P solubilization has already been reported experimentally for BIRD-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.

[41] and KT2440 [45], indicating that all *P. 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* [46]. 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 [46]. 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. However, we observed that *P. alloputida* lacks the main genes involved in indole-3-acetic acid synthesis (e.g. *ipdC*, *iaaM*, and *iaaH*), indicating an incomplete or nonfunctional pathway. Moreover, *P. alloputida* also lacks AcdS, an enzyme that counteract ethylene stress response, a result that has been experimentally confirmed in KT2440 [47].

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 [4,12]. One of the toluene-degrading pathways includes the *todABCDE* operon and the *todST* regulator. The *p*-cymene compound can be degraded by means of the *cymAaAbBCDER* or the *cmtAaAbAcAdBCDEFGHI* operon [48]. We found a genomic island of approximately 48 kbp harboring all these genes in F1, DOT-T1E, UV4, UV4/95, YKD221, and NBRC_111125 genomes (Fig. S9). All these isolates, except NBRC_111125, were experimentally confirmed to degrade toluene. Further, *P. alloputida* F1 is well-known to grow on toluene [4]; YKD221 was isolated from contaminated industrial soil and degrades *cis*-dichloroethene [49]; DOT-T1E is an isolate known to grow on different carbon sources [50] and; UV4 and UV4/95 conduct important industrial biotransformation of arenes, alkenes, and phenols [51]. Interestingly, genes in this genomic island presented a very similar genetic context (Fig. S7), with an upstream arm-type integrase associated with bacteriophages. We were unable to precisely define the *att* 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 [30]. 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 [30], corroborating the additional and important function to extrude antibiotics and organic solvents in all *P. alloputida* isolates. TtgDEF is located in the same genomic island of *tod* genes. This complex can expel toluene, but not antibiotics [52], reinforcing the variety of molecules that can be extruded by RND efflux pumps and the need to explore the structural basis of this specificity, not only in *P. alloputida* isolates, but also in other bacteria.

3. Concluding remarks

Through a remarkable metabolic versatility, *P. putida* species can thrive in a wide variety of niches. In this work, we explored the genetic diversity of *P. alloputida* and characterized its population structure for the first time. Through a large-scale genomic analysis, we identified a major problem with *P. putida species* classification, including several reference strains that likely belong to new species, as also suggested elsewhere [6]. *P. alloputida* has an open pangenome dominated by low-frequency genes. The population structure of this species has at least 7 clonal complexes that were verified by cgMLST and STRUCTURE ancestry simulations.

We analyzed genes of clinical and biotechnological interest. The low-frequency acquired resistance genes are predominant in plasmids from a few clinical strains. *P. alloputida* lacks key genes for indole-3-acetic acid production. 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 [7] with a minimum threshold of 90% completeness. The Kh7^T (GCA_900291035.1) was used as a reference with mash v.2.2.2 [53] to find genomes with distances up to 0.05. We used mashtree [8] to generate the distance tree. The ANI analysis was performed with pyani 0.2.10 [54]. 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 [55] to avoid bias in the identification of protein families. Plasmids were analyzed with PLSD v2020_06_29 [34].

4.2. Pangenome characterization

We inferred the *P. alloputida* pangenome using Roary 3.13.0 [56], 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 [23], retaining the 30 principal components and 3 discriminant functions. Pangenome openness and fluidity were conducted with micropan [57] 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 [58]. SNPs were retrieved with

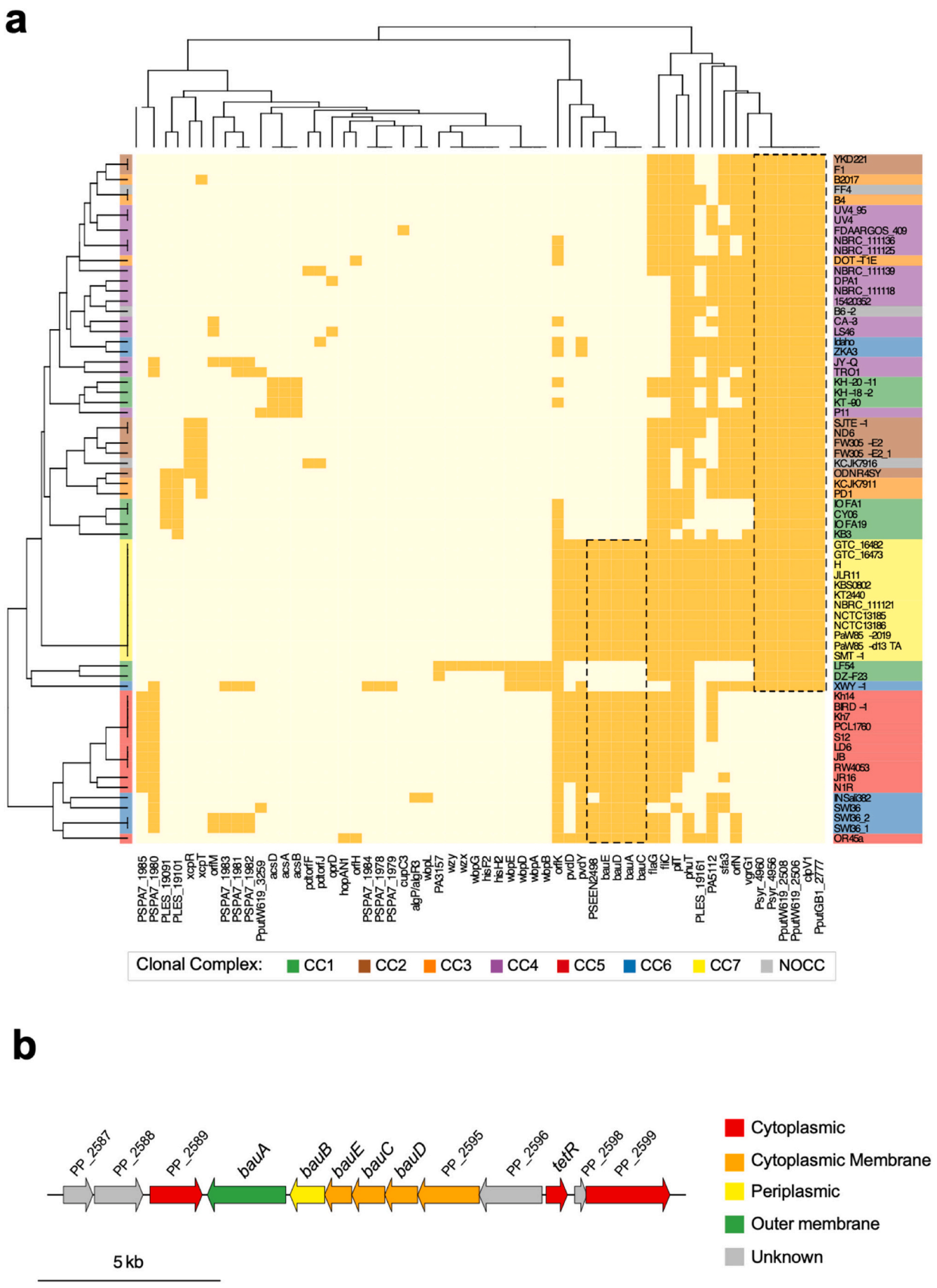

Fig. 6. Acquired virulome composition. a. Matrix with presence (dark squares) and absence (light squares) profiles of virulence genes. Rows represent strains colored based on Clonal Complex that they belong. Columns are virulence genes identified. b. *P. alloputida* KT2440 *bauABCDE* genetic context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

snp-sites v2.3.3 [59] and SNP alignment was used as input to RAxML v8 [60] 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 [61].

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 [24]. We used BLASTN [62] 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 [25] 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 [26]. 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 [27] to predict antibiotic resistance genes. The virulence factor database (VFDB) [35] 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 [63]. 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 [63] with 50% minimum coverage for query and subject and 60% minimum identity.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2021.09.008.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgements

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

References

[1] A.C. Parte, J. Sarda Carbasse, J.P. Meier-Kolthoff, L.C. Reimer, M. Goker, List of prokaryotic names with standing in nomenclature (LPSN) moves to the DSMZ, Int. J. Syst. Evol. Microbiol. 70 (2020) 5607–5612.

[2] M. Mulet, J. Lalucat, E. Garcia-Valdes, DNA sequence-based analysis of the *Pseudomonas* species, Environ. Microbiol. 12 (2010) 1513–1530.

[3] D.C. Volke, P. Calero, P.I. Nikel, Pseudomonas putida, Trends Microbiol. 28 (2020) 512–513.

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

[5] K.N. Timmis, *Pseudomonas putida*: a cosmopolitan opportunist par excellence, Environ. Microbiol. 4 (2002) 779–781.

[6] V. Keshavarz-Tohid, J. Vacheron, A. Dubost, C. Prigent-Combaret, P. Taheri, S. Tarighi, S.M. Taghavi, Y. Moenne-Loccoz, D. Muller, Genomic, phylogenetic and catabolic re-assessment of the *Pseudomonas putida* clade supports the delineation of *Pseudomonas alloputida* sp. nov., *Pseudomonas inefficax* sp. nov., *Pseudomonas persica* sp. nov., and *Pseudomonas shirazica* sp. nov, Syst Appl Microbiol 42 (2019) 468–480.

[7] M. Seppey, M. Manni, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness, Methods Mol. Biol. 1962 (2019) 227–245.

[8] L.S. Katz, T. Griswold, S.S. Morrison, J.A. Caravas, S. Zhang, H.C. den Bakker, X. Deng, H.A. Carleton, Mashtree: a rapid comparison of whole genome sequence files, Journal of Open Source Software 4 (2019).

[9] M. Richter, R. Rossello-Mora, Shifting the genomic gold standard for the prokaryotic species definition, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 19126–19131.

[10] L.M. Bobay, The prokaryotic species concept and challenges, in: H. Tettelin, D. Medini (Eds.), The Pangenome: Diversity, Dynamics and Evolution of Genomes, Cham (CH), 2020, pp. 21–49.

[11] K. Yonezuka, J. Shimodaira, M. Tabata, S. Ohji, A. Hosoyama, D. Kasai, A. Yamazoe, N. Fujita, T. Ezaki, M. Fukuda, Phylogenetic analysis reveals the taxonomically diverse distribution of the *Pseudomonas putida* group, J. Gen. Appl. Microbiol. 63 (2017) 1–10.

[12] K.E. Nelson, C. Weinel, I.T. Paulsen, R.J. Dodson, H. Hilbert, V.A. Martins dos Santos, D.E. Fouts, S.R. Gill, M. Pop, M. Holmes, L. Brinkac, M. Beanan, R. T. DeBoy, S. Daugherty, J. Kolonay, R. Madupu, W. Nelson, O. White, J. Peterson, H. Khouri, I. Hance, P. Chris Lee, E. Holtzapple, D. Scanlan, K. Tran, A. Moazzez, T. Utterback, M. Rizzo, K. Lee, D. Kosack, D. Moestl, H. Wedler, J. Lauber, D. Stjepandic, J. Hoheisel, M. Straetz, S. Heim, C. Kiewitz, J.A. Eisen, K.N. Timmis, A. Dusterhoft, B. Tummler, C.M. Fraser, Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440, Environ Microbiol 4 (2002) 799–808.

[13] S. Taghavi, C. Garafola, S. Monchy, L. Newman, A. Hoffman, N. Weyens, T. Barac, J. Vangronsveld, D. van der Lelie, Genome survey and characterization of endophytic bacteria exhibiting a beneficial effect on growth and development of poplar trees, Appl. Environ. Microbiol. 75 (2009) 748–757.

[14] R.A. Rosson, K.H. Nealson, Manganese binding and oxidation by spores of a marine bacillus, J. Bacteriol. 151 (1982) 1027–1034.

[15] M. Fernandez, M. Porcel, J. de la Torre, M.A. Molina-Henares, A. Daddaoua, M. A. Llamas, A. Roca, V. Carriel, I. Garzon, J.L. Ramos, M. Alaminos, E. Duque, Analysis of the pathogenic potential of nosocomial *Pseudomonas putida* strains, Front. Microbiol. 6 (2015) 871.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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. Oloni, D.J. Speicher, A. Florescu, B. Singh, M. Faltyn, A. Hernandez-Koutoucheva, A.N. Sharma, E. Bordeleau, A. C. Pawlowski, H.L. Zubyk, D. Dooley, E. Griffiths, F. Maguire, G.L. Winsor, R. G. Beiko, F.S.L. Brinkman, W.W.L. Hsiao, G.V. Domselaar, A.G. McArthur, CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database, Nucleic Acids Res 48 (2020) D517–D525.

[28] P.D. Lister, D.J. Wolter, N.D. Hanson, Antibacterial-resistant *Pseudomonas aeruginosa*: clinical impact and complex regulation of chromosomally encoded resistance mechanisms, Clin. Microbiol. Rev. 22 (2009) 582–610.

[29] H.P. Schweizer, Efflux as a mechanism of resistance to antimicrobials in *Pseudomonas aeruginosa* and related bacteria: unanswered questions, Genet. Mol. Res. 2 (2003) 48–62.

[30] A. Rojas, E. Duque, G. Mosqueda, G. Golden, A. Hurtado, J.L. Ramos, A. Segura, Three efflux pumps are required to provide efficient tolerance to toluene in *Pseudomonas putida* DOT-T1E, J. Bacteriol. 183 (2001) 3967–3973.

[31] Z.X. Tian, X.X. Yi, A. Cho, F. O'Gara, Y.P. Wang, 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 (2016), e1005932.

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

[33] P. Sanchez, J.F. Linares, B. Ruiz-Diez, E. Campanario, A. Navas, F. Baquero, J. L. Martinez, Fitness of in vitro selected *Pseudomonas aeruginosa* nalB and nfxB multidrug resistant mutants, J. Antimicrob. Chemother. 50 (2002) 657–664.

[34] V. Galata, T. Fehlmann, C. Backes, A. Keller, PLSDB: a resource of complete bacterial plasmids, Nucleic Acids Res. 47 (2019) D195–D202.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] P.I. Nikel, V. de Lorenzo, *Pseudomonas putida* as a functional chassis for industrial biocatalysis: from native biochemistry to trans-metabolism, Metab. Eng. 50 (2018) 142–155.

[44] 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.

[45] 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.

[46] 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.

[47] B.R. Glick, B. Todorovic, J. Czarny, Z. Cheng, J. Duan, B. McConkey, Promotion of plant growth by bacterial ACC deaminase, Critical Reviews in Plant Sciences 26 (2007).

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

[49] K. Yonezuka, N. Araki, J. Shimodaira, S. Ohji, A. Hosoyama, M. Numata, A. Yamazoe, D. Kasai, E. Masai, N. Fujita, T. Ezaki, M. Fukuda, Isolation and characterization of a bacterial strain that degrades cis-dichloroethenein the absence of aromatic inducers, J. Gen. Appl. Microbiol. 62 (2016) 118–125.

[50] C. Daniels, P. Godoy, E. Duque, M.A. Molina-Henares, J. de la Torre, J.M. Del Arco, C. Herrera, A. Segura, M.E. Guazzaroni, M. Ferrer, J.L. Ramos, Global regulation of food supply by *Pseudomonas putida* DOT-T1E, J. Bacteriol. 192 (2010) 2169–2181.

[51] T. Skvortsov, P. Hoering, K. Arkhipova, R.C. Whitehead, D.R. Boyd, C.C.R. Allen, Draft genome sequences of *Pseudomonas putida* UV4 and UV4/95, Toluene Dioxygenase-Expressing Producers of cis-1,2-dihydrodiols, Genome Announc 6 (2018).

[52] G. Mosqueda, J.L. Ramos, 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 (2000) 937–943.

[53] B.D. Ondov, T.J. Treangen, P. Melsted, A.B. Mallonee, N.H. Bergman, S. Koren, A. M. Phillippy, Mash: fast genome and metagenome distance estimation using MinHash, Genome Biol. 17 (2016) 132.

[54] L. Pritchard, H.R. Glover, S. Humphris, J.G. Elphinstone, I.K. Toth, Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens, Anal. Methods 8 (2016) 10–24.

[55] T. Seemann, Prokka: rapid prokaryotic genome annotation, Bioinformatics 30 (2014) 2068–2069.

[56] A.J. Page, C.A. Cummins, M. Hunt, V.K. Wong, S. Reuter, M.T. Holden, M. Fookes, D. Falush, J.A. Keane, J. Parkhill, Roary: rapid large-scale prokaryote pan genome analysis, Bioinformatics 31 (2015) 3691–3693.

[57] L. Snipen, K.H. Liland, micropan: an R-package for microbial pan-genomics, BMC Bioinformatics 16 (2015) 79.

[58] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, Mol. Biol. Evol. 30 (2013) 772–780.

[59] A.J. Page, B. Taylor, A.J. Delaney, J. Soares, T. Seemann, J.A. Keane, S.R. Harris, SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments, Microb Genom 2 (2016), e000056.

[60] A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (2014) 1312–1313.

[61] I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v4: recent updates and new developments, Nucleic Acids Res. 47 (2019) W256–W259.

[62] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.

[63] R.C. Edgar, Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26 (2010) 2460–2461.

Supplementary figures

Phylogenetic analysis and population structure of *Pseudomonas alloputida*

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

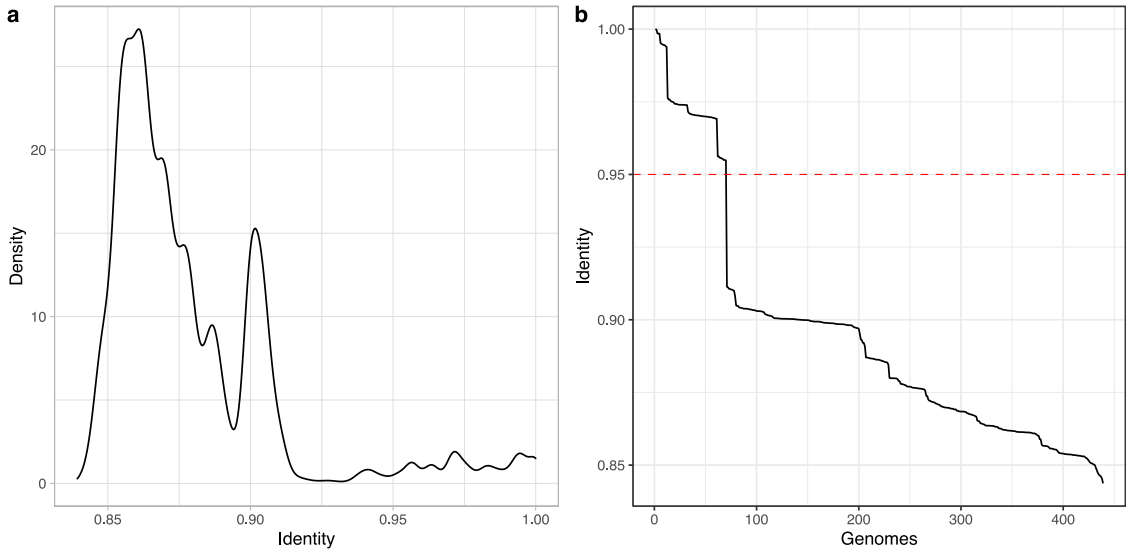

Figure S1. Average Nucleotide Identity distribution across *Pseudomonas putida* group. **a.** Identity density and **b.** ranked identity distribution in *P. putida* group from *P. putida* Kh7^{T}. Red dotted line represents the threshold used to define species based on average nucleotide identity.

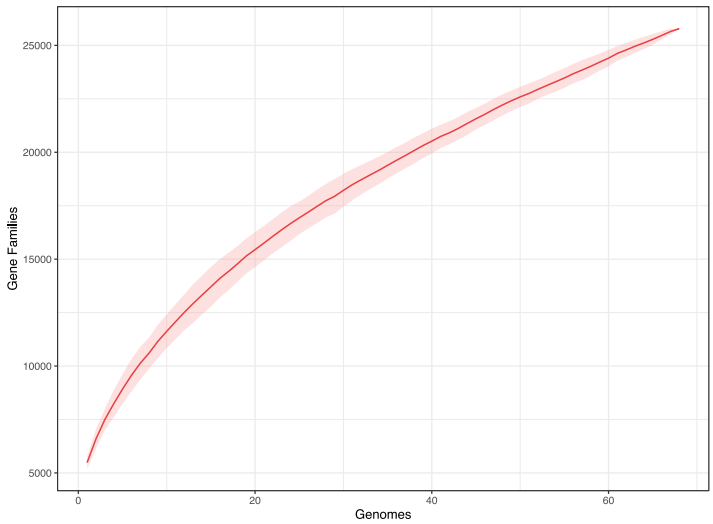

Figure S2. Cumulative curve of the *P. alloputida* pangenome. Gene families are in function of the number of isolates added sequentially. The slope (α) of the curve is 0.417, indicating an open pangenome.

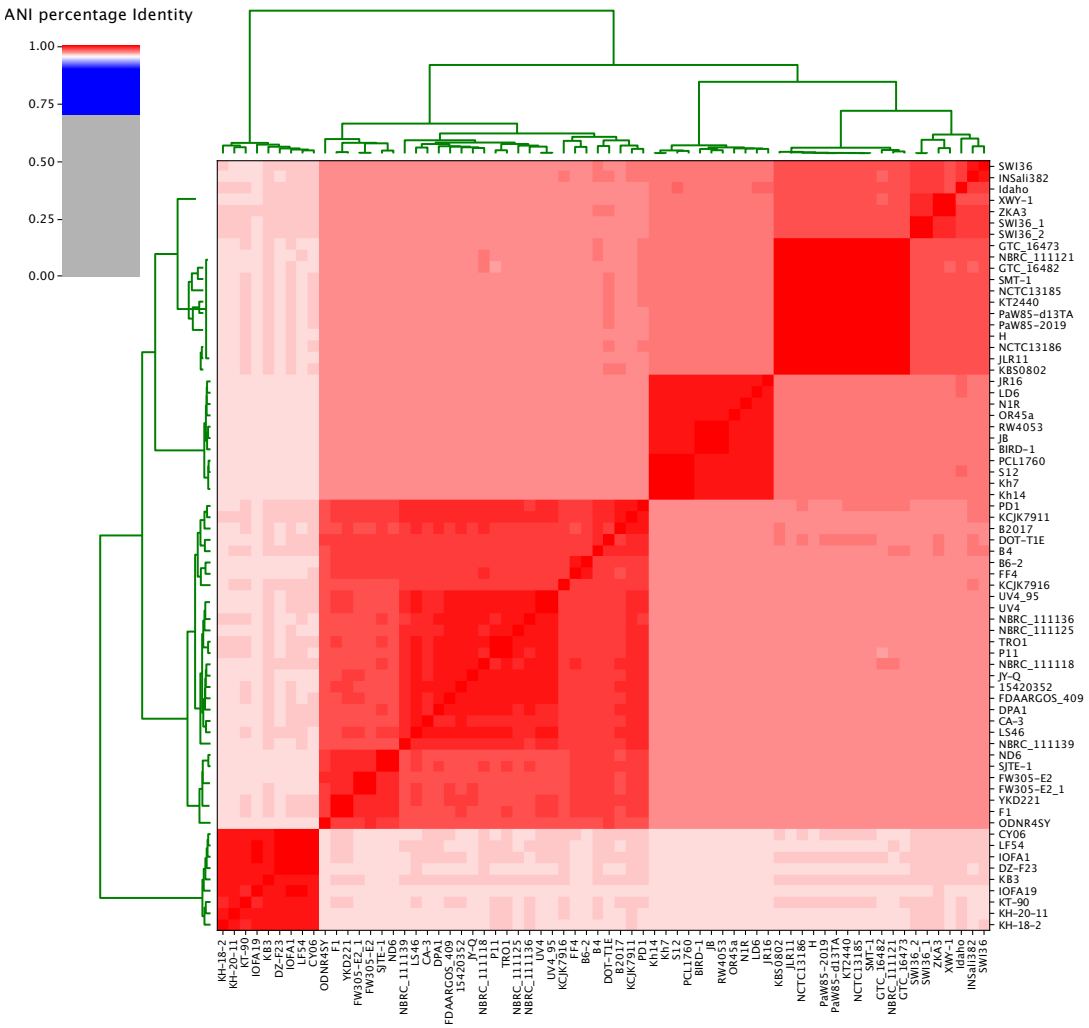

Figure S3. Average nucleotide identity analysis of the 68 isolates used in this study. Pairwise comparison indicating major groups in *P. alloputida*.

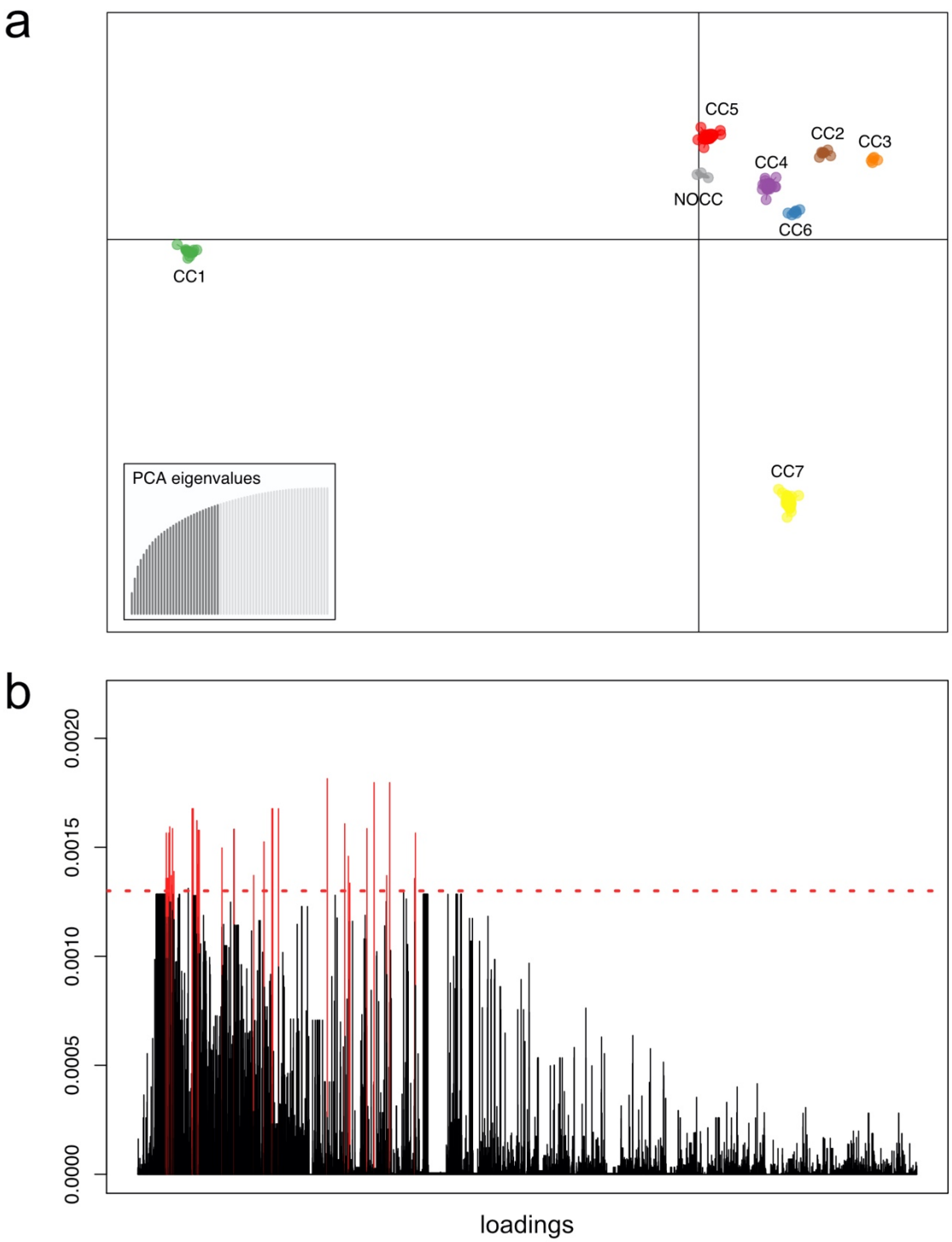

Figure S4. Discriminant Analysis of Principal Components of accessory genes present in 5% to 95% of the isolates. **a.** Clustering pattern of Clonal Complexes using the first two principal components of DAPC **b.** Loading plot. Red lines represent those gene families above the threshold of 0.0013 (red dotted line) and contributed more for the observed clustering patterns.

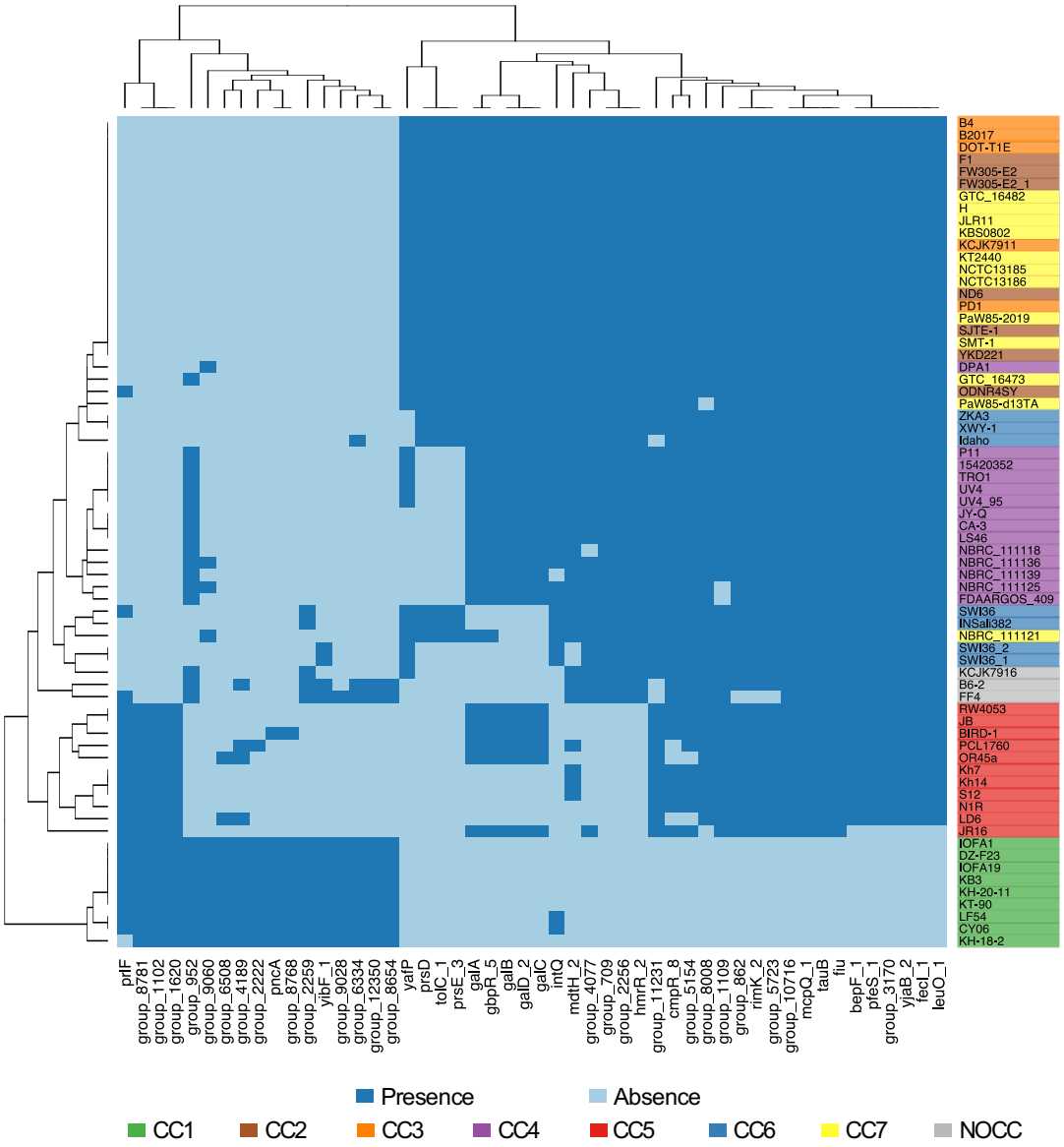

Figure S5. Heatmap from presence/absence profiles for top 50 genes detected to contribute for Clonal Complex clustering. Colors represent Clonal Complexes.

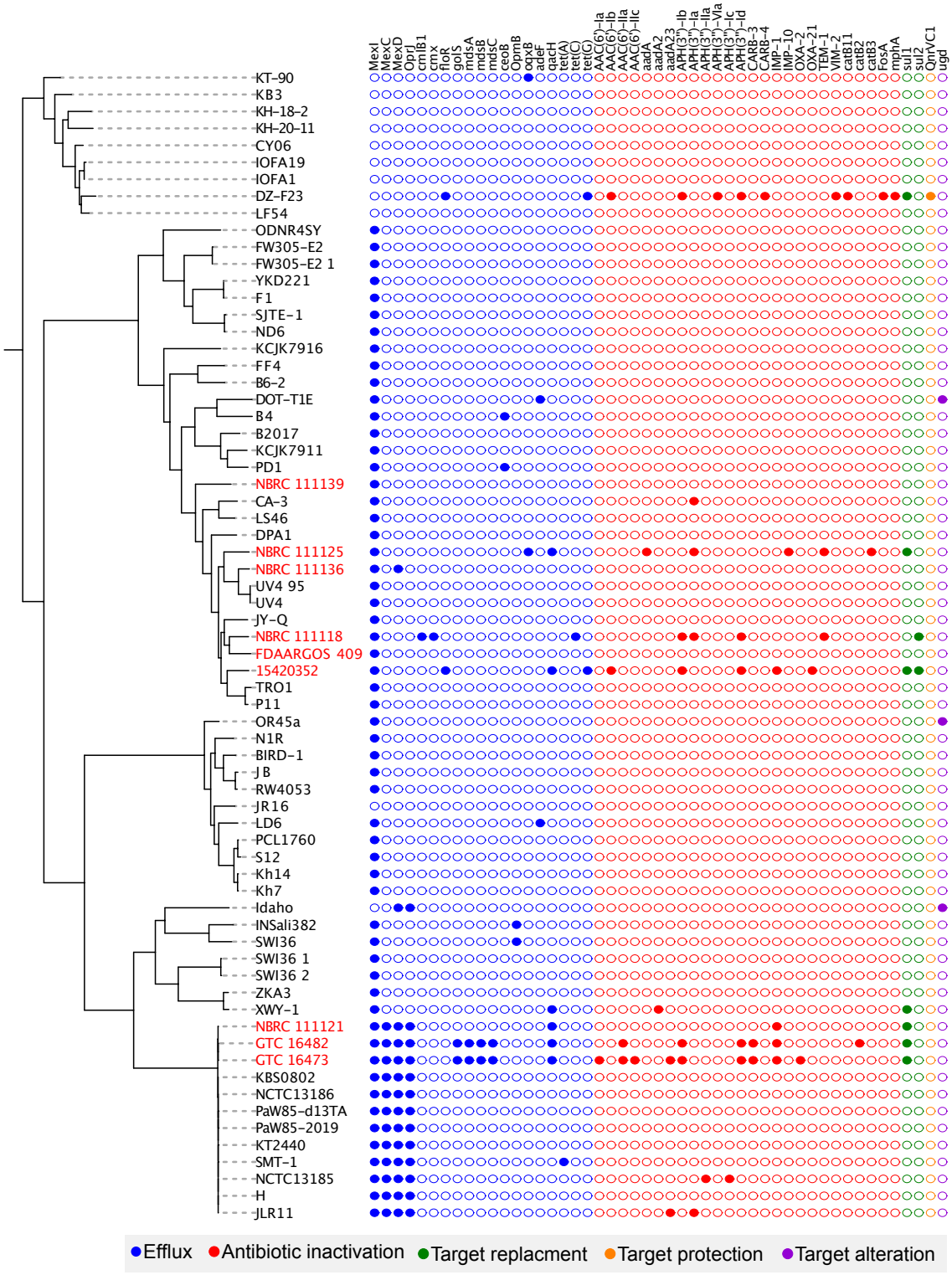

Figure S6. Accessory resistance genes identified in *P. alloputida*. Filled circles represent presence of a given gene and colors indicate the antibiotic resistance mechanism for each gene. Clinical strains are written in red.

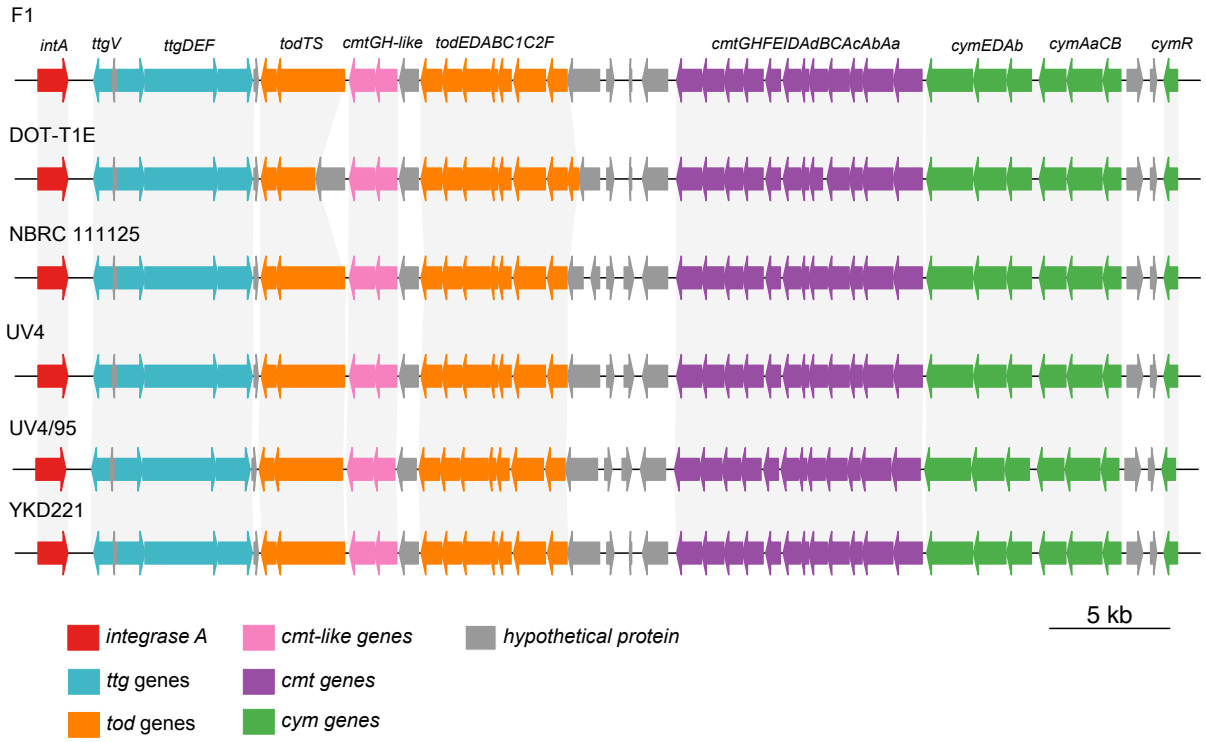

Figure S7. Genetic context of the genomic island containing *tod* genes in *P. alloputida* isolates. The gene coding for arm-type integrase A is upstream the genomic island, indicating a horizontal gene transfer for each represented isolate.

3.2 Network analysis of ten thousand genomes shed light on *Pseudomonas* diversity and classification

In our prior work on *P. alloputida*'s population structure[17], we employed an identity network approach, which illuminated a very structured network with well-defined communities within the *Pseudomonas putida* group. By using a 95% identity threshold to define species, the network was structured to the point that the number of detected communities was the same as the number of network components.

The second article of this thesis addresses how identity networks change their structure across different thresholds, using *Pseudomonas* genus as a model[16]. We employed an extensive dataset of 10,035 *Pseudomonas* genomes, including type strains, to construct a genomic identity network. We observed a network stabilization around 95% identity. This study also uncovers taxonomic inconsistencies and reveals that a substantial proportion of *Pseudomonas* genomes deposited in GenBank are misclassified. A phylogenetic analysis using single-copy genes revealed the presence of at least 14 distinct *Pseudomonas* groups, suggesting that *Pseudomonas* is an admixture of different genera and its taxonomy should be revisited.

Microbiological Research 254 (2022) 126919

Contents lists available at ScienceDirect

Microbiological Research

journal homepage: www.elsevier.com/locate/micres

Network analysis of ten thousand genomes shed light on *Pseudomonas* diversity and classification

Hemanoel Passarelli-Araujo [a,b,*], Glória Regina Franco [a], Thiago M. Venancio [b,*]

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

ARTICLE INFO

Keywords:
Phylogenomics
Pseudomonads
Taxonomy
Community detection

ABSTRACT

The growth of sequenced bacterial genomes has revolutionized the assessment of microbial diversity. *Pseudomonas* is a widely diverse genus, containing more than 254 species. Although type strains have been employed to estimate *Pseudomonas* diversity, they represent a small fraction of the genomic diversity at a genus level. We used 10,035 available *Pseudomonas* genomes, including 210 type strains, to build a genomic distance network to estimate the number of species through community identification. We identified taxonomic inconsistencies with several type strains and found that 25.65 % of the *Pseudomonas* genomes deposited on Genbank are misclassified. The phylogenetic tree using single-copy genes from representative genomes in each species cluster in the distance network revealed at least 14 *Pseudomonas* groups, including the *P. alcaligenes* group proposed here. We show that *Pseudomonas* is likely an admixture of different genera and should be further divided. This study provides an overview of *Pseudomonas* diversity from a network and phylogenomic perspective that may help reduce the propagation of mislabeled *Pseudomonas* genomes.

1. Introduction

Biological networks have been an essential analytical tool to better understand microbial diversity and ecology (Coutinho et al., 2015; Layeghifard et al., 2019). A network is a set of connected objects, in which objects can be represented as nodes and connections as edges. Networks provide a simple and powerful abstraction to evaluate the importance of individual or clustered nodes in maintaining a given system. Coupled with whole-genome sequencing, it can refine our knowledge about genetic relationships of diverse bacteria such as *Pseudomonas*.

Pseudomonas is a genus within the *Gammaproteobacteria* class, whose members colonize aquatic and terrestrial habitats. These bacteria are involved in plant and human diseases, as well as in biotechnological applications such as plant growth-promotion and bioremediation (Silby et al., 2011). The genus *Pseudomonas* was described at the end of the nineteenth century based on morphology, and its remarkable nutritional versatility was recognized thereafter (Palleroni, 2010). The metabolic diversity of pseudomonads, combined with biochemical tests to describe species, culminated in a chaotic taxonomic situation (Palleroni, 2010).

In 1984, the genus was revised and subdivided into five groups based on DNA-DNA and rRNA-DNA hybridization (Palleroni et al., 1894), with group I retaining the name *Pseudomonas*. Over the past 30 years, other molecular markers such as housekeeping genes have been used to mitigate the issues of *Pseudomonas* taxonomy (Ait Tayeb et al., 2005; Mulet et al., 2012; Gomila et al., 2015). Based on the 16S rRNA gene sequences, the genus is divided into three main lineages represented by *Pseudomonas pertucinogena*, *Pseudomonas aeruginosa*, and *Pseudomonas fluorescens* (Peix et al., 2018). These lineages comprise groups of different species – both lineages and groups receive the name of the representative species. Currently, there are 254 *Pseudomonas* species with validated names according to the List of Prokaryotic Names with Standing in the Nomenclature (LPSN) (Parte et al., 2020). However, although the genus division into lineages and groups has facilitated the classification of new species, the remnants of the *Pseudomonas* misclassification still linger in public databases (Gomila et al., 2017; Tran et al., 2017).

The explosion in the availability of complete genomes for both cultured and uncultured microorganisms has improved the classification of several bacteria, including *Pseudomonas* (Gomila et al., 2015; Parks et al., 2018). One of the gold standards for species circumscription is the digital whole-genome comparison by Average Nucleotide Identity (ANI)

* Corresponding authors at: Av. Alberto Lamego 2000, P5 sala 217, Parque Califórnia, Campos dos Goytacazes, RJ, 28013-602, Brazil.
E-mail addresses: hemanuel.passarelli@gmail.com (H. Passarelli-Araujo), thiago.venancio@gmail.com (T.M. Venancio).

https://doi.org/10.1016/j.micres.2021.126919
Received 5 October 2021; Accepted 10 November 2021
Available online 15 November 2021

(Bobay, 2020). Since using only genomes from type strains might bias the analysis and provide an unrealistic picture of microbial diversity, we aimed to estimate the *Pseudomonas* diversity using all available genomes through a network approach. Here, we provide new perspectives on *Pseudomonas* diversity by exploring the structure of the genomic distance network and the phylogenetic tree from representative genomes. This work also provides novel insights into the misclassification and phylogenetic borders of *Pseudomonas*.

2. Methods

2.1. Dataset collection and annotation

We recovered 11,025 genomes of *Pseudomonas* from Genbank in June 2020. Genome quality was evaluated with BUSCO v4.0.6 (Seppey et al., 2019) using the *Pseudomonadales* dataset. We defined completeness as 100 % minus the percentage of missing genes, and contamination as the fraction of duplicated genes. Quality was defined as completeness – 5 x contamination (Parks et al., 2018). Genomes with more than 400 contigs were removed, and contigs shorter than 500bp were discarded from the remaining genomes. We used mash v2.2.2 (Ondov et al., 2016) to calculate the pairwise distances between those 10,035 genomes with quality higher than 80 % using sketches of 1000 and 5000. Regarding the type strains, we used all species with available genomes and validated taxonomic names according to the LPSN (Parte et al., 2020) in March 2021. The pairwise distances between type strains were performed using pyani v0.2.10 (Pritchard et al., 2016). We reannotated the genomes with prokka v1.14 (Seemann, 2014) to allow a systematic large-scale genome comparison.

2.2. Network analysis

By using the pairwise Mash distances, we generated the corresponding network and obtained thestructural properties such as density, transitivity, and number of components with the igraph package (Csardi and Nepusz, 2006). We used the label propagation algorithm to detect communities (Raghavan et al., 2007). The representative genome for each community was defined based on three conditions: i) if the community has only one type strain, the type strain was considered the representative genome; ii) if the community has more than one type strain, the first described type strain was chosen; iii) else, we randomly chose a genome in a community (seed = 1996) and assigned the community name with the notation *Pseudomonas sppX*, where *X* is the community number.

2.3. Phylogeny and POCP index

We used OrthoFinder v2.5.2 (Emms and Kelly, 2019) to obtain the orthogroups from community representative genomes. All single-copy genes were aligned with MAFFT v7.467 (Katoh and Standley, 2013) and concatenated to reconstruct the *Pseudomonas* phylogeny with IQ-TREE v2.1.2 (Minh et al., 2020). The best-fit model detected through ModelFinder (Kalyaanamoorthy et al., 2017) was LG + F+ I + G4. One thousand bootstrap replicates were generated to assess the significance of internal nodes. Phylogenetic trees were visualized and annotated using ggtree (Yu, 2020). We tracked MRCA nodes for *Pseudomonas* groups definition using treeio (Yu, 2020).

The Percentage of Conserved Proteins (POCP) between two genomes was calculated using the formula $\frac{C_1+C_2}{T_1+T_2}$, where C is the number of conserved proteins and T is the total number of proteins (Qin et al., 2014). The number of conserved proteins was obtained from the orthologs matrix A_{ij} generated by OrthoFinder, where each entry $(i.j)$ is the total number of genes in species i that have orthologues in species j. The graphs were generated and visualized using igraph (Csardi and Nepusz, 2006) and ggnetwork v0.5.8 (Briatte, 2020), respectively. The GTDB classification was obtained in April 2021 (http://gtdb.ecogenomic.org/).

3. Results

3.1. Dataset collection

We obtained 11,025 genomes from GenBank in June 2020. After evaluating the quality of each genome (see methods for more details) and removing fragmented genomes, 10,035 genomes passed in the 80 % quality threshold (Fig. S1). The size of the retrieved genomes ranged from 3.0–9.4 Mb. We used 238 type strains with available genomes and names validly published according to the *List of Prokaryotic Names with Standing in Nomenclature* in March 2021. The genome size and GC content of type strains ranged from 3,022,325 bp and 48.26 % (*P. caeni*) to 7,375,852 bp and 62.79 % (*P. saponiphila*) (Table S1). According to the NCBI classification, the top four abundant species in our dataset are *P. aeruginosa* (n = 5,088), *P. viridiflava* (n = 1,509), *Pseudomonas* sp. (n = 1,083), and *P. syringae* (n = 435) (Table S2).

3.2. Genome-based analysis reveals the presence of synonymous Pseudomonas species

The misclassification of some *Pseudomonas* type strains has been reported by several studies (Gomila et al., 2015; Hesse et al., 2018; Lalucat et al., 2020; Passarelli-Araujo et al., 2021). Type strains play an essential role in taxonomy by anchoring species names as unambiguous points of reference (Hugenholtz et al., 2021). In this context, the term "synonym" refers to the situation where the same taxon receives different scientific names. We used 238 type strain genomes to evaluate the presence of synonymous species in *Pseudomonas*. The ANI was computed for all type strains to construct an identity network that was further used to check the linkage between genomes based on a 95 % ANI threshold (Fig. 1). Since 95 % has been accepted as species delimitation threshold (Bobay, 2020), connections between type strains indicate synonymous names or subspecies.

We identified 30 connected genomes in the ANI network (Fig. 1). Four of these connected genomes are expected because they represent *P. chlororaphis* and its subspecies. Of the 26 remaining connected species, 15 have been previously reported, such as that in the group containing *P. amygdali*, *P. ficuserectae*, and *P. savastanoi* (Hesse et al., 2018; Lalucat et al., 2020). Here, we observed 11 connections, including the one between *P. panacis* and *P. marginalis* with 97.34 % identity, suggesting that *P. panacis* is a later synonym of *P. marginalis*.

3.3. The Pseudomonas genomic distance network is highly structured

In networks, identifying communities plays an important role in understanding network structure. We used all 10,035 *Pseudomonas* genomes to construct a distance network to estimate the number of *Pseudomonas* species from the number of communities detected in this network. Since alignment-based methods to estimate genome similarity (e.g. ANI) are computationally expensive due to the algorithm quadratic time complexity (Backurs and Indyk, 2015), it becomes impractical for thousand genomes. Therefore, we estimated the Mash distance that strongly correlates with ANI and can be rapidly computed for large datasets (Ondov et al., 2016).

Mash distances are computed by reducing large sequences to small and representative sketches (Ondov et al., 2016). We estimated the pairwise Mash distance for all genomes using sketch sizes of 1000 and 5000, which converged to similar distance values (Fig. S2a). However, we observed that the greater the distance between two *Pseudomonas* genomes, the more divergent the distance estimation (Fig. S2b), although the distribution is similar (Fig. S2c). The final distance between two genomes was given as the average distance value from both sketch sizes. We used the reciprocal Mash distance (1 - Mash) to estimate

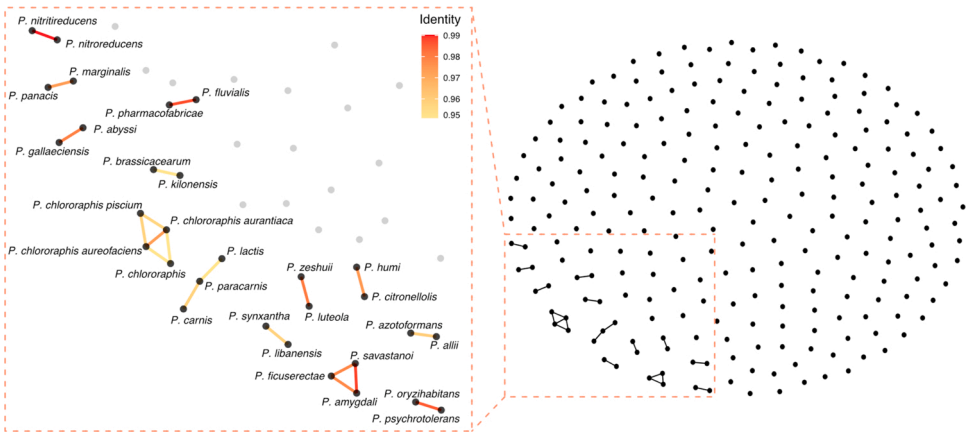

Fig. 1. Type strain validation based on Average Nucleotide Identity. Each node in the network represents a type strain genome and nodes are connected if they share at least 95 % identity. The left panel is a magnified representation of the connected nodes, with edges colored according to percent identity between the nodes.

the ANI for all 10,035 genomes.

We generated a weighted *Pseudomonas* network considering nodes as genomes and edges as the identity between two genomes. Although the 95 % ANI value has been widely accepted to delineate species, we evaluated how different thresholds affect network structure by assessing density, transitivity, and the number of connected components (Fig. 2). The network density, i.e., the ratio of the number of edges and the number of possible edges, decreased throughout the interval but stabilized between 90 % and 97 % ANI, keeping the network structure almost unchanged (Fig. 2a). To estimate how structured the network was with different ANI thresholds, we also computed the average network transitivity (also called average clustering coefficient) (Fig. 2b). The average transitivity is the normalized sum over all local transitivities (the probability of a given node having adjacent nodes interconnected). The high transitivity values revealed that the *Pseudomonas* network is highly structured (i.e., formed by tightly connected clusters) (Fig. 2b). This structured profile was observed before for the *P. putida* group network (Passarelli-Araujo et al., 2021), indicating that communities in *Pseudomonas* distance networks rarely overlap.

To decrease the influence of overrepresented species (e.g., *P. aeruginosa*) on the topological network statistics, we also computed the variation in the number of components (Fig. 2c). A connected component in a network is a subset of nodes connected via a path. At 70 % identity, we had a single giant connected component. Expectedly, the number of connected components increased with the identity threshold because of the emergence of smaller components or even orphan nodes. Interestingly, connected components with more than ten nodes arose only above 81 % identity threshold and stabilized close to 95 %, highlighting that the 95 % ANI threshold is accurate for species demarcation.

We used the *Pseudomonas* network discarding connections lower than 95 % identity to estimate the number of species from the number of communities in the network. We detected 573 communities by using the label propagation algorithm (Raghavan et al., 2007). This number is similar to the number of connected components at 95 % identity threshold (n = 570), further supporting that the *Pseudomonas* distance network is highly structured, containing non-overlapping communities. By considering each community as a different *Pseudomonas* species, we evaluated the distribution of type strains in these communities.

Seventeen communities had more than one type strain in the same cluster, indicating the existence of later heterotypic synonyms, as shown in Fig. 1. For each community, we assigned only one representative genome (see methods for more detail). For example, in the community containing *P. amygdali, P. ficuserectae, and P. savastanoi,* we maintained *P. amygdali* as the representative strain and the others were considered later heterotypic synonyms, as previously proposed (Gomila et al., 2017). We observed that only 210 communities (36.64 %) had representative genomes from validly described species, reinforcing the underestimation of the number of *Pseudomonas* species if only the type strains are considered.

Regarding the community's sizes, *P. aeruginosa* corresponds to the largest community, comprising 5116 genomes (Fig. 3, Table S3). Most communities had few genomes. Although large communities tend to have type strains, 61 type strains (29.04 %) are single nodes (Fig. 3, Table S3), further demonstrating that estimating the diversity of *Pseudomonas* only by type strains severely underestimates diversity. For example, the community containing *Pseudomonas spp7* has 122 genomes and is potentially a new genomospecies.

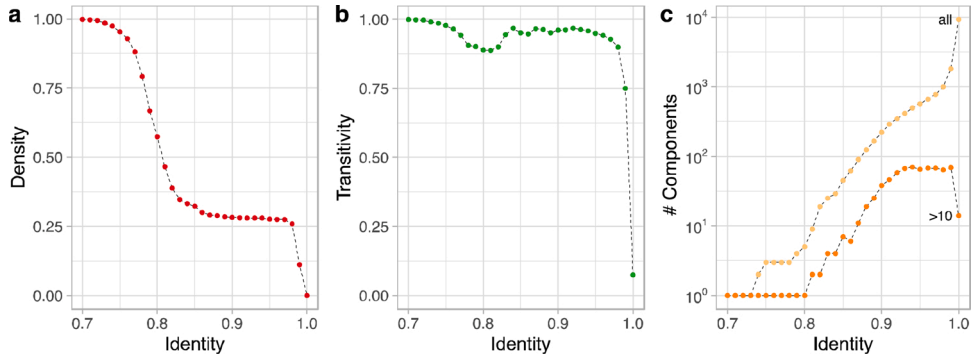

Fig. 2. *Pseudomonas* distance network structure evolution. a) Proportion of present connections (network density) and b) average transitivity change over different identity (1 – Mash) cut-off values. c) Number of network components detected with different identity thresholds. Light orange dots represent the total number of components, whereas the dark dots represent only components with more than ten nodes.

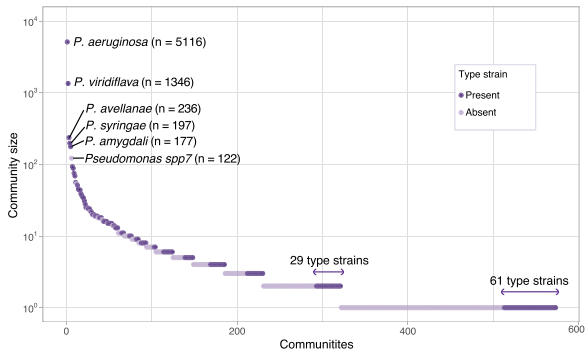

Fig. 3. *Pseudomonas* community sizes. Dark and light purple dots represent communities with and without type strains, respectively. The names and number of genomes are displayed in those communities with more than 100 genomes. The y-axis is in log scale.

3.4. Comparison with NCBI classification highlights Pseudomonas misclassification

After delimiting the species by the community detection approach, we compared them with the classification available in NCBI Taxonomy (Schoch et al., 2020). Briefly, we computed how many genomes were deposited with a given species name and how many genomes were identified for that species by our network approach. Of the 10,035 genomes used in this work, 25.65 % were misclassified in NCBI Taxonomy (Table S5). This proportion includes species considered as later synonyms that should be reclassified (e.g. *P. savastanoi*), non-classified genomes (*Pseudomonas* sp.), and those genomes that are unconnected to the expected species cluster. The most poorly classified species were *P. brassicacearum* (95.65 %), *P. fluorescens* (95.23 %), *P. stutzeri* (94.58 %), and *P. putida* (88.70 %). This high rate of misclassification is linked to the type strain determined for each species. For example, the critical misclassification problem of *P. putida* has been recently reported by us (Passarelli-Araujo et al., 2021). The *P. putida* NBRC 14164^{T} type strain forms an isolated community in the network with only 15 genomes. On the other hand, the community of *P. alloputida* Kh7^{T} harbors 69 genomes, constituting the largest community in the *P. putida* group. Thus, most of the genomes deposited as *P. putida* are actually from *P. alloputida*. Regarding the misclassification of *P. stutzeri*, 122 genomes fall into the community represented by *Pseudomonas spp7*, a potentially new genomospecies mentioned above.

We also assessed the impact of our approach defining the species-level taxonomy of the 1,083 non-classified *Pseudomonas* genomes available in Genbank (*Pseudomonas sp.*). Interestingly, 511 *Pseudomonas* sp. genomes (47.18 %) were distributed among 97 communities containing type strains (Table S6). The species that received the most genomes were *P. glycinae* (n = 35), *P. lactis* (n = 34), and *P. mandelii* (n = 31).

3.5. The Pseudomonas phylogeny reveals at least fourteen groups

To reduce the influence of overrepresented species, we used the 573 representative genomes from each community to retrieve orthologous genes and reconstruct the *Pseudomonas* phylogeny. The *Cellvibrio japonicus* Ueda 107^{T} was used as an outgroup. We identified 31,094 orthogroups, of which 168 were present in all species, including 30 single-copy genes. We used the single-copy genes to reconstruct the *Pseudomonas* phylogeny and identify the main *Pseudomonas* groups (Fig. 4).

The main *Pseudomonas* groups have been previously characterized using housekeeping genes such as 16S rDNA, *gyrB*, *rpoB*, and *rpoD* from type strains (Gomila et al., 2015; Hesse et al., 2018). To delineate each group, we retrieved those representative genomes (species) within previously-described groups (Table S7). We then tracked the Most Recent Common Ancestor (MRCA) for those species in the *Pseudomonas* phylogenetic tree to include uncharacterized representative genomes as well. For example, the *P. lutea* group comprises three known species: *P. abietaniphila*, *P. graminis*, and *P. lutea* (Gomila et al., 2015). By tracking the corresponding MRCA node, we ensured the monophyly and included *P. bohemica* and 12 uncharacterized species in this group

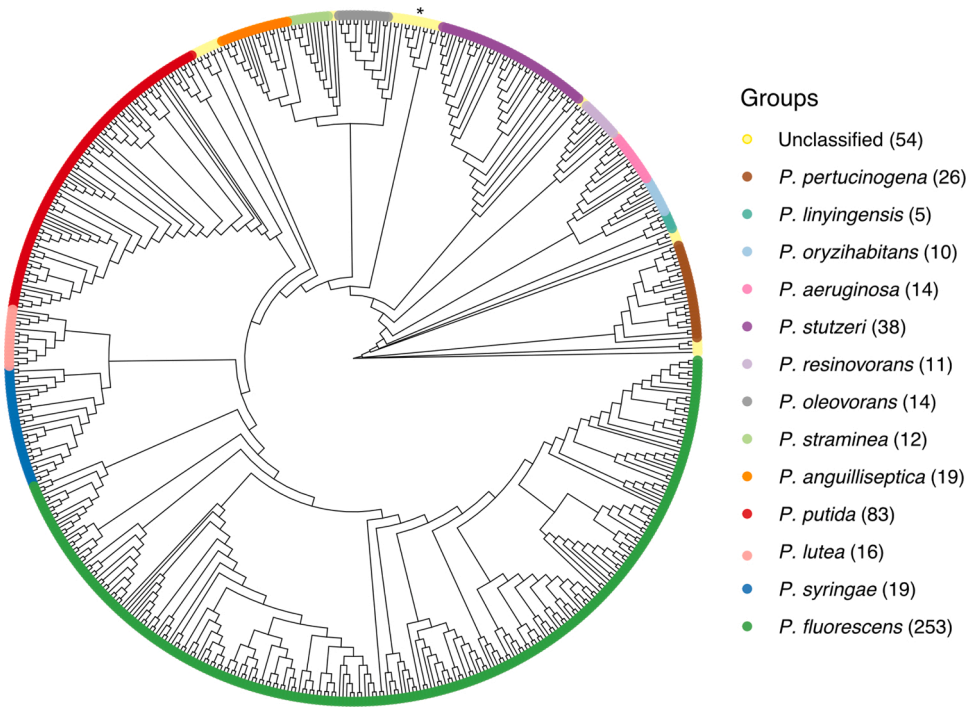

Fig. 4. Phylogenetic tree mapping Pseudomonas groups. Maximum-likelihood phylogenetic tree using core single-copy genes in representative genomes from 573 communities detected in the *Pseudomonas* network. Colors indicate *Pseudomonas* groups. The number of genomes in each group is in parenthesis. The asterisk highlights the *P. alcaligenes* group described here. The outgroup is *Cellvibrio japonicus* Ueda 107^{T}.

(Table S3). This approach allowed a more accurate characterization of both recently described type strains and other uncharacterized species (Fig. 4, Table S3). We identified the 13 main *Pseudomonas* groups and one new group with 10 genomes and three type strains: *P. alcaligenes*, *P. fluvialis*, and *P. pohangensis* (Fig. 4, Table S8). Since *P. alcaligenes* is the firstly-described type strain in this group (Monias, 1928), we named this group as *P. alcaligenes* group.

3.6. Lineage and genus boundaries

The genus *Pseudomonas* has three recognized lineages: *P. pertucinogena*, *P. aeruginosa*, and *P. fluorescens*. The *P. pertucinogena* lineage is composed of a single phylogenetic group. The *P. aeruginosa* lineage comprises 6 phylogenetic groups (*P. oryzihabitans*, *P. stutzeri*, *P. oleovorans*, *P. resinovorans*, *P. aeruginosa,* and *P. linyingensis*). The *P. fluorescens* lineage also comprises 6 phylogenetic groups (*P. fluorescens*, *P. lutea*, *P. syringae*, *P. putida*, *P. anguilliseptica*, and *P. straminea*); *P. fluorescens* group is further divided into 8 or 9 phylogenetic subgroups (Hesse et al., 2018). In this work, 70.38 % of the communities (species) belong to the *P. fluorescens* lineage, 16.72 % to *P. aeruginosa*, and 4.52 % to *P. pertucinogena*; 8.36 % were unclassified communities. We observed that, unlike the *P. pertucinogena* and *P. fluorescens* lineages, the *P. aeruginosa* lineage is polyphyletic (Fig. 5a).

We used the Genome Taxonomy Database (GTDB) approach (Parks et al., 2018) to evaluate whether *Pseudomonas* should be divided into different genera. GTDB proposes a framework to classify genomes in higher taxonomic ranks (e.g. genus). By using the GTDB classification, *Pseudomonas* should be divided into 17 genera named generically with "Pseudomonas" followed by a letter (e.g. "*Pseudomonas_A*"), with *P. aeruginosa* group retaining the name *Pseudomonas*. We found a high correspondence between *Pseudomonas* groups and the proposed genera, with few inconsistencies (Fig. 5a, Table S8). According to the GTDB classification, the *P. fluorescens* lineage, together with the *P. oleovorans* group and the here described *P. alcaligenes* group, would form a single genus called *Pseudomonas_E* (Fig. 5a), which corresponds to 77.52 % of the species (communities) estimated in our study.

We also used the Percentage of Conserved Proteins (POCP) index to evaluate the relationships between lineages (Fig. 5b) and complement the GTDB approach. Briefly, the POCP index measures the proportion of shared proteins between two genomes (Qin et al., 2014). The original proposal is that genomes belong to the same genus if they share at least half of their proteins (Qin et al., 2014). By using 50 % as a threshold, we observed that only the outgroup *C. japonicus* and other four genomes do not belong to the main POCP network component with all lineages. However, we observed two main clusters by using a 60 % threshold to link communities (Fig. 5b).

Apart from *P. anguilliseptica* and *P. straminea* groups, the *P. fluorescens* lineage forms an isolated component in the network (Fig. 5b). The *P. pertucinogena* and *P. aeruginosa* lineages are in the same component, but linked by a few connections, including a bridge via a *P. caeni* genome. The outgroup *C. japonicus* is an orphan in the network, as well as *P. kirkiae*. The species *P. boreopolis*, *P. cissicula*, and *P. geniculata* were also isolated. These three species have already been recognized as belonging to the genus *Xanthomonas* (Anzai et al., 2000). Nevertheless, they remain classified as *Pseudomonas* in Genbank and are still labeled as validly published with a correct name in LPSN.

4. Discussion

The *Pseudomonas* genus underwent several taxonomic reclassifications over the years. Here, we used 10,035 *Pseudomonas* genomes to estimate the genus diversity through network analysis and community detection. We observed that several type strains are later synonyms and should be officially revised, as also noted elsewhere (Gomila et al., 2015; Hesse et al., 2018).

Regarding the *Pseudomonas* network, we observed that the number of detected communities is very close to the number of network components at a 95 % identity threshold. Combined with the stabilization of density and high transitivity around this threshold, we conclude that the *Pseudomonas* network is highly structured. This structured network profile has also been noted previously reported for the *P. putida* group (Passarelli-Araujo et al., 2021).

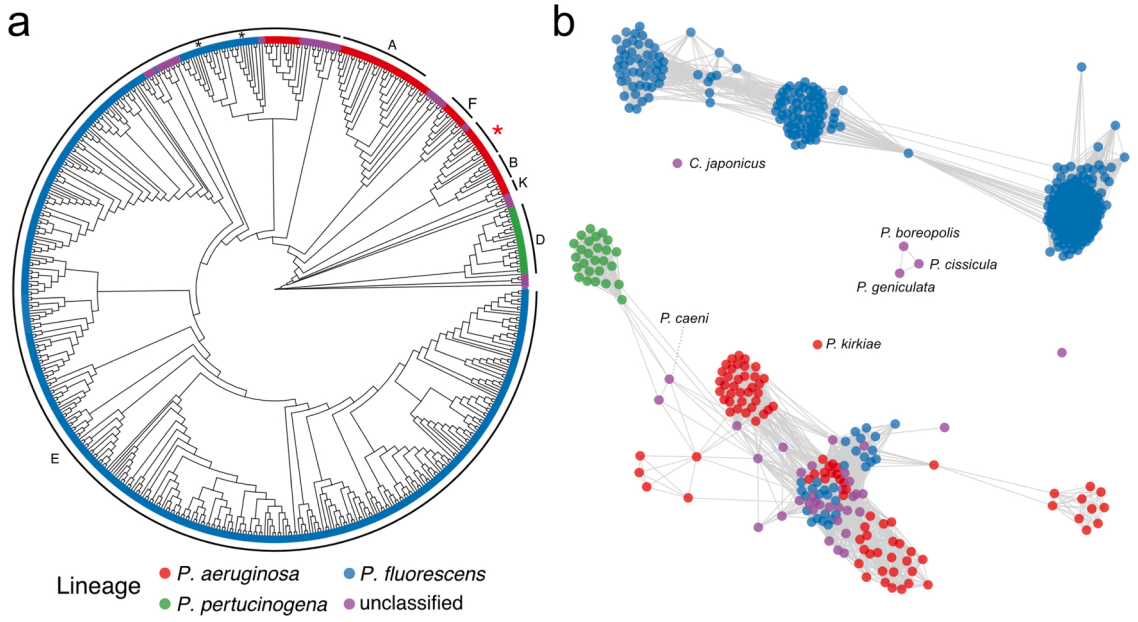

Fig. 5. *Pseudomonas* phylogenetic tree with proposed genus boundaries and Percentage of Conserved Proteins (POCP) network. a) Phylogenetic tree annotated with *Pseudomonas* lineages. The outer letters indicate the annotation adopted by the Genome Taxonomy Database (GTDB). The genus proposed to keep the name *Pseudomonas* is marked with a red asterisk. Other genera proposed by GTDB adopt the nomenclature "*Pseudomonas*" followed by a letter (e.g. *Pseudomonas_E*); for clarity, only the letters and those proposed genera with more than five communities are displayed. b) Network based on POCP index using a 60 % threshold. Colors represent lineages. Blue nodes embedded in the component with genomes of the *Pseudomonas aeruginosa* lineage belong to the groups *P. anguilliseptica* and *P. straminea*; these two groups are marked in the phylogenetic tree with black asterisks.

Considering each community as a different genomospecies, we identified 573 communities, way more than the 233 *Pseudomonas* species with validly published names. Moreover, we found 61 orphan type strains in the network, indicating that the diversity estimated using only type strains is highly underestimated. In addition, this work shows that 25.65 % of the *Pseudomonas* genomes are misclassified. This is a matter of concern, as misclassified genomes in public repositories can introduce noise to pangenome studies, reduce strain typing accuracy, and propagate labeling errors to several studies, including those characterizing new species.

Here, we also showed potential new genomospecies. For example, the community assigned as *Pseudomonas spp7* contains 122 genomes, and it is a sister group of *P. stutzeri*. The high misclassification proportion of *P. stutzeri* (Table S5) can be explained by the presence of this new closely-related species. Such inconsistencies could be mitigated through a standardized taxonomic framework, as previously proposed (Hugenholtz et al., 2021). However, there is still resistance to define species based solely on genome sequences, even with the massive number of available genomes (Hugenholtz et al., 2021). Therefore, isolating and characterizing members from *Pseudomonas spp7* community will allow the consolidation of this new species.

Although previous works provided insights about what would be considered *Pseudomonas* (Ozen and Ussery, 2012; Gomila et al., 2015; Hesse et al., 2018), how to delimit the *Pseudomonas* genus remains an open question. We tried to address this problem by using GTDB classification and POCP index network, two approaches proposed to delimit genera. The GTDB results indicate that the *P. fluorescens* lineage and the *P. oleovorans* and *P. alcaligenes* groups would constitute a genus with the generic name *Pseudomonas_E* (Fig. 4). However, the POCP index network at 60 % shows that *P. straminea* and *P. anguilliseptica* groups are closer to *P. aeruginosa* than to *P. fluorescens* lineage (Fig. 4b). Aiming for a parsimonious separation, we propose that the *P. fluorescens* lineage, excluding the *P. straminea* and *P. anguilliseptica* groups, should be considered a new genus. Furthermore, by the GTDB results, the *Pseudomonas* groups from the *P. aeruginosa* lineage should also be revised to assess whether they are new genera, as the *P. aeruginosa* lineage itself is polyphyletic. Prioritizing the GTDB approach here should provide the best alternative because it normalizes taxonomic ranks and ensures group monophyly (Parks et al., 2018).

5. Conclusion

In this study, we estimated the *Pseudomonas* diversity using a network approach. We show that type strains represent less than half of the estimated number of species, and that many of them are orphans in the network. We discovered new genomospecies and groups, such as *Pseudomonas spp7* and *P. alcaligenes*, respectively. Although genus delineation is somewhat complex, we propose the *Pseudomonas* genus division by combining GTDB classification and POCP index. To fully understand the *Pseudomonas* diversity, it will be important to focus on each group and characterize species from communities without type strains. This study provides a state-of-the-art classification to delimit bacterial species, which we expect to serve as a guide for future studies with *Pseudomonas spp*, reducing the problems caused by misclassified genomes.

Author contributions

Conceptualization: Hemanoel Passarelli-Araujo and Thiago M. Venancio; Formal analysis: Hemanoel Passarelli-Araujo; Data Visualization: Hemanoel Passarelli-Araujo; Resources: Thiago M. Venancio and Glória R. Franco; Writing: Hemanoel Passarelli-Araujo, Thiago M. Venancio, and Glória R. Franco; Supervision: Thiago M. Venancio and Glória R. Franco.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgements

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

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.micres.2021.126919.

References

Ait Tayeb, L., Ageron, E., Grimont, F., Grimont, P.A., 2005. Molecular phylogeny of the genus Pseudomonas based on rpoB sequences and application for the identification of isolates. Res. Microbiol. 156 (5-6), 763–773.
Anzai, Y., Kim, H., Park, J.Y., Wakabayashi, H., Oyaizu, H., 2000. Phylogenetic affiliation of the pseudomonads based on 16S rRNA sequence. Int. J. Syst. Evol. Microbiol. 50 (Pt 4), 1563–1589.
Backurs, A., Indyk, P., 2015. Edit distance cannot be computed in strongly subquadratic time (Unless SETH is False). SIAM J. Sci. Comput. 47, 10.
Bobay, L.M., 2020. The prokaryotic species concept and challenges. In: Tettelin, H., Medini, D. (Eds.), The Pangenome: Diversity, Dynamics and Evolution of Genomes, pp. 21–49. Cham (CH).
Briatte, F., 2020. Ggnetwork: Geometries to Plot Networks With 'ggplot2': R Package Version 0.5.8..
Coutinho, F.H., Meirelles, P.M., Moreira, A.P., Paranhos, R.P., Dutilh, B.E., Thompson, F. L., 2015. Niche distribution and influence of environmental parameters in marine microbial communities: a systematic review. PeerJ 3, e1008.
Csardi, G., Nepusz, T., 2006. The igraph software package for complex network research. InterJournal. Complex Systems:1695.
Emms, D.M., Kelly, S., 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20 (1), 238.
Gomila, M., Pena, A., Mulet, M., Lalucat, J., Garcia-Valdes, E., 2015. Phylogenomics and systematics in Pseudomonas. Front. Microbiol. 6, 214.
Gomila, M., Busquets, A., Mulet, M., Garcia-Valdes, E., Lalucat, J., 2017. Clarification of taxonomic status within the Pseudomonas syringae species group based on a phylogenomic analysis. Front. Microbiol. 8, 2422.
Hesse, C., Schulz, F., Bull, C.T., Shaffer, B.T., Yan, Q., Shapiro, N., Hassan, K.A., Varghese, N., Elbourne, L.D.H., Paulsen, I.T., Kyrpides, N., Woyke, T., Loper, J.E., 2018. Genome-based evolutionary history of Pseudomonas spp. Environ. Microbiol. 20 (6), 2142–2159.
Hugenholtz, P., Chuvochina, M., Oren, A., Parks, D.H., Soo, R.M., 2021. Prokaryotic taxonomy and nomenclature in the age of big sequence data. ISME J.
Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14 (6), 587–589.
Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30 (4), 772–780.
Lalucat, J., Mulet, M., Gomila, M., Garcia-Valdes, E., 2020. Genomics in bacterial taxonomy: impact on the genus Pseudomonas. Genes (Basel) 11 (2).
Layeghifard, M., Li, H., Wang, P.W., Donaldson, S.L., Coburn, B., Clark, S.T., Caballero, J. D., Zhang, Y., Tullis, D.E., Yau, Y.C.W., Waters, V., Hwang, D.M., Guttman, D.S., 2019. Microbiome networks and change-point analysis reveal key community changes associated with cystic fibrosis pulmonary exacerbations. NPJ Biofilms Microbiomes 5 (1), 4.
Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37 (5), 1530–1534.
Monias, B.L., 1928. Classification of Bacterium alcaligenes pyocyaneum and fluorescens. J. Infect. Dis. 43 (4), 330–334.
Mulet, M., Gomila, M., Scotta, C., Sanchez, D., Lalucat, J., Garcia-Valdes, E., 2012. Concordance between whole-cell matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry and multilocus sequence analysis approaches in species discrimination within the genus Pseudomonas. Syst. Appl. Microbiol. 35 (7), 455–464.
Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., Phillippy, A.M., 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17 (1), 132.
Ozen, A.I., Ussery, D.W., 2012. Defining the Pseudomonas genus: where do we draw the line with Azotobacter? Microb. Ecol. 63 (2), 239–248.
Palleroni, N.J., 2010. The Pseudomonas story. Environ. Microbiol. 12 (6), 1377–1383.

Palleroni, N.J., Genus, I., Migula, Pseudomonas, 1894. Bergey's manual of systematic. Bacteriology 1984 (1), 59.
Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A., Hugenholtz, P., 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. 36 (10), 996–1004.
Parte, A.C., Sarda Carbasse, J., Meier-Kolthoff, J.P., Reimer, L.C., Goker, M., 2020. List of Prokaryotic names with standing in Nomenclature (LPSN) moves to the DSMZ. Int. J. Syst. Evol. Microbiol. 70 (11), 5607–5612.
Passarelli-Araujo, H., Jacobs, S.H., Franco, G.R., Venancio, T.M., 2021. Phylogenetic analysis and population structure of Pseudomonas alloputida. Genomics 113 (6), 3762–3773.
Peix, A., Ramirez-Bahena, M.H., Velazquez, E., 2018. The current status on the taxonomy of Pseudomonas revisited: an update. Infect. Genet. Evol. 57, 106–116.
Pritchard, L., Glover, H.R., Humphris, S., Elphinstone, J.G., Toth, I.K., 2016. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Anal. Methods 8 (1), 10–24.
Qin, Q.L., Xie, B.B., Zhang, X.Y., Chen, X.L., Zhou, B.C., Zhou, J., Oren, A., Zhang, Y.Z., 2014. A proposed genus boundary for the prokaryotes based on genomic insights. J. Bacteriol. 196 (12), 2210–2215.
Raghavan, U.N., Albert, R., Kumara, S., 2007. Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. 76 (3 Pt 2), 036106.
Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., McVeigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J.P., Sun, L., Turner, S., Karsch-Mizrachi, I., 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford) 2020.
Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30 (14), 2068–2069.
Seppey, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: assessing genome assembly and annotation completeness. Methods Mol. Biol. 1962, 227–245.
Silby, M.W., Winstanley, C., Godfrey, S.A., Levy, S.B., Jackson, R.W., 2011. Pseudomonas genomes: diverse and adaptable. FEMS Microbiol. Rev. 35 (4), 652–680.
Tran, P.N., Savka, M.A., Gan, H.M., 2017. In-silico taxonomic classification of 373 genomes reveals species misidentification and new genospecies within the genus Pseudomonas. Front. Microbiol. 8, 1296.
Yu, G., 2020. Using ggtree to visualize data on tree-like structures. Curr. Protoc. Bioinformatics 69 (1), e96.

SUPPLEMENTARY FIGURES

Network analysis of ten thousand genomes shed light on *Pseudomonas* diversity and classification

Hemanoel Passarelli-Araujo[1,2,*], Glória Regina 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.

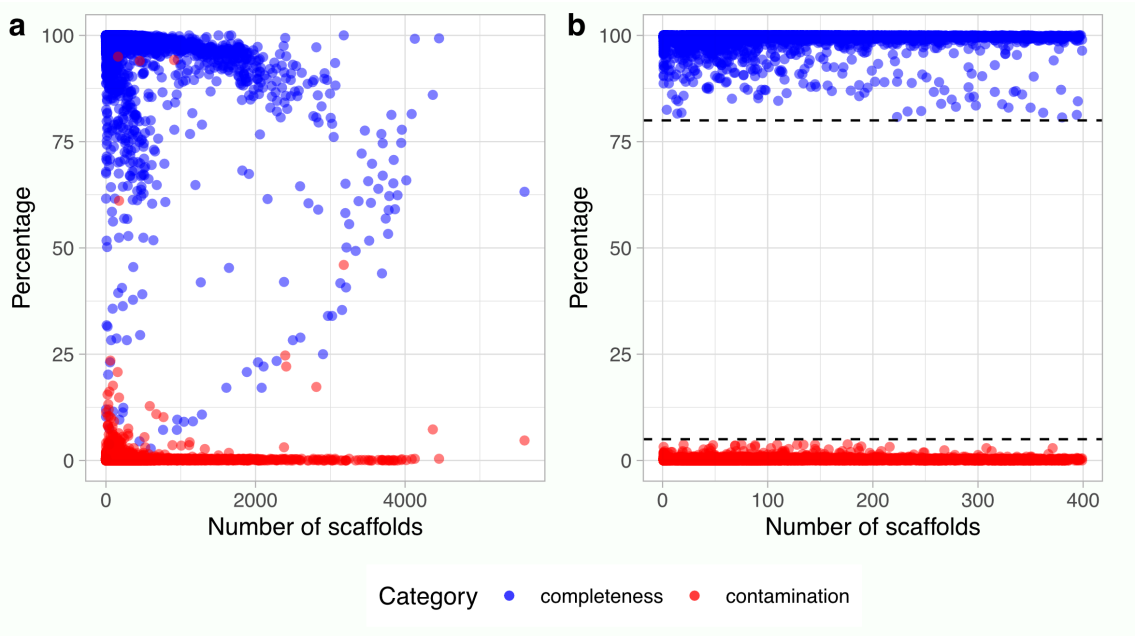

Figure S1. BUSCO estimation for completeness and contamination for all *Pseudomonas* genomes. a) Distribution for all 11,025 *Pseudomonas* genomes. b) Genomes used in this study after discarding genomes based on 80% quality threshold and fragmentation higher than 400 scaffolds (see methods).

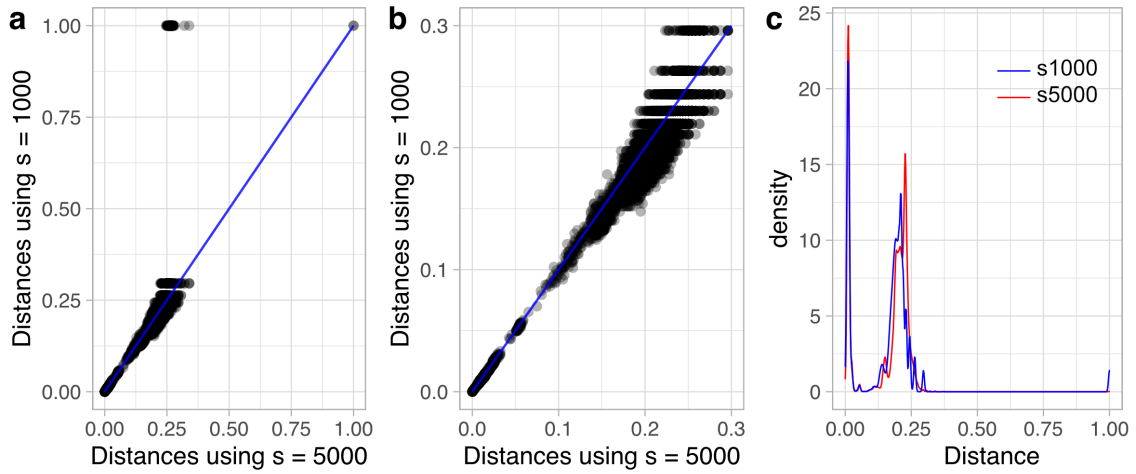

Figure S2. Mash distance statistics. a) Comparison of estimated Mash distance using sketches sizes of 1000 and 5000. b) Mash distances restricted to the interval [0.0, 0.3] in both axes. c) Mash distance distribution for each sketch size.

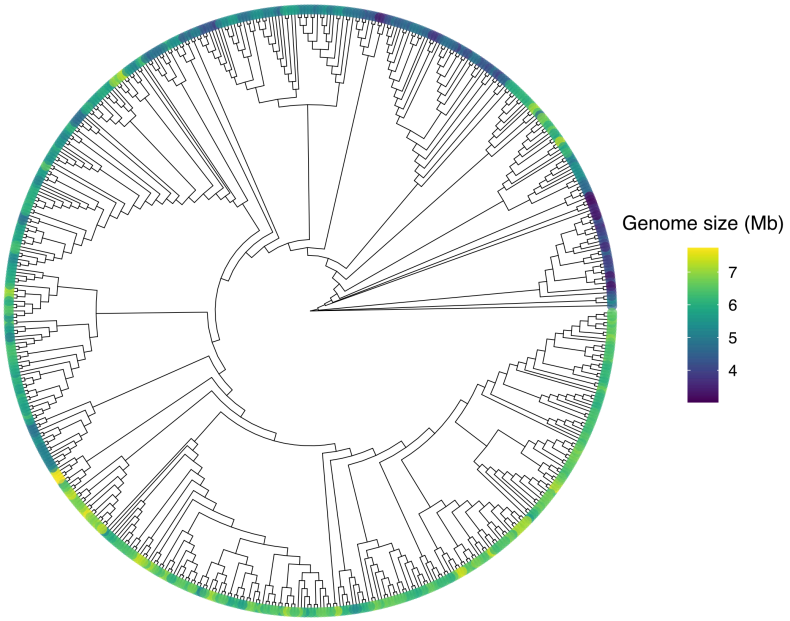

Figure S3. Genome size distribution for *Pseudomonas* communities. Maximum-likelihood phylogenetic tree using core single-copy genes in representative genomes from 573 communities detected in the *Pseudomonas* distance network. Colors indicate the genome size distribution.

3.3 Unveiling bacterial genetic discontinuity across different species provide insights into genetic and ecological diversity

From the two previous works, we observed that well-structured identity networks across *Pseudomonas* species. This third article delves into the quantification of bacterial genetic discontinuity and its ecological significance beyond Pseudomonas to evaluate whether this phenomenon is observed in other species.

Bacterial genetic discontinuity refers to abrupt genomic identity shifts among species. In this article, a dataset comprising 210,129 bacterial genomes is harnessed to systematically quantify genetic discontinuity patterns across diverse bacterial species. The research reveals clear breakpoints in genomic identity distributions and establishes a significant correlation between pangenome saturation and genetic discontinuity. Closed pangenomes are associated with more pronounced genetic breaks, exemplified by *Mycobacterium tuberculosis*.

Moreover, machine learning techniques identify key features that impact genetic discontinuity prediction. This study significantly advances our understanding of bacterial genetic patterns and their ecological implications, offering insights into species boundaries among prokaryotes.

Unveiling bacterial genetic discontinuity across different species provide insights into genetic and ecological diversity

Hemanoel Passarelli-Araujo[1,2,*], Thiago M. Venancio[3,*], William P Hanage[1]

[1]Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, Massachusetts, USA

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

[3]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

Running title: Quantifying the extent and ecological impact of Bacterial Genetic Discontinuity

Email: HPA: hemanuel.passarelli@gmail.com; TMV: thiago.venancio@gmail.com

Abstract

Bacterial genetic discontinuity, representing abrupt breaks in genomic identity among species, is crucial for grasping microbial diversity and evolution. Advances in genomic sequencing have enhanced our ability to track and characterize genetic discontinuity in bacterial populations. However, exploring systematically whether bacterial diversity exists as a continuum or into discrete species groups remains a challenge in microbial ecology. Here, we aimed to quantify the genetic discontinuity (δ) and investigate their ecological relevance. We harnessed a dataset comprising 210,129 genomes to systematically explore genetic discontinuity patterns across several distantly related species. Our findings revealed clear breakpoints in genomic identity distributions. By delving into pangenome characteristics, we uncovered a significant association between pangenome saturation and genetic discontinuity. Closed pangenomes were associated with more pronounced breaks, exemplified by *Mycobacterium tuberculosis*. Additionally, through a machine learning approach, we detected key features that impact genetic discontinuity prediction. Our study enhances the understanding of bacterial genetic patterns and their ecological implications, offering insights into species boundaries for prokaryotes.

Key-words: pangenome; machine learning; speciation; genetic rate of change; bacterial ecology

INTRODUCTION

Bacteria exhibit remarkable genetic makeup and ecological versatility, thriving in diverse niches worldwide. Plummeting sequencing costs have led to a wealth of genomic data, enabling extensive exploration of genetic diversity and evolutionary relationships across bacterial species. However, an essential inquiry in microbial ecology pertains to whether bacterial diversity exists as a continuum or as distinct species groups[1, 2, 3].

The definition of bacterial species faces challenges because bacteria can exchange genetic material through horizontal gene transfer (HGT)[4], potentially blurring the species boundaries. This complexity has led to divergent views on bacterial species existence: while it was once thought that excessive recombination would preclude their species formation[5], a contemporary perspective suggests that the gene flow patterns can even delineate species[6]. Besides, recent studies have revealed a clear genetic discontinuity across bacterial genomes, supporting the existence of discrete genetic clusters (species)[7, 8, 9, 10].

Genetic discontinuity refers to the occurrence of a significant difference in genetic makeup between populations or groups of organisms[11], thereby signifying potential boundaries between distinct species. This discontinuity can occur over time through natural selection, genetic drift, or geographic isolation. Besides, genetic discontinuity can be an important factor in determining whether populations should be classified as separate species[8, 10].

Defining bacterial species is beyond a human desire to catalog bacterial diversity; it is vital for understanding how evolutionary forces shape genetic lineages[12, 13]. Furthermore, proper classification impacts practical applications in industry, agriculture, and medicine. For instance, *Gardnerella vaginalis* illustrates the clinical relevance of naming individuals properly. Formerly grouped under *G. vaginalis*, the division into multiple species has revealed diverse health associations[14], including species linked to bacterial vaginosis to those found in healthy vaginal microbiomes[15]. Therefore, the previous classification of all *Gardnerella* as *G. vaginalis* limited the ability of clinicians to assess when and whether the presence of *Gardnerella* indicated a health risk.

One way to assess species boundaries is by estimating the genetic relatedness between genomes. A robust method to classify bacterial species is based on the Average Nucleotide Identity (ANI) estimate, with organisms belonging to the same species if the possess around 95% ANI or more among themselves[4, 5]. Since estimating ANI for thousands of genomes is computationally expensive, alternative methods were developed to accommodate the growing genomic dataset[10, 16].

Despite observed breaks in genetic identity distributions for various species, quantifying the magnitude of these breaks and their ecological implications remains a challenge. Here, we address the intricate nature of bacterial diversity through a genomic distance network approach and pangenome analysis. We aimed to quantify the extent of genetic discontinuity within and between bacterial populations to determine whether

intrinsic genetic boundaries can provide a more ecologically relevant basis for species redefinition. Here, we seek not to pigeonhole bacterial species, but to examine the presence of genetic boundaries, quantify their extent, and explore their ecological implications for species classification.

Results

Dataset information

We obtained 258,603 genomes from RefSeq that were filtered following the GTDB protocol[17]. After removing low quality and fragmented genomes, we retained of 210,129 genomes to explore bacterial genetic discontinuity (Table S1). According to the NCBI classification, the top four abundant species in our dataset are *Escherichia coli* (n = 22,853), *Staphylococcus aureus* (n = 12,747), *Klebsiella pneumoniae* (n = 10,387), and *Salmonella enterica* (n = 9,755) (Table S1).

Over 44 billions of comparisons were performed to construct an identity matrix M. The next step was defining communities in this network. Representative genomes for each community were then selected by removing edges below 95% identity in M, resulting in 7,122 communities identified using label propagation – a proxy for species number (see methods for more details). Notably, 84.84% of communities contained fewer than 10 genomes, consistent with prior observations in the genus *Pseudomonas*[18]. For instance, a previous study found that 29% of officially recognized *Pseudomonas* type strains appeared as isolated nodes in a similar network analysis, highlighting the substantial underestimation of diversity when relying solely on type strains[18]. By focusing on communities with over 50 genomes (3.85%), we obtained a set of 261 representative genomes, enabling meaningful comparative genomic analyses subsequentially.

We reannotated a total of 45,550 genomes, representing nine phyla according to GTDB classification (Figure 1). Out of 942,094 genes detected, 95.1% were successfully assigned across 39,552 orthogroups. Moreover, species-specific orthogroups represented approximately 0.9% of the genes. Single-copy genes were used to reconstruct the phylogenetic tree. Except for *Proteobacteria,* the tree indicated monophyly in all other phyla based on GTDB classification, contrasting with the polyphyly observed based on NCBI classification, as noted before[17]. The GC content in these genomes ranged from 25.9% in *Mesomycoplasma hyorhinis* to 73.4% in *Streptomyces albidoflavus* (Table S2). Notably, *Clostridioides difficile* exhibited the greatest number of CRISPR arrays (median = 8) in a community of 500 genomes, with at least one array each. The dataset included genomes with diverse sizes, from the smallest

commensal *Metamycoplasma hominis* (0.7 Mb) to the larger free-living bacterium *Burkholderia cepacia* (8.5 Mb).

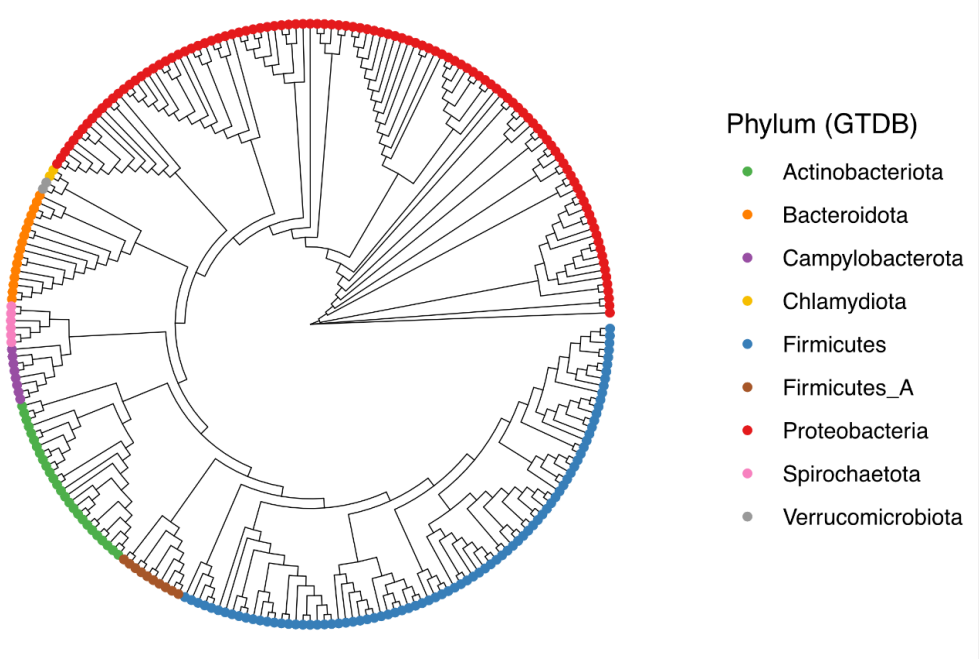

Figure 1. Phylogenetic tree of representative species. The phylogenetic tree reconstructed from single-copy genes of 261 representative species used in this work to explore bacterial genetic discontinuity.

Clear genetic discontinuity revealed across bacterial species

To explore genetic discontinuity across different species, we employed an egocentric-based strategy, where a representative genome serves as a "bait" node within the network and the identity of all other genomes from it is calculated. This method allowed us to assess the ranked identity distribution from each representative genome (Figure 2a), revealing clear breakpoints within the distribution. For instance, considering the representative species of *Acinetobacter baumannii*, the 4968th genome maintains a 97.27% identity. However, the identity of the 4969th genome drops drastically to 93.34%, exemplifying the observed genetic discontinuity or genetic break.

We systematically quantified the genetic discontinuity by estimating how rapidly the genomic identity decayed as we moved through the sorted identity array. We took the first derivative of the distribution, offering a measure of variability in genomic similarity that we named as Genetic Rate of Change (GRC) (Figure 2b). The maximum value of GCR resulted in the genetic discontinuity metric δ, which characterizes the steepest change in genomic identity (see methods). For instance, the genetic break observed from 97.27% to 93.34% in *A. baumannii*, corresponds to $\delta = 0.9727 - 0.9334 = 0.0393$ for this species (Table 1, Figure 2b).

Eight species were selected to showcase bacterial discontinuity, encompassing both pathogenic and non-pathogenic strains across various phyla and lifestyles (Figure

2b). Notably, *Chlamydia trachomatis* and *M. tuberculosis* exhibited pronounced δ, indicating substantial shifts in genetic similarity. Conversely, *Helicobacter pylori* represented few instances where a species lacks a clear genetic discontinuity, suggesting a blurred genetic boundary possibly influenced by its evolutionary history and lifestyle.

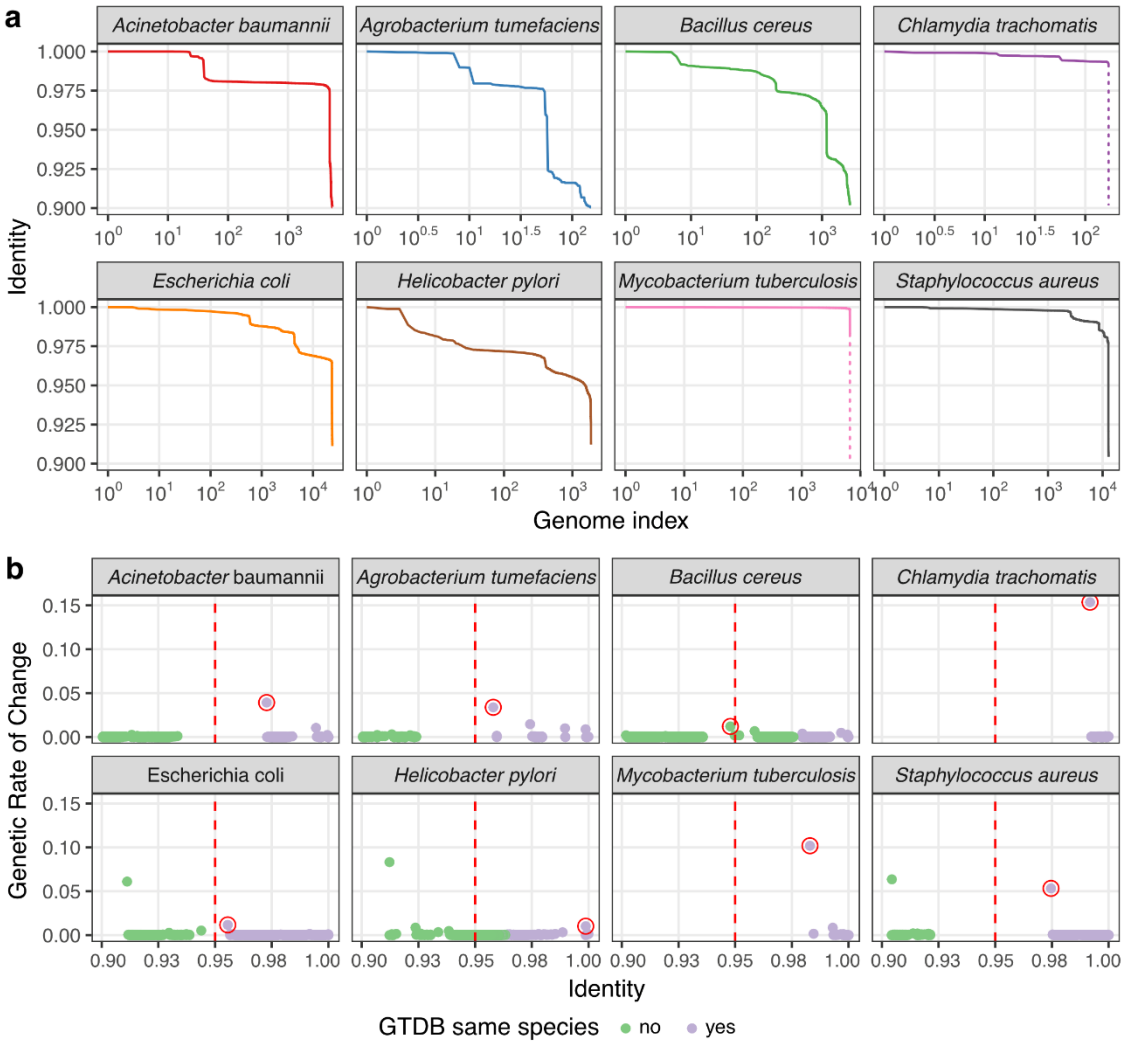

Figure 2. Genetic discontinuity properties of ten selected species. a) distribution of ranked genomic identities, revealing breakpoints around 95%. Vertical dashed lines for Chlamydia trachomatis and Mycobacterium tuberculosis indicate breaks beyond 90% from their closest genomes. The x-axis is represented in log-scale. **b)** Genetic Rate of Change is depicted across different identity values, with the break-associated point emphasized by a red circle. This point represents the key measure of genetic discontinuity (δ) examined in this work (see methods for comprehensive information). Genomes are colored based on GTDB classification.

Higher genetic discontinuity associates with allopatric lifestyle

To explore the ecological implications of bacterial discontinuity, we analyzed the pangenome of the 261 species mentioned above, which provides valuable insights into their lifestyles and evolution[19, 20]. The pangenome encompasses the core genome

(genes present in all isolates), the accessory genome (genes in more than one but not all isolates), and isolate-specific genes. Pangenome openness, measured by the saturation coefficient (α), indicates the extent to which new gene families are detected in the pangenome as more genomes are included. Higher α values suggest gene pool saturation (the addition of new genomes contributes fewer new detected genes), while lower α values imply a more flexible genomic repertoire.

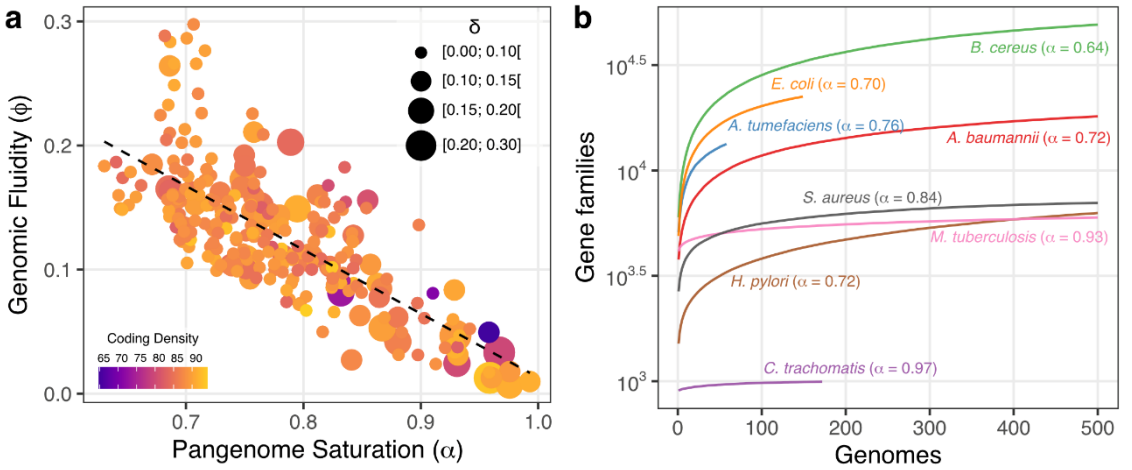

Figure 3. Pangenome properties of representative species. a) genomic fluidity in function of pangenome saturation with a linear regression line given by $\phi = -0.51148 \times \alpha + 0.52509$. The size of the dots corresponds to different levels of genetic discontinuity, grouped into four categories. Additionally, the color of the dots indicates the coding density of each representative genome. **b)** pangenome openness of ten selected species, with α highlighted for reference.

We used the pangenome openness to indirectly assess the species lifestyle[19] (Table 1). A low saturation coefficient (open pangenome) suggests a flexible genomic repertoire, characteristic of sympatric populations that frequently exchange genes to adapt to various environments. Conversely, a high saturation coefficient (closed pangenomes) is associated with allopatric populations adapted to specific niches with limited gene exchange due to physical isolation or genetic incompatibility[19, 20].

In our investigation, we identified a noteworthy correlation between pangenome openness and genomic fluidity (ϕ) (Figure 3a) – a measure of genomic dissimilarity at the gene level [21]. Specifically, we found a negative correlation, indicating that species with closed pangenomes exhibit lower genomic fluidity, as previously noted[22]. Furthermore, we observed a pronounced increase in genetic discontinuity as the pangenome saturation coefficient rises.

C. trachomatis and *Bacillus cereus* exhibit distinct pangenome characteristics that reflect their contrasting lifestyles. *C. trachomatis* exhibited a closed pangenome (α = 0.97), indicating a limited capacity for gene acquisition through HGT. This suggests a relatively stable genome and a more specialized lifestyle, features associated with an obligate intracellular pathogenic behavior[23]. This pattern is frequent among species with high genomic discontinuity, closed pangenomes, allopatric lifestyles, and highly

conserved pangenomes (Table 1). In contrast, *B. cereus* displays an open pangenome (α = 0.64), indicating a high propensity for gene acquisition and genomic diversity. This suggests a more versatile lifestyle, potentially enabling *B. cereus* to occupy various ecological niches and adapt to changing environments. The variations in pangenome openness observed in these two species provide valuable insights into their lifestyles and on how their pangenomes evolve in the context of genetic discontinuity.

Species	Lifestyle	δ	α	Core Prop.
Coxiella burnetii	Obligate intracellular	0.291	0.930	0.672
Treponema pallidum	Obligate pathogen	0.256	0.959	0.855
Mycoplasmoides pneumoniae	Obligate intracellular	0.228	0.967	0.755
Metamycoplasma hominis	Obligate intracellular	0.175	0.832	0.471
Chlamydia trachomatis	Obligate intracellular	0.154	0.976	0.895
Mycobacterium tuberculosis	Obligate pathogen	0.102	0.932	0.639
Staphylococcus aureus	Opportunistic pathogen	0.053	0.844	0.306
Acinetobacter baumannii	Opportunistic pathogen	0.039	0.073	0.152
Agrobacterium tumefaciens	Plant pathogen	0.034	0.764	0.312
Pseudomonas syringae	Plant pathogen	0.030	0.708	0.189
Rhizobium leguminosarum	Symbiont	0.024	0.704	0.230
Klebsiella pneumoniae	Opportunistic pathogen	0.016	0.730	0.161
Bacillus cereus	Free-living	0.012	0.640	0.085
Escherichia coli	Free-living	0.011	0.696	0.149
Helicobacter pylori	Free-living	0.010	0.725	0.189

Table 1: Genomic and ecological characteristics of 15 representative species. Genetic discontinuity (δ) represents the steepest change (break) in genomic identity distribution from a representative genome. The saturation coefficient (α) indicates pangenome openness: the higher the α value, the more closed the pangenome is. Core proportion refers to the ratio between the number of core genes and pangenome size of each species.

Uncovering most influential features to predict bacterial genetic discontinuity

We aimed evaluate the importance of different features in predicting bacterial genetic discontinuity. To address the asymmetrical nature of δ distribution (Figure 4a) and the impact of the number of genomes to estimate the pangenome openness, we employed a quantile regression that allowed us to assess the influence of pangenome saturation on genetic discontinuity while controlling for the number of genomes used to compute α. We found a significant impact across all quantiles examined (Figure 4b). Notably, as the quantile increased, the impact of pangenome saturation in δ became more pronounced. For instance, in the top quantile ($\tau = 0.95; [0.1040 - 0.291]$), an increase of one unit in alpha corresponded to a 0.44-unit rise in δ. These results shed light on the pivotal role of pangenome saturation in shaping genetic discontinuity patterns.

Beyond pangenome features, we also used a rich set of features ranging from taxonomical classification to orthology assessment to model bacterial discontinuity through a machine learning approach (see methods). Among the six tested methods, Linear Regression and Random Forest demonstrated superior performance in terms of root mean squared error, mean absolute error and quantile loss over different quantiles (Figure S1). Random Forest was chosen due to its ability to handle data distribution and collinearity without imposing strict assumptions. After hyperparameter tuning via k-fold cross-validation and grid search, we delved into feature importance using SHAP values to predict δ.

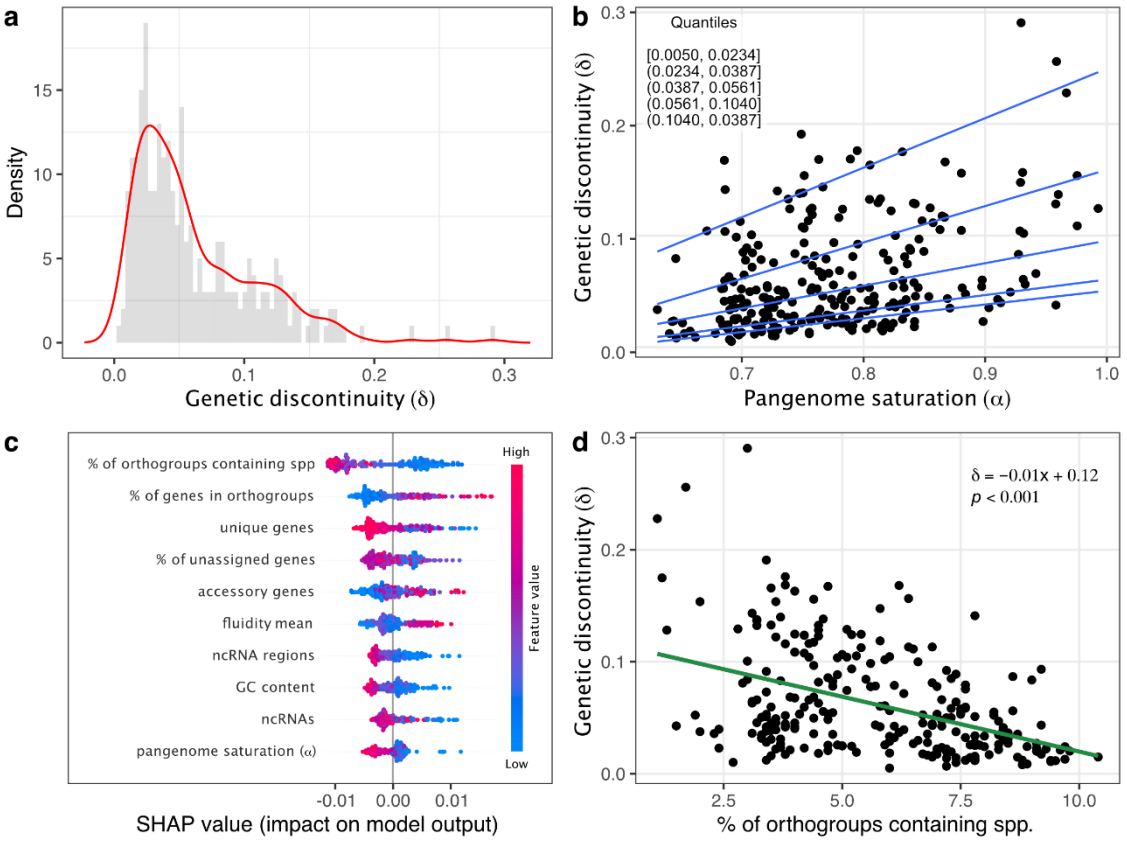

Figure 4. Genetic discontinuity modeling. a) Probability distribution of genetic discontinuity (δ) estimated from representative genomes. **b)** quantile regression analysis of δ as a function of pangenome saturation (α), highlighting a positive impact of α on δ across different quantiles (0.15, 0.25, 0.50, 0.75, 0.95). **c)** SHAP values derived from Random Forest Regression model, indicating the importance of ten features in predicting δ. Each representative genome is displayed, along with the positive or negative impact of each feature. **d)** linear regression between δ and the percentage of orthogroups containing species, the feature with the highest impact on predicting δ. The equation of the regression line and the associated p-value are also shown.

The most significant variable affecting the prediction of genetic discontinuity was the "Percentage of Orthogroups Containing Species" (Figure 4c). This metric gauges the proportion of orthogroups containing at least one gene from a given species. For example, if at least one gene from species A is present in 90 orthogroups from a total of 100, the percentage of orthogroups containing species A would be 90%. Moreover, this

feature negatively impacts δ (Figure 4d; p-value <0.001). This metric helps associate the presence and representation of a species within orthogroups with genetic discontinuity, as it indicates how frequently genes from that species are involved in shared functional contexts across different organisms.

Discussion

In this study, we systematically quantified genetic discontinuity (δ), which reflects significant shifts or abrupt changes in genetic similarity compared to a representative genome. Our work offers insights into the ecological roles of bacterial discontinuity and its implications for species classification. We carefully considered whether rigid taxonomic boundaries capture the fluidity and evolutionary dynamics that shape species' histories, especially considering organisms where genetic exchange, adaptation, and hybridization prevail[24, 25].

Philosophically, species classification raises the issue of Aristotelian essentialism – the idea that there are inherent qualities that define a species[26, 27]. The act of assigning species to specific categories becomes an exercise in grappling with the fundamental question of what defines a species. Is it solely genetic similarity, shared phenotypic traits, ecological niche, or something deeper that eludes our current understanding? In this work, we deployed genomics and ecology approaches to assess species boundaries. We used the essentialism idea as a prior to select representative genomes and explore whether the resulting genetic variation exhibited continuity or discreteness. To be agnostic about the choice of representative genomes, we employed a network approach where we could also retrieve understudied genomes to represent a given community, avoiding the limitation to explore species based only on well-studied type strains[18].

Bacterial genetic discontinuity has already been observed in studies comprising thousands of genomes by using different approaches[7, 8, 18]. The remaining inquiry was how to measure genetic discontinuity and unveil its ecological significance, while accounting for potential external factors such as sampling biases. To address this, we devised a novel metric (δ) by examining the maximum value in the first derivative distribution of genome identity derived from a representative genome. We observed delta values spanning from 0.005 (*Acinetobacter pittii*) to 0.290 (*Coxiella burnetii*) – species with remarkably distinct lifestyles. For the extreme case *C. burnetii,* the idea of $\delta = 0.29$ means that the most similar genome in a different community within a network of over 200 thousand genomes shares only 29% genetic identity with the representative genome of the *C. burnetii* community, which comprises 65 genomes. This identity value is way beyond of what we expect to distinguish species and approaches to those used to define higher taxonomical ranks such as genera and families[28, 29]. *C. burnetti* is an obligate intracellular pathogen responsible for causing the Q fever disease in humans[30]. Its allopatric lifestyle may explain its high genetic discontinuity.

After identifying breaks that varied according species lifestyles (Table 1), the next challenge was to assign an ecological meaning to them. The pangenome analysis is essential to gain insights into lifestyle and evolution a species[19, 20]. A key result we found here was that the greater the bacterial discontinuity, the more closed the pangenome was, always controlling for the number of genomes used to estimate the saturation coefficient. Besides *C. burnetti*, *M. tuberculosis* and *C. trachomatis* also illustrate situations where the magnitude of the break is related to lifestyles, especially regarding the HGT dynamics.

Conversely, species with ubiquitous or environmental lifestyles such as *E. coli* and *B cereus*, presented smaller breaks, but still well-defined boundaries. Those bacteria are known for their ability to thrive in various environments, including soil, water, and plant surfaces. Their diverse ecological niches might lead to more continuous genetic variation, exhibiting less pronounced breaks. In contrast, *Helicobacter pylori* posed an intriguing challenge with regards to genetic discontinuity. Our analysis revealed either an ambiguous or non-existent genetic break in this species, rendering it difficult to draw a line to delineate its boundaries. The absence of a discernible genetic break in *H. pylori* emphasizes the need for a more nuanced understanding of species boundaries and genetic cohesion, and suggests the presence of unique evolutionary dynamics that warrants further investigation.

Beyond the use of pangenome features to predict bacterial discontinuity, orthogroups assessment may be vital to understand genetic discontinuity. By using both the SHAP values and ExtraTreess Regressor to retrieve feature importance, the "percentage of orthogroups containing species", assigned by orthofinder, was the most important variable. This variable reveals insights into genetic interconnectedness and shared functions among bacterial species within an ecological niche. A higher percentage indicates shared traits due to the frequent co-occurrence, suggesting ecological overlap. Conversely, a lower percentage implies species specialization, indicating distinct ecological roles. For example, species with closed pangenomes such as *C. trachomatis* had genes present in only 2% of the total number of orthogroups.

In conclusion, our study highlights the significance of bacterial genetic discontinuity in understanding microbial diversity and evolution. We have shown that closed pangenomes and pronounced genetic breaks correspond to specific bacterial lifestyles, offering insights into microbial adaptation. Furthermore, our findings emphasize the role of orthogroups in characterizing genetic discontinuity and ecological dynamics within bacterial communities. This study contributes to a more nuanced understanding of bacterial diversity, emphasizing dynamic genetic relationships and the need to reevaluate traditional species classifications in microbial ecology.

Methods

Data collection and network analysis

We download a dataset comprising 210,129 genomes available on RefSeq as of September 2022. To ensure data quality, we retained genomes with fewer than 500 scaffolds and utilized the GTDB quality control[17] to exclude genomes displaying low quality or contamination. Filtered genomes were used to perform tens of billions of comparisons with mash v2.2.2 [16] to construct a weighted network (M) with igraph v1.5.1 [31]. M comprises genomic relationships among the set of genomes (g), with edges (e) corresponding to the genomic identity between pairs of genomes, quantified as the inverse of Mash distance (1 - Mash).

The weighted network was used to select representative genomes to infer the genetic identity patters. To select representative genomes, we subsetted the network M to create M':

$$M' = \{g, e' \mid e' \in e \; and \; e' \geq 0.95\}$$

Therefore, M' represents a species network containing the same set of genomes, but retaining edges above 95% identity, a threshold used to define species[4]. We employed the label propagation algorithm[32] to detect communities in M' that represent species. Only communities containing more than 50 genomes were used. This criterion was adopted to mitigate downstream modeling errors and enhance confidence in the biological significance of species representation.

Representative genomes of each species were chosen based on the following criteria: (i) prior designation as representative in both GTDB and RefSeq databases, (ii) representative at least in GTDB, or (iii) fewest scaffolds if the previous criteria were not met. Ties were resolved by random selection. This yielded a subset of 261 representative genomes (t) used for subsequent analysis.

Genome annotation, pangenome, and phylogenetic analysis

Genomes assigned to communities containing representative genomes were annotated using Bakta v1.5.1[33]. To ensure consistency in annotation, all genomes were reannotated within the same framework, mitigating potential discrepancies. In cases where communities exceeded 500 genomes, we implemented a downsampling strategy. Genomes showing 99.5% or higher identity were removed. For communities still surpassing this threshold, the remaining genomes were randomly selected.

We employed Panaroo v1.2.10[34] in the moderate mode to obtain the pangenome. The R packages Pagoo[35] and Micropan[36] were used to estimate the pangenome openness and genomic fluidity, respectively. We used Orthofinder v2.5.4[37] to obtain the orthogroups from representative genome communities. All single-copy genes were with Mafft v7.505[38] and concatenated to reconstruct the phylogenetic tree with IQ-Tree v2.1.4[39], incorporating ModelFinder[40] to identify the best fitting model.

One thousand bootstrap replicates were generated to assess the significance of internal nodes. The phylogenetic tree was visualized with ggtree[41].

Genetic discontinuity estimation (δ)

To estimate and quantify the genetic discontinuity across species, we employed an egocentric-based approach using representative genomes as "baits" within the network M. For each genome (i) in the representative subset (t), i served as an egocentric node to calculate its genomic distance ($d_i(g)$) from all other genomes g in M. These distances $d_i(g)$ were sorted in descending order to retrieve genomes most similar to the representative genome i. The sorted array (D_i) yielded a ranked list of genomic similarities. For each index j in D_i, the corresponding genomic identity ($I_i(j)$) was determined, quantifying the similarity between the representative genome i and the genome at index j in the array.

To assess the Genetic Rate of Change, we calculated the first derivative of the genomic identity ($I_i(j)$) with respect to j. The first derivative Δ_i was calculated as the change in genomic identity between two consecutive indices j and $j+1$ in D_i, divided by index difference:

$$\Delta_i = \frac{I_i(j) - I_i(j+1)}{j+1-j} = I_i(j) - I_i(j+1)$$

This rate of change Δ_i offered insights into how rapidly the genomic identity changed as we moved through the sorted array (D_i), offering a measure of variability in genomic similarity.

Finally, let δ represent the genetic discontinuity, reflecting the idea of a break or sharp change in the genetic identity between a species and its closest relative. For each species, δ was defined as the maximum Genetic Rate of Change above 94%. This threshold was adopted to exclude breaks representing higher taxonomical classifications, focusing solely on the species level. For instance, consider hypothetical genera G1 and G2: δ captures the steepest change in genomic identity from species in G1, not those from G1 to G2 (see Figure 1). Thus, δ can be calculated as:

$$\delta = \max\{\Delta_i | \, I_i \in [0.94, 1.00]\}$$

Additionally, we also incorporated GTDB species classification as a second prior to enhance accuracy in modeling δ for downstream analyses.

Machine Learning modeling

The outcome prediction task was formulated as a regression problem. We tested four different ML models to predict the genetic discontinuity (δ) for each species: Linear

Regression, Lasso Regression, Support Vector Regressor, Random Forest Regressor, and Gradient Booster Regressor. This analysis was employed using the Scikit-learn v1.0.2[42] and XGBoost v1.5.2[43] python[44] libraries.

We considered 261 species and 31 features categorized into four main groups: taxonomical, orthology-related, annotation-based, and pangenome metrics. Taxonomical features encompassed GTDB classification for each species (phylum, class, order, family, and genus). Ortholog-related features refers to six metrics obtained after detecting orthogroups for reference genomes (e.g., proportion of species-specific orthogroups). Categorical features with more than two categories were represented by a set of dummy variables, with one variable for each category.

Annotation-based features were retrieved from all 45,550 reannotated genomes. Continuous variables, including %GC content and coding density, were represented by their median value to account for their asymmetrical probability distribution. For discrete variables such as the number of CRISPR arrays, we calculated their relative frequency, indicating the likelihood of a species carrying such elements. This yielded 13 annotation-based features. Seven pangenome metrics, encompassing pangenome openness, genomic fluidity, and core genome proportion (core genes to pangenome size ratio), were included in the feature dataset.

Given the relatively low number of observations, the entire dataset was employed for training the model. We utilized the Extra Tree Regression feature selection method to reduce dimensionality, improve the estimator's accuracy, and boost the model performance. This algorithm employs randomized feature selection and ensemble averaging to make predictions, helping identify influential features, reduce overfitting, and enhance the model's performance[45]. Also, we adopted k-fold cross-validation to mitigate dataset size limitations and to evaluate model performance metrics (Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Quantile Loss).

To tune hyperparameters in the final ML model, we conducted a grid search with k-fold cross-validation utilizing 5 folds. We used the RMSE as the model score metric. We retrieved the importance of variables on explaining the model, by adopting the SHAP (*Shapley Additive exPlanations*) technique. the essence of SHAP is to measure the feature contribution of each individual to the outcome and whether the feature has a positive or negative impact on predictions[46].

Declaration of Competing Interest

The authors declare no conflict of interest.

Supplementary Information

Supplementary Tables can be found at: https://github.com/Passarelli-bio/Data_Bacterial_Discontinuity

Acknowledgments

We thank the Fulbright Brasil Commission for funding HP-A studies at Harvard T.H. Chan School of Public Health in 2022 and 2023. TMV research group is funded by Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ; grant E-26/201.117/2022), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES; Finance Code 001), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

1. Caro-Quintero, A. & Konstantinidis, K.T. Bacterial species may exist, metagenomics reveal. *Environ Microbiol* **14**, 347-355 (2012).

2. Cohan, F.M. Systematics: The Cohesive Nature of Bacterial Species Taxa. *Curr Biol* **29**, R169-R172 (2019).

3. Shapiro, B.J. *et al.* Population genomics of early events in the ecological differentiation of bacteria. *Science* **336**, 48-51 (2012).

4. Bobay, L.M. The Prokaryotic Species Concept and Challenges. In: Tettelin, H. & Medini, D. (eds). *The Pangenome: Diversity, Dynamics and Evolution of Genomes*: Cham (CH), 2020, pp 21-49.

5. Doolittle, W.F. & Papke, R.T. Genomics and the bacterial species problem. *Genome Biol* **7**, 116 (2006).

6. Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J. & Polz, M.F. A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. *Cell* **178**, 820-834 e814 (2019).

7. Knight, D.R. *et al.* Major genetic discontinuity and novel toxigenic species in Clostridioides difficile taxonomy. *Elife* **10** (2021).

8. Olm, M.R. *et al.* Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems* **5** (2020).

9. Passarelli-Araujo, H., Jacobs, S.H., Franco, G.R. & Venancio, T.M. Phylogenetic analysis and population structure of Pseudomonas alloputida. *Genomics* **113**, 3762-3773 (2021).

10. Jain, C., Rodriguez, R.L., Phillippy, A.M., Konstantinidis, K.T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114 (2018).

11. Hanage, W.P., Fraser, C. & Spratt, B.G. Sequences, sequence clusters and bacterial species. *Philos Trans R Soc Lond B Biol Sci* **361**, 1917-1927 (2006).

12. Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G. & Hanage, W.P. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323**, 741-746 (2009).

13. Hugenholtz, P., Chuvochina, M., Oren, A., Parks, D.H. & Soo, R.M. Prokaryotic taxonomy and nomenclature in the age of big sequence data. *ISME J* **15**, 1879-1892 (2021).

14. Potter, R.F., Burnham, C.D. & Dantas, G. In Silico Analysis of Gardnerella Genomospecies Detected in the Setting of Bacterial Vaginosis. *Clin Chem* **65**, 1375-1387 (2019).

15. Hill, J.E., Albert, A.Y.K. & Group, V.R. Resolution and Cooccurrence Patterns of Gardnerella leopoldii, G. swidsinskii, G. piotii, and G. vaginalis within the Vaginal Microbiome. *Infect Immun* **87** (2019).

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

17. Parks, D.H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **36**, 996-1004 (2018).

18. Passarelli-Araujo, H., Franco, G.R. & Venancio, T.M. Network analysis of ten thousand genomes shed light on Pseudomonas diversity and classification. *Microbiol Res* **254**, 126919 (2022).

19. Rouli, L., Merhej, V., Fournier, P.E. & Raoult, D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect* **7**, 72-85 (2015).

20. Brockhurst, M.A. *et al.* The Ecology and Evolution of Pangenomes. *Curr Biol* **29**, R1094-R1103 (2019).

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

22. Henaut-Jacobs, S., Passarelli-Araujo, H. & Venancio, T.M. Comparative genomics and phylogenomics of Campylobacter unveil potential novel species and provide insights into niche segregation. *Mol Phylogenet Evol* **184**, 107786 (2023).

23. Stelzner, K., Vollmuth, N. & Rudel, T. Intracellular lifestyle of Chlamydia trachomatis and host-pathogen interactions. *Nat Rev Microbiol* **21**, 448-462 (2023).

24. Arnold, B.J., Huang, I.T. & Hanage, W.P. Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol* **20**, 206-218 (2022).

25. Soucy, S.M., Huang, J. & Gogarten, J.P. Horizontal gene transfer: building the web of life. *Nat Rev Genet* **16**, 472-482 (2015).

26. Austin, C. Aristotelian essentialism: essence in the age of evolution. *Synthese* **194**, 2539-2556 (2017).

27. Shtulman, A. & Schulz, L. The relation between essentialist beliefs and evolutionary reasoning. *Cogn Sci* **32**, 1049-1062 (2008).

28. Deloger, M., El Karoui, M. & Petit, M.A. A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. *J Bacteriol* **191**, 91-99 (2009).

29. Qin, Q.L. *et al.* A proposed genus boundary for the prokaryotes based on genomic insights. *J Bacteriol* **196**, 2210-2215 (2014).

30. Seshadri, R. *et al.* Complete genome sequence of the Q-fever pathogen Coxiella burnetii. *Proc Natl Acad Sci U S A* **100**, 5455-5460 (2003).

31. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Systems**, 1695 (2006).

32. Raghavan, U.N., Albert, R. & Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **76**, 036106 (2007).

33. Schwengers, O. *et al.* Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb Genom* **7** (2021).

34. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* **21**, 180 (2020).

35. Ferres, I. & Iraola, G. An object-oriented framework for evolutionary pangenome analysis. *Cell Rep Methods* **1**, 100085 (2021).

36. Snipen, L. & Liland, K.H. micropan: an R-package for microbial pan-genomics. *BMC Bioinformatics* **16**, 79 (2015).

37. Emms, D.M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019).

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

39. Minh, B.Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020).

40. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. & Jermiin, L.S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587-589 (2017).

41. Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc Bioinformatics* **69**, e96 (2020).

42. Buitinck, L. *et al.* API design for machine learning software: experiences from the scikit-learn project. *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108-122 (2013).

43. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794 (2016).

44. Rossum, V. & Drake, F. Python 3 Reference Manual. Scotts Valley, California: CreateSpace; 2009.

45. Ferri, J., Pavel, P. & Hatef, M. Comparative Study of Techniques for Large-Scale Feature Selection. *Pattern Recognition in Practice, IV: Multiple Paradigms, Comparative Studies and Hybrid Systems* (2001).

46. Lundberg, S. & Lee, S. A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing Systems* (2017).

Supplementary Figure

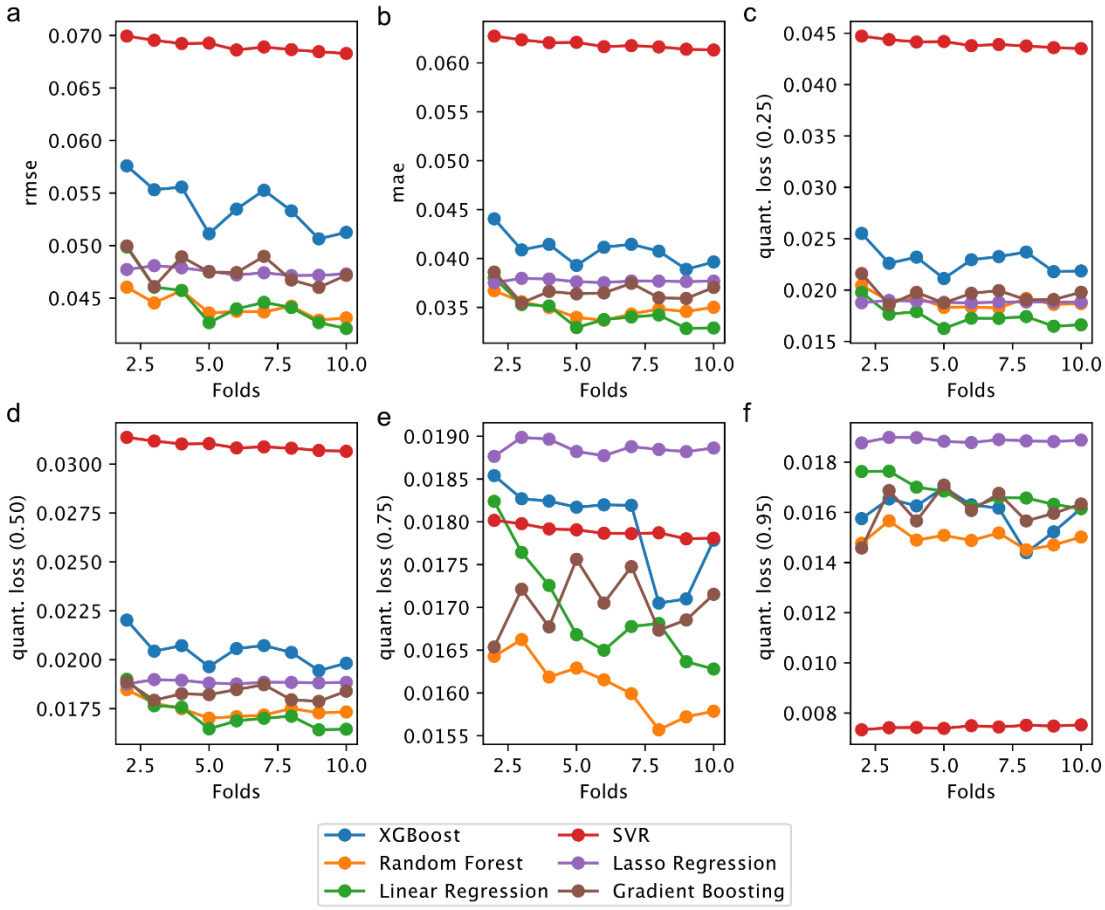

Figure S1. Model evaluation through k-fold cross validation using different metrics. a) Root Mean Squared Error (RMSE). **b)** Mean Absolute Error (MAE). **c-f)** Quantile Loss across four quantiles ($\tau = [0.25, 0.50, 0.75, 0.95]$).

4 Integrative Discussion

In this work, we quantified bacterial genetic discontinuity and explored how it can be perceived across different taxonomic levels. At the species level, we used the *P. alloputida* as a model and characterized its population structure for the first time. By using core genome Multilocus sequence typing (cgMLST) techniques and STRUCTURE ancestry simulations, we detected at least seven clonal complexes in the population. These clonal complexes represent genetically distinct groups of strains with specific ecological adaptations and evolutionary histories.

When analyzing genes of clinical and biotechnology interest, we observed that low-frequency resistance genes were usually found in plasmids and genes responsible for degrading aromatic compounds were likely transmitted horizontally. When we expanded the taxonomic level to explore the *P. putida* group, we observed that the ANI network was highly structured with easily discernible communities that would represent new species.

When we considered the distance of the more than 400 genomes in relation to the type strain for *P. alloputida* (Kh7^{T}), a clear break in distribution was detected, which motivated the exploration of this phenomenon in other species. Independently, Knight et. Al (2021)[18] also observed this genetic discontinuity around 95% identity using a different methodology, suggesting a broader applicability of this genetic discontinuity phenomenon.

As *Pseudomonas* is a genus that comprises bacteria ranging from those with clinical (*P. aeruginosa*), agricultural (*P. syringae*) and biotechnological (*P. putida*) interest, our next step was to explore how bacterial discontinuity could be perceived at the genus level. Therefore, from more than ten thousand genomes and different thresholds to define species, we found that networks maintained strong structures and facilitated community detection, emphasizing the persistence of genetic discontinuity.

Community sizes varied, with *P. aeruginosa* being the largest with 5116 genomes. However, 61 type strains (29.04%) existed as single nodes, underscoring the inadequacy of traditional reliance on type strains for capturing

genomic diversity within bacterial genera. Notably, the *Pseudomonas spp7* community, housing 122 genomes, may represent a new genomospecies.

Another key result of this article with relevance to the understanding of bacterial genetic discontinuity was that 25.65% of explored genomes were misclassified, highlighting the complex nature of the Pseudomonas genus, which comprises an admixture of other genera that needs to be further explored.

In our last article exploring bacterial genetic discontinuity, we used over 220,000 genomes to quantify and attribute the ecological significance for the genetic breaks. Also, we delve into the multifaceted dimensions of bacterial genetic discontinuity, elucidating its ecological, taxonomic, and evolutionary implications.

The ecological relevance of genetic discontinuity emerged in its association with microbial lifestyles, as seen in pangenome estimates and ecological studies. Closed pangenomes, indicative of stable niches, correlated with pronounced genetic breaks. This trend extended to species like *M. tuberculosis* and *Coxiella burnetii*, where genetic discontinuity reflected their distinct lifestyles. Defining bacterial species in light of genetic discontinuity poses a challenge, necessitating new approaches.

Moreover, the ecological implications of genetic discontinuity transcended species classification. The percentage of orthogroups containing species revealed interconnectedness and shared traits among bacterial species within niches, offering insights into ecological dynamics. Further exploration of this metric within the context of bacterial genetic discontinuity is essential. Collectively, these articles illustrate how genetic discontinuity can be observed across various taxonomic levels, enhancing our understanding of bacterial diversity and evolution.

5 Research Perspectives and Conclusion

As we advance our understanding of bacterial genetic discontinuity, several avenues for future research and challenges emerge. Further exploration of the role of horizontal gene transfer in shaping genetic discontinuity patterns across diverse bacterial taxa is warranted. Integrating ecological, genomic, and metagenomic approaches will offer a comprehensive understanding of genetic distribution in microbial communities. Taxonomic frameworks must also adapt to genomic diversity and lateral gene transfer.

In conclusion, our investigation into bacterial genetic discontinuity provides new insights to understand microbial ecology and evolution. By systematically quantifying genetic discontinuity patterns and exploring their ecological, taxonomic, and practical implications, we highlight the fascinating and unique nature of bacterial genetic diversity.

6 References

1. Shapiro, B.J., Leducq, J.B. & Mallet, J. What Is Speciation? *PLoS Genet* **12**, e1005860 (2016).

2. 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).

3. Arnold, B.J., Huang, I.T. & Hanage, W.P. Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol* **20**, 206-218 (2022).

4. Brockhurst, M.A. *et al.* The Ecology and Evolution of Pangenomes. *Curr Biol* **29**, R1094-R1103 (2019).

5. Hanage, W.P. Fuzzy species revisited. *BMC Biol* **11**, 41 (2013).

6. Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G. & Hanage, W.P. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323**, 741-746 (2009).

7. Jain, C., Rodriguez, R.L., Phillippy, A.M., Konstantinidis, K.T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114 (2018).

8. Luo, C. *et al.* Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proc Natl Acad Sci U S A* **108**, 7200-7205 (2011).

9. Goris, J. *et al.* DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **57**, 81-91 (2007).

10. Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J. & Polz, M.F. A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. *Cell* **178**, 820-834 e814 (2019).

11. Olm, M.R. *et al.* Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems* **5** (2020).

12. Deloger, M., El Karoui, M. & Petit, M.A. A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. *J Bacteriol* **191**, 91-99 (2009).

13. Bobay, L.M. The Prokaryotic Species Concept and Challenges. In: Tettelin, H. & Medini, D. (eds). *The Pangenome: Diversity, Dynamics and Evolution of Genomes*: Cham (CH), 2020, pp 21-49.

14. Coutinho, F.H. *et al.* Niche distribution and influence of environmental parameters in marine microbial communities: a systematic review. *PeerJ* **3**, e1008 (2015).

15. Layeghifard, M. *et al.* Microbiome networks and change-point analysis reveal key community changes associated with cystic fibrosis pulmonary exacerbations. *NPJ Biofilms Microbiomes* **5**, 4 (2019).

16. Passarelli-Araujo, H., Franco, G.R. & Venancio, T.M. Network analysis of ten thousand genomes shed light on Pseudomonas diversity and classification. *Microbiol Res* **254**, 126919 (2022).

17. Passarelli-Araujo, H., Jacobs, S.H., Franco, G.R. & Venancio, T.M. Phylogenetic analysis and population structure of Pseudomonas alloputida. *Genomics* **113**, 3762-3773 (2021).

18. Knight, D.R. *et al.* Major genetic discontinuity and novel toxigenic species in Clostridioides difficile taxonomy. *Elife* **10** (2021).