Phylogroup-specific variation shapes pangenome dynamics in *Pseudomonas aeruginosa*

João Botelho[1,2]*, Leif Tüffers[2,3], Janina Fuss[4], Florian Buchholz[2], Christian Utpatel[5], Jens Klockgether[6], Stefan Niemann[7], Burkhard Tümmler[6,8], Hinrich Schulenburg[1,2]*

[1]Antibiotic resistance group, Max-Planck Institute for Evolutionary Biology, Plön, Germany;

[2]Evolutionary Ecology and Genetics, University of Kiel, Kiel, Germany;

[3]Department of Infectious Diseases and Microbiology, University of Lübeck, Lübeck, Germany;

[4]Institute of Clinical Molecular Biology, Christian Albrechts University and University Hospital Schleswig-Holstein, Kiel, Germany;

[5]Molecular and Experimental Mycobacteriology, Research Center Borstel – Leibniz Lung Center, Borstel, Germany;

[6]Clinic for Paediatric Pneumology, Allergology, and Neonatology, Hannover Medical School (MHH), Hannover, Germany;

[7]Borstel Research Centre, National Reference Center for Mycobacteria, Borstel, Germany;

[8]Biomedical Research in Endstage and Obstructive Lung Disease Hannover (BREATH), German Center for Lung Research, Hannover Medical School, Hannover, Germany.

To whom correspondence should be addressed. Email: botelho@evolbio.mpg.de. Correspondence may also be addressed to hschulenburg@zoologie.uni-kiel.de.

ABSTRACT

Pseudomonas aeruginosa is a human opportunistic pathogen consisting of three phylogroups (hereafter named A, B, and C) of unevenly distributed size. Here, we assessed phylogroup-specific evolutionary dynamics in a collection of publicly available and newly sequenced genomes of *P. aeruginosa*. We explored to what extent antimicrobial resistance (AMR) genes, defence systems, and virulence genes vary in their distribution across regions of genome plasticity (RGPs) and "masked" (RGP-free) genomes, and to what extent this variation differs among the phylogroups. We found that members of phylogroup B possess larger genomes, contribute a comparatively larger number of gene families to the pangenome, but show lower abundance of CRISPR-Cas systems. Furthermore, AMR and defence systems are pervasive in RGPs and integrative and conjugative/mobilizable elements (ICEs/IMEs) from phylogroups A and B, and the abundance of these two types of cargo genes is often significantly correlated. Moreover, multiple inter- and intra-phylogroup interaction events occur at the accessory genome content level, suggesting that recombination events are frequent. Finally, we provide here a panel of phylogenetically diverse *P. aeruginosa* strains that can be used as a reference set for future functional analyses. Altogether, our results highlight distinct pangenome characteristics of the three phylogroups of *P. aeruginosa*, that are possibly influenced by variation in the abundance of CRISPR-Cas systems and that are shaped by the differential distribution of AMR and other defence systems.

INTRODUCTION

Pseudomonas aeruginosa is a ubiquitous metabolically versatile γ-proteobacterium. This Gram-negative bacterium is also an opportunistic human pathogen commonly linked to life-threatening acute and chronic infections (1). It belongs to the ESKAPE pathogens collection (2), highlighting its major contribution to nosocomial infections across the globe and its ability to “escape” antimicrobial therapy because of the widespread evolution of antimicrobial resistance (AMR) (3). This species is also often found to be multi- as well as extensively drug resistant (MDR and XDR, respectively) (4), making it difficult and in some cases even impossible to treat. For this reason, *P. aeruginosa* is placed by the World Health Organization (WHO) in the top priority group of most critical human pathogens, for which new treatment options are urgently required (5). These efforts rely on an in-depth understanding of the species biology and its evolutionary potential, which may be improved through a functional analysis of whole genome sequencing data.

The combined pool of genes belonging to the same bacterial species is commonly referred to as the pangenome. Frequently, only a small proportion of these genes is shared by all species members (the core genome). On the contrary, a substantial proportion of the total pool of genes is heterogeneously distributed across the members (the accessory genome). Following Koonin and Wolf (6), the pangenome can be divided into 3 categories: i) the persistent or softcore genome, for gene families present in the majority of the genomes; ii) the shell genome, for those present at intermediate frequencies and that are gained and lost rather slowly; iii) the cloud genome, for gene families present at low frequency in all genomes and that are rapidly gained and lost (7). Clusters of genes that are part of the accessory genome (i.e, the shell and cloud genome) are often located in so-called regions of genome plasticity (RGPs), genomic loci apparently prone to insertion of foreign DNA. By harbouring divergent accessory DNA in different strains, these loci can represent highly variable genomic regions. The shell and cloud genomes are also characterized by mobile genetic elements (MGEs) that are capable of being laterally transferred between bacterial cells, including plasmids, integrative and conjugative/mobilizable elements (ICEs/IMEs), and prophages (8, 9). These MGEs can mediate the shuffling of cargo genes that may provide a selective advantage to the recipient cell, such as resistance to antibiotics, increased pathogenicity, and defense systems against foreign DNA (10–12).

Most pangenome studies described to date have characterized gene frequencies across the whole species dataset without accounting for biased sampling or the population structure of the genomes in the dataset. This is particularly relevant in species consisting of multiple phylogroups with unevenly distributed members. As recently reported for *Escherichia coli* (13), genes classified as part of the accessory genome using traditional pangenome approaches are in fact core to specific phylogroups. Since *P. aeruginosa* is composed of three different-sized phylogroups (hereafter referred to as phylogroups A, B, and C as per the nomenclature proposed by Ozer *et al* (14)), characterized by high intraspecies functional variability (15, 16), it is likely that evolution in these phylogroups is driven by specific sets of genes found in the majority of members within the groups, but not across groups.

The aim of the current study is to enhance our understanding of the pangenome of the human pathogen *P. aeruginosa* by specifically assessing phylogroup-specific characteristics and genome dynamics, including data from more than 2000 genomes. We explore to what extent particular groups of cargo genes, such as those encoding AMR, virulence, and defence systems, vary in their distribution across RGPs and “masked” (RGP-free) genomes, and to what extent this variation differs among the phylogroups. Our data set includes new full genome sequences of a representative set of *P. aeruginosa* strains, the ‘major clone type strain panel’. This set of strains was previously isolated by the Tümmler lab (Hanover, Germany) from both clinical and environmental samples (17). It encompasses the most common clone types in the contemporary population (18–20) and provides a manageable, focused resource for in-depth functional analyses.

RESULTS

The *P. aeruginosa* phylogeny is composed of three phylogroups

Our phylogenomic characterization was based on 2009 assembled *P. aeruginosa* genomes, including 1991 publicly available genomes (following quality control and distance filtering, **Table S1**) and additionally 18 genomes for the ‘major clone type strain panel’ (**Table S2**) (17). Analysis of the softcore-genome alignment of these genomes identified three phylogroups, as previously reported (14, 21) (**Figure 1**). The two major reference isolates are part of the larger phylogroups: PAO1 (22) is part of phylogroup A (n=1531), while the PA14 strain falls into phylogroup B (n=435). Phylogroup C includes a substantially smaller number of members (n=43) (**Table S1**). Members of the phylogroup C were recently subdivided into either 2 (14) or 3 clusters, including the distantly related PA7 cluster (21). In this work, however, the PA7 cluster was excluded, and we focused our analysis on only the remainder of phylogroup C, since genomes from the PA7 cluster were too distantly related to the other genomes. In fact, the PA7 strain was first described as a taxonomic outlier of this species (23), and genomes belonging to this cluster were recently proposed to belong to a new *Pseudomonas* species (24). To test the impact of recombination on the softcore-genome alignment, we used ClonalFrameML to reconstruct the phylogenomic tree with corrected branch lengths. The segregation of *P. aeruginosa* into three phylogroups was maintained, resulting in a tree with decreased branch lengths and with identical number of members assigned to each phylogroup (**Figure S1**). Genomes from the Tümmler panel sequenced in this study were positioned in the wider context of the *P. aeruginosa* population and were widely distributed, with 12 strains in phylogroup A, 5 in phylogroup B, and 1 in phylogroup C (**Figure 1** and **Table S2**). Our results show that *P. aeruginosa* consists of three asymmetrical phylogroups and that the segregation of the 2009 genomes into phylogenetically distinct groups is not an artefact of recombination.

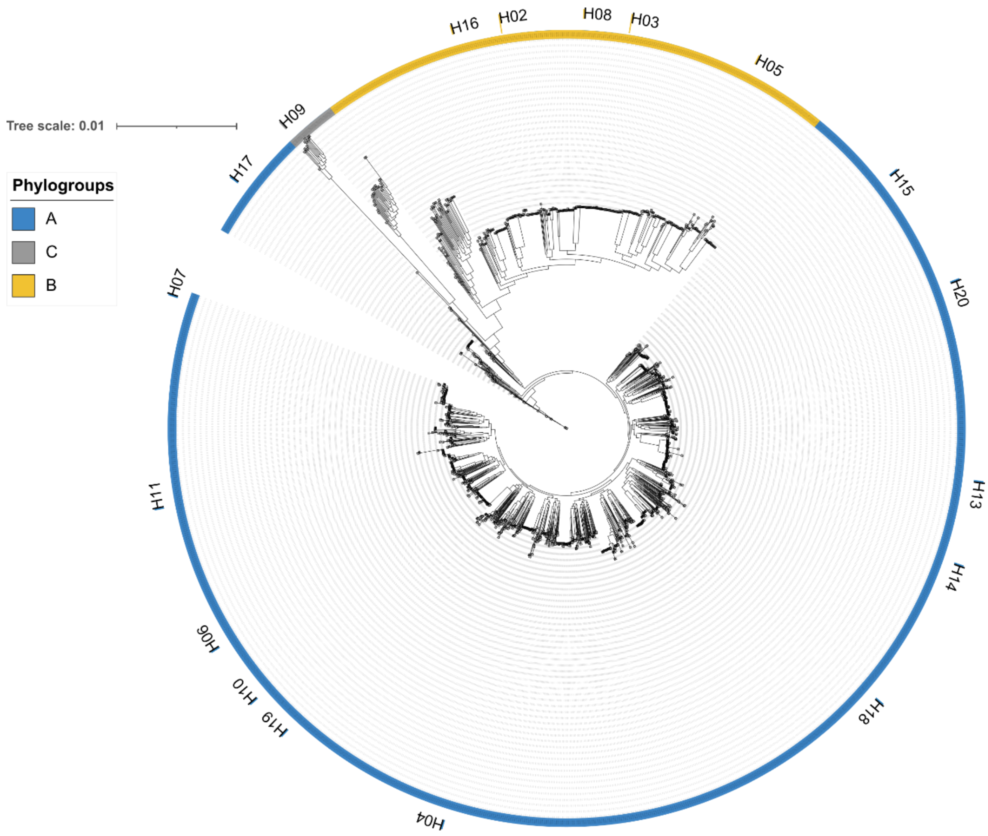

Figure 1. Maximum-likelihood tree of the softcore-genome alignment of all *P. aeruginosa* isolates used in this study (n=2009). The scale bar represents the genetic distance. Arcs in blue represent phylogroup A, yellow B, and grey C. The phylogenetic placement of the 'major clone type strain panel' sequenced in this study are highlighted in the tree, with the strain name next to strips coloured according to the phylogroup.

Phylogroup B contributes comparatively more gene families to the pangenome than the other two phylogroups

We next built a pangenome for the whole species, and separate pangenomes for each of the three phylogroups. This latter approach is important to take phylogenetic subdivisions of the species into account, which is additionally critical because the three phylogroups in our collection have substantially different sample sizes. We observed that the number of persistent gene families in the larger phylogroups A and B were similar to those found in the whole species, while the phylogroup C contained a substantially smaller number of persistent gene families (**Table S3**).

The pangenome of bacterial species is usually classified in two types: open pangenomes and closed ones (25). Since *P. aeruginosa* is an example of a bacterial species with open pangenome (14), i.e., the sequencing of new genomes will increase pangenome size, we explored the contribution of each phylogroup to the pangenome. To ensure comparability among the three phylogroups in our first analysis, we randomly drew 43 genomes from each phylogroup (thus, including the total sample size of the smallest phylogroup C), and observed that there is more diversity in the accessory genes of phylogroup B with regarding to the functions contributed by the acquired genes (**Figure 2A** and **Table S4**). In our second analysis, we focused only on the two larger phylogroups A and B, for which we randomly drew in each case 100 genomes and found the trend unchanged (**Figure 2B** and **Table S5**).

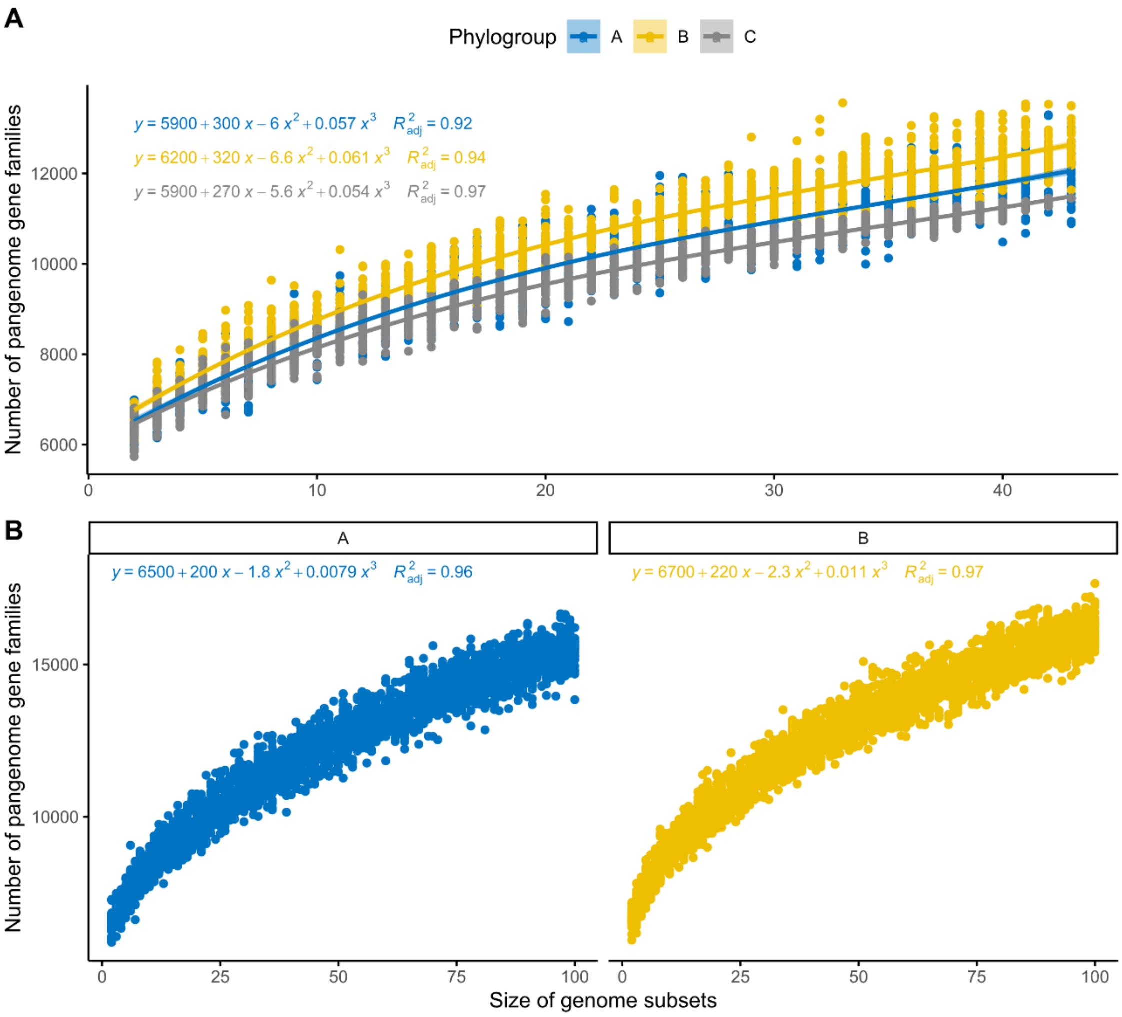

Figure 2. Rarefaction curves of the pangenome gene families for each phylogroup. All curves were inferred using polynomial regression lines. Curves in blue represent phylogroup A, yellow B, and grey C. **A)** The curves were generated by randomly re-sampling 43 genomes from each phylogroup several times and then plotting the average number of pangenome families found on each genome. **B)** Rarefaction curves were plotted with 100 random genomes from phylogroups A and B.

We then explored if specific gene families are pervasive across single or multiple phylogroups. We found 14 phylogroup-specific softcore gene families in phylogroup C, and two gene families that are exclusively found in the softcore genomes of phylogroups A and B (**Figure S2** and **Table S6**). Most gene families uniquely found on the softcore genome of phylogroup C were part of the Xcp type-II secretion system (T2SS), which is one of two complete and functionally distinct T2SS present in this species (**Table S7**). The Xcp system is encoded in a cluster containing 11 genes (*xcpP–Z*), as well as an additional *xcpA/pilD* gene found elsewhere in the genome (26). These genes were also found in the majority of the genomes from phylogroups A and B (**Table S8**), but the encoded proteins were too distantly related to those from phylogroup C. A similar pattern was observed for the two gene families indicated exclusively for phylogroups A and B, for which we also found distantly related orthologues in phylogroup C. Altogether, these results highlight that phylogroup B differs from the other two by contributing a comparatively larger number of gene families to the pangenome, possibly suggesting that phylogroup B members have larger genomes.

Phylogroup B genomes are significantly larger and most carry no CRISPR-Cas systems

A comparison of genome lengths revealed significantly larger genome sizes for phylogroup B than the other two phylogroups (**Figure 3A**, p-value < 2.2e-16). We then extracted the RGPs from each phylogroup, and found a total of 57901 RGPs across the three phylogroups. After removing the RGPs, we noted that the resulting "masked" genomes from phylogroup B are still significantly larger than those from the other two phylogroups (**Figure 3B**, p-value < 2.2e-16), pointing to a potentially higher number of genes conserved across genomes from this phylogroup. Still, the difference in genome size between phylogroups A and B is mainly explained by differences in accessory genome size (**Figure S3**). Masked genomes from phylogroup C are significantly smaller than genomes from the other two phylogroups, which is consistent with the smaller number of persistent gene families identified in this phylogroup (**Table S3**). We also explored the difference in GC content between genomes from different phylogroups, and we observed that the GC content from genomes in phylogroup B is significantly lower than genomes from other phylogroups (**Figure S4**, p-value < 2.2e-16).

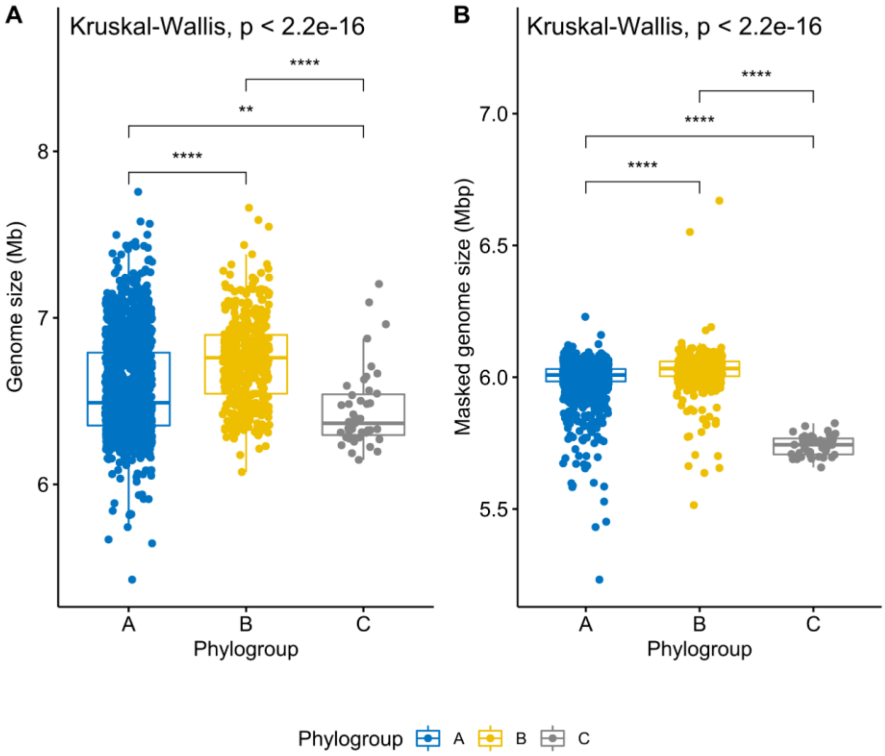

Figure 3. Boxplots representing the variation in genome size **(A)** and masked genome size **(B)** across the three phylogroups. Values above 0.05 were considered as non-significant (ns). Stars indicate significance level: * p <= 0.05, ** p <= 0.01, *** p <= 0.001, and **** p <= 0.0001. Boxplots in blue represent phylogroup A, yellow B, and grey C.

We next assessed whether presence of the defence CRISPR-Cas system is associated with genome size variation. Since CRISPR-Cas systems are important to defend bacteria against foreign DNA (12, 27), we expected that genomes carrying these systems would be smaller, while those devoid of these systems would accumulate mobile elements and hence be larger. We subdivided genomes from each of the three phylogroups into two groups depending on whether they contain or lack CRISPR-Cas systems, respectively (CRISPR-Cas^{pos}, CRISPR-Cas^{neg}). We indeed found that genomes with CRISPR-Cas systems are significantly smaller than those without (**Figure 4A**, p-values 8.3e-05 and 0.00025 for the phylogroup A and B comparisons, respectively), supporting the hypothesis that CRISPR-Cas systems can constrain horizontal gene transfer in *P. aeruginosa* (28–30). While the number of CRISPR-Cas^{pos} and CRISPR-Cas^{neg} genomes in phylogroups A and C is evenly distributed, phylogroup B genomes without CRISPR-Cas (n=279) were nearly two times more prevalent than those that carried these systems (n=156, **Table S1**). Interestingly, we observed that the difference between CRISPR-Cas^{pos} and CRISPR-Cas^{neg} in phylogroup B was no longer significant when only considering the masked genome size (**Figure 4B**). In line with this finding, we observed that the cumulative size of all RGPs was higher in genomes without CRISPR-Cas systems across phylogroups A and B (**Figure S5**). The absence of these defence systems in most genomes from phylogroup B may help to explain the observed larger size.

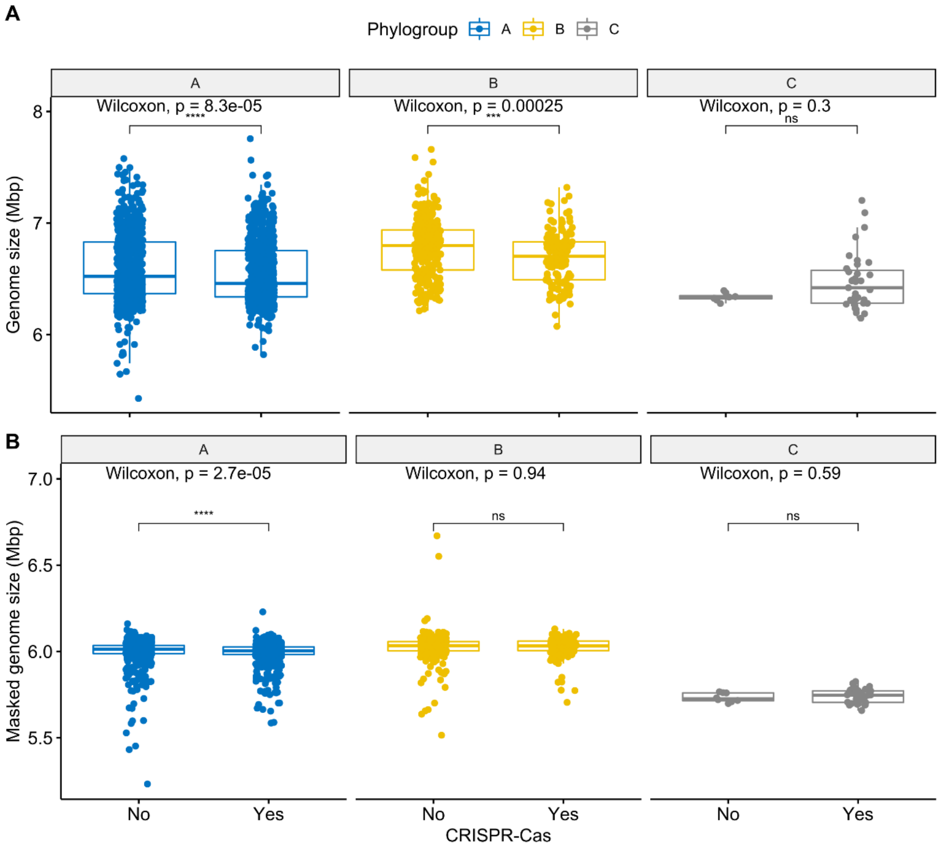

Figure 4. Boxplots representing the variation in genome size **(A)** and masked genome size **(B)** across pairs of conspecific genomes from the same phylogroup with and without CRISPR-Cas systems. Values above 0.05 were considered as non-significant (ns). Stars indicate significance level: * p <= 0.05, ** p <= 0.01, *** p <= 0.001, and **** p <= 0.0001. Boxplots in blue represent phylogroup A, yellow B, and grey C.

A wider diversity of CRISPR-Cas systems was found in genomes from phylogroup A, including I-C, I-E, I-F, IV-A1, and IV-A2 (**Figure S6** and **Table S9**). These CRISPR-Cas subtypes were all found in genomes from phylogroup B, with the exception of the IV-A2. Curiously, only subtypes I-E and I-F were present in genomes from phylogroup C. Type IV CRISPR-Cas systems were found almost exclusively on plasmids, and recent work revealed that they participate in plasmid–plasmid warfare (12, 31). The type I-C CRISPR–Cas subtype is typically encoded on ICEs and is also involved in competition dynamics between mobile elements (29, 32). Overall, our findings show that phylogroup B genomes are significantly larger and have a wider pool of accessory genes than those from the other two phylogroups, possibly driven by the lower prevalence of CRISPR-Cas systems in phylogroup B.

AMR and defence systems are overrepresented in RGPs from phylogroups A and B

We next sought to compare the relative frequency of proteins encoded in RGPs from different phylogroups. We observed that most functional categories are conserved across phylogroups. However, proteins coding for replication, recombination and repair functions are more prevalent in phylogroups A and B RGPs than those from phylogroup C (**Figure 5A**). Since these proteins are frequently involved in mobilization, this finding may suggest that genomes in these phylogroups have more functional mobile elements, with the ability to be horizontally transferred, while the RGPs in phylogroup C may be derived from remnants of mobile elements that can no longer be mobilized.

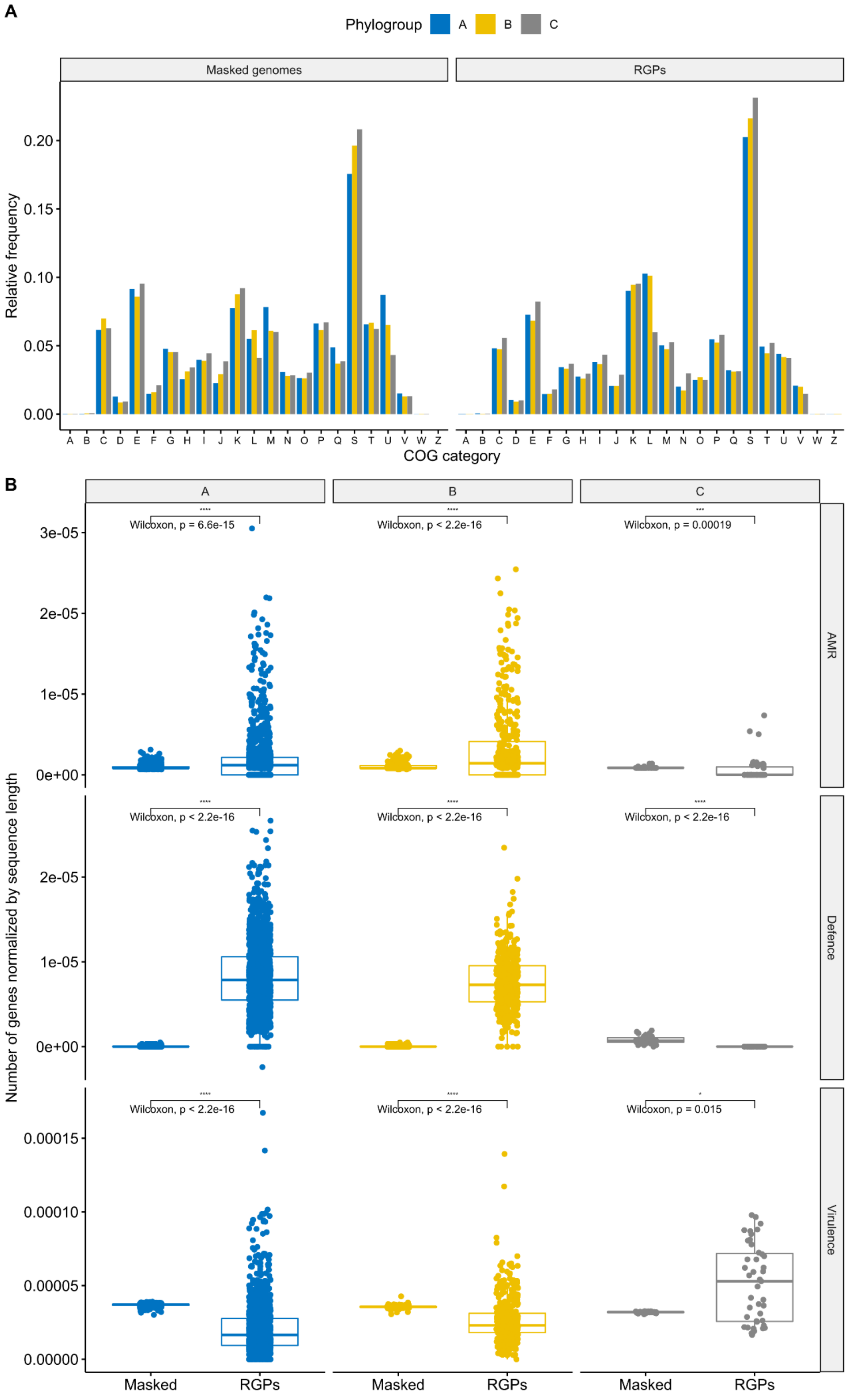

Figure 5. Distribution of functional categories across RGPs and masked genomes from the different phylogroups. Bar and boxplots in blue represent phylogroup A, yellow B, and grey C. **A.** Relative frequencies of cluster of orthologous groups categories. The relative frequencies were calculated by dividing the absolute counts for each category by the total number of clustered proteins found in each of the six groups. The functional categories are indicated by capital letters, including: A, RNA processing and modification; B, chromatin structure and dynamics; C, energy production and conversion; D, cell cycle control and mitosis; E, amino acid metabolism and transport; F, nucleotide metabolism and transport; G, carbohydrate metabolism and transport; H, coenzyme metabolism; I, lipid metabolism; J, translation; K, transcription; L, replication, recombination and repair; M, cell wall/membrane/envelop biogenesis; N, cell motility; O, post-translational modification, protein turnover, chaperone functions; P, inorganic ion transport and metabolism; Q, secondary structure; R, general functional prediction only; S, function unknown; T, signal transduction; U, intracellular trafficking and secretion; V, defence mechanisms; W, extracellular structures; Z, cytoskeleton. **B** Boxplots of the variation in the number of AMR genes, defence systems, and virulence genes found in RGPs and masked genomes across the three phylogroups. Absolute counts of genes and systems were normalized to RGP and masked genome sequence lengths in each strain. Values above 0.05 were considered as non-significant (ns). Stars indicate significance level: * p <= 0.05, ** p <= 0.01, *** p <= 0.001, and **** p <= 0.0001.

We next assessed to what extent RGPs and masked genomes vary in prevalence of genes for three types of functions, which are often encoded on MGEs, including virulence, defence systems, and AMR. Since the cumulative size of all RGPs is substantially smaller than that of masked genomes (**Table S1**), the number of virulence genes, defence systems, and AMR genes were normalized to the sequence length of the RGPs and masked genomes for each strain. We observed that the gene prevalence for these functions is conserved across masked genomes from different phylogroups, while virulence genes, defence systems, and AMR genes are unevenly distributed in RGPs. (**Figure 5B**).

Two important virulence factors were only present in some genomes from phylogroup C, and absent from the other two phylogroups (**Table S8**). These genes (*exlA* and *exlB*) encode hemolysins, and when genomes from phylogroup C carry these genes, the typical T3SS machinery found in most bacteria (encoding the toxins ExoS, ExoY, ExoT, and ExoU) is absent from these genomes, supporting previous reports that these are mutually exclusive (33). In agreement with previous findings (14), we further found that two important genes encoding type-III secretion system effector proteins (*exoS* and *exoU*) were unevenly distributed across the phylogroups: the *exoS* gene was pervasive among genomes from phylogroup A (99.5%, 1524/1531) and the majority of phylogroup C strains (28/43), while the *exoU* gene was overrepresented in genomes from phylogroup B (408/435) and nearly absent in genomes from the other two phylogroups (**Table S8**). As expected (34), some virulence genes were exclusively found on RGPs (i.e., absent from masked genomes): flagellar-associated proteins *fleI/flag*, *flgL*, *fliC* and *fliD*, as well as *wzy*, which codes for an O-antigen chain length regulator. All these virulence genes were found in RGPs from both phylogroups.

In agreement with the important role of MGEs as vectors for AMR genes in *P. aeruginosa* (9, 35), we found that AMR genes are overrepresented in RGPs from phylogroups A and B (**Figure 5B, Figure S7**). We then calculated the relative proportion of different AMR classes across RGPs from the three phylogroups, and found that most AMR classes were overrepresented across RGPs from phylogroup B (**Figure S8**). This result is consistent with our finding that RGPs play a significant role in the larger genome sizes from this phylogroup (**Figure S3**). Point mutations linked to resistance to beta-lactams and quinolones were observed for all phylogroups (**Table S10**).

A wide array of defence systems with a patchy distribution in closely related and distantly related strains was recently characterized in *P. aeruginosa*, suggesting high rates of horizontal gene transfer (36). According to this hypothesis, we would expect to observe an abundance of defence systems in RGPs, when compared with masked genomes. Similar to our results for AMR genes, we found that defence systems are indeed overrepresented in RGPs from phylogroups A and B (**Figure 5B**). Defence systems such as the globally distributed restriction-modification and CRISPR-Cas systems were common in RGPs from both phylogroups. Some rarer systems such as cyclic-oligonucleotide-based anti-phage signalling systems (CBASS) (37), Zorya, Gabija, Druantia (38), abortive infection (39), and bacteriophage exclusion (BREX) (40) were also observed in RGPs from phylogroups A and B (**Figure S9** and **Table S11**). In contrast, dGTPases were absent from both phylogroups. Our results reveal that AMR and defence systems are pervasive in RGPs from phylogroups A and B, and the majority of AMR classes are overrepresented in RGPs from phylogroup B.

AMR and defence systems are prevalent in ICEs/IMEs from phylogroups A and B

Given that the distribution and clustering of defence systems in *P. aeruginosa* is not dependent on the phylogenetic distance between all strains (36), and considering the high prevalence of ICEs/IMEs in this species (41), we explored the potential role of these elements as defence islands. To accurately detect these MGEs, we focused our analysis on complete genomes. We noted that 12.6% of our collection consisted of complete genomes (254/2009), including 172 genomes from phylogroup A, 78 from phylogroup B, and 4 genomes from phylogroup C (**Table S1**). 215 out of the 254 complete genomes harboured a total of 477 ICEs and 76 IMEs (**Table S12**). These ICEs/IMEs were present in 136 genomes from phylogroup A, 77 from phylogroup B, and 2 from phylogroup C. Thus, ICEs/IMEs were pervasive in strains from phylogroup B (77/78) and in the majority of strains from phylogroup A (136/172).

Nearly half of the ICEs/IMEs carried at least one AMR gene (228/553), with the ciprofloxacin-modifying *crpP* gene and the sulphonamide-resistance *sul1* gene being most frequent (**Table S13**). Indeed, the *crpP* gene was recently shown to be widely dispersed across ICEs from *P. aeruginosa* (42). Around one third of the ICEs/IMEs (193/553) carried at least one defence system, resulting in a total of 250 defence systems across the ICEs/IMEs and including 27 different types (**Figure S10** and **Table S13**). The most frequently found defence subtypes were CBASS-III and restriction-modification type-II (37, 39). Virulence genes were present in a smaller proportion of the ICEs/IMEs (99/553) and showed higher variation in abundance across ICEs/IMEs than AMR genes and defence systems (**Figure S11**). The *exoU* encoding for the effector protein and the *spcU* gene encoding for its chaperone were the most frequent virulence genes, all in ICEs/IMEs from phylogroup B (**Table S13**).

We next explored to what extent the prevalence of these three functional groups is correlated across ICEs/IMEs from the two larger phylogroups A and B. We observed that genes encoding resistance to fluoroquinolones were negatively correlated with genes involved in resistance to other antibiotic classes, and also with specific defence systems as restriction-modification and CBASS (**Figure 6A**). ICEs/IMEs from phylogroup B carrying fluoroquinolone-encoding resistance genes were also negatively associated with genes from the type-III secretion systems (**Figure 6B**). In contrast, genes encoding resistance to distinct antibiotic classes (e.g., beta-lactams, aminoglycosides, and sulphonamides) were often positively correlated in the ICEs/IMEs from both phylogroups, consistent with the previous observations that these genes tend to be co-localized in genetic structures named integrons (43). Virulence genes involved in flagellar motility were also often correlated, either additionally with (phylogroup B) or without (phylogroup A) genes involved in chemotaxis (44). Defence systems BREX and AbiEii (39, 40) were positively correlated in ICEs/IMEs from phylogroup B. Our results show that AMR and defence systems are densely populated in ICEs/IMEs from phylogroups A and B, and positive correlations between these functional groups can be observed in both phylogroups.

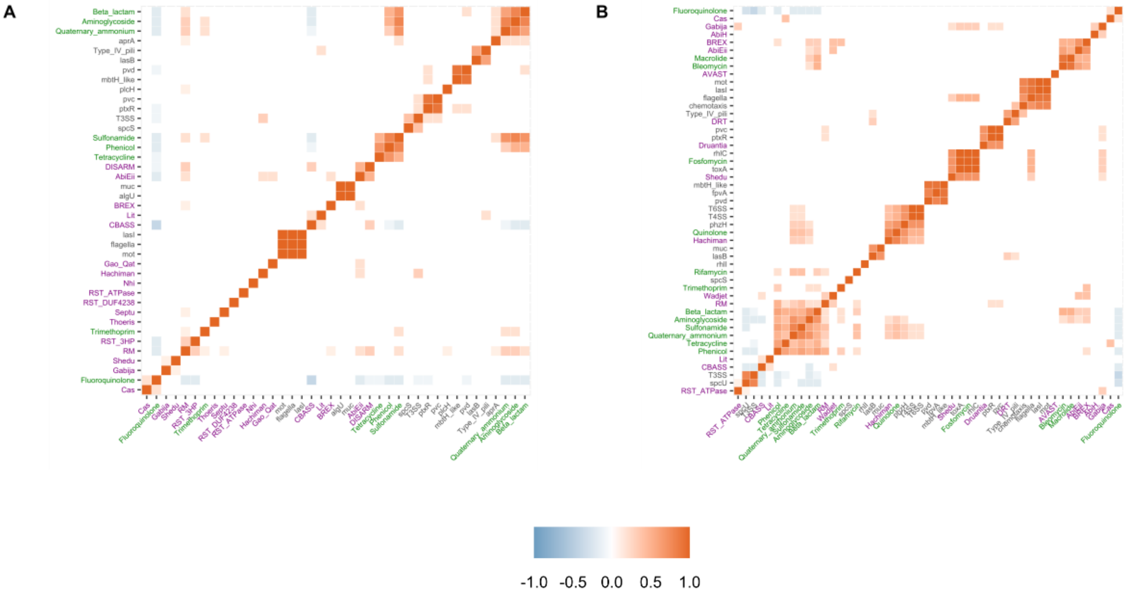

Figure 6. Correlation plots between AMR classes, virulence genes, and defence systems across ICEs/IMEs from phylogroup A **(A)** and phylogroup B **(B)**. The distribution of cargo genes across ICEs/IMEs was converted into a presence/absence matrix. Correlation matrices were

ordered using the hierarchical clustering function. Positive correlations are shown in different shades of red, while negative correlations are shown in different shades of blue. AMR genes and point mutations encoding resistance to particular AMR classes are part of the AMRFinder database (45), defence systems of defense-finder (36), and virulence genes of the VFDB (46). Virulence gene labels are coloured in black, AMR in green, and defence systems in purple.

Recombination events are pervasive between ICEs/IMEs from different phylogroups

We next used an alignment-free sequence similarity comparison of the ICEs/IMEs to infer an undirected network. The density plot showed a right-skewed distribution of pairwise distance similarities where the vast majority of ICE/IME pairs shared little similarity, with a Jaccard Index value below 0.5 (**Figure S12**), in accordance with the high diversity frequently observed across MGEs (47). To reduce the density and increase the sparsity of the network, we used the mean Jaccard Index between all pairs of RGPs as a threshold (0.12184). The network assigned 95.8% (530/553) of the ICEs/IMEs into 15 clusters (**Figure 7**). Nearly half of the ICEs/IMEs were grouped in cluster 1 (259/530, **Table S14**), which includes representatives of the three phylogroups.

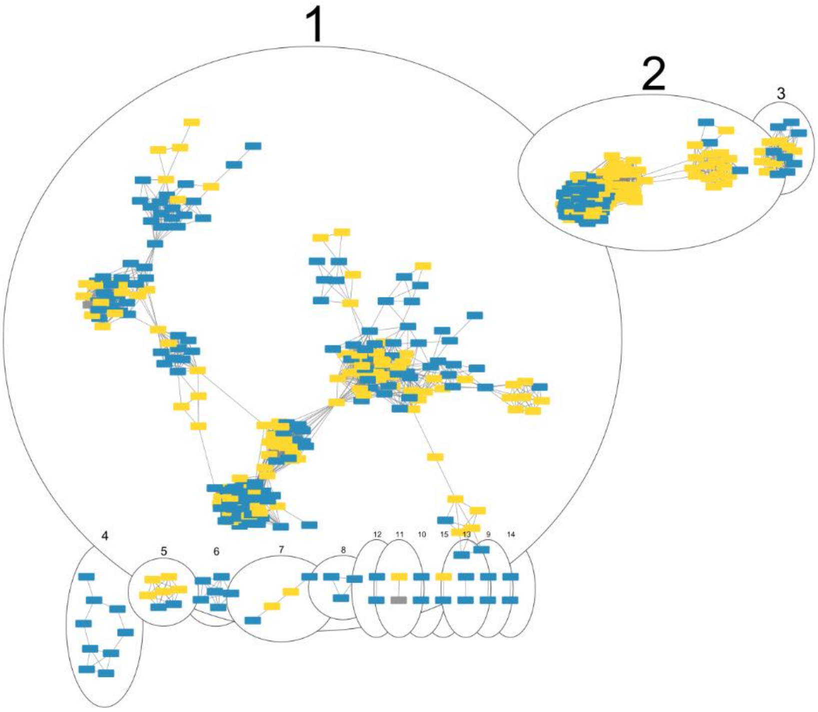

Figure 7. Network of clustered ICEs/IMEs from the three phylogroups, using the mean Jaccard Index between all pairs of ICEs/IMEs as a threshold. Each ICE/IME is represented by a node, connected by edges according to the pairwise distances between all ICE/IME pairs. Numbered ellipses represent ICEs/IMEs that belong to the same cluster. The network has a clustering coefficient of 0.794, a density of 0.099, a centralization of 0.217, and a heterogeneity of 0.785. ICEs/IMEs from phylogroup A are coloured in blue, from phylogroup B in yellow, and from phylogroup C in grey.

We then focused our analysis on the RGPs we extracted from all phylogroups (57901 RGPs in total). We filtered out RGPs smaller than 10kb, and we used the resulting 32744 RGPs to calculate the Jaccard Index between all pairs of RGPs. To reduce the density and increase the sparsity of the network, we used as a threshold the mean value (0.0919429) of the estimated pairwise distances between the 32744 RGPs identified in this study. The network assigned 99.7% (32651/32744) of the RGPs larger than 10kb into 51 clusters (**Figure S13**). While the majority of the RGP clusters were homogeneous for a given phylogroup, we also observed DNA sharing events between different phylogroups. These findings suggest that recombination events between the accessory genome is common between and within phylogroups in *P. aeruginosa*.

DISCUSSION

In this work, we explored the pangenome of the opportunistic human pathogen *P. aeruginosa* in consideration of its three main phylogroups. This approach allowed us to characterize defining properties of each phylogroup. In particular, we identified genes that are prevalent in the small phylogroup C, and absent from members of the two larger phylogroups. These genes would have been classified to be part of the accessory genome in conventional analysis of the pangenome of the species as a whole. In contrast, our refined approach suggests that these genes have an evolutionary advantage in a specific genetic context that is particular to this phylogroup (48). Moreover, phylogroup C is also clearly distinct from the other two phylogroups A and B in having a significantly smaller genome size and a low relative frequency of AMR and defence systems across RGPs. In addition, our results suggest that the larger accessory genome observed in phylogroup B is likely driven by the comparatively lower occurrence of CRISPR-Cas systems. Remarkably, genomes devoid of CRISPR-Cas systems in phylogroups A and B were significantly larger than those with these systems, a trend that was no longer observed when only considering the non-RGP (“masked”) genomes. This strongly supports the hypothesis that CRISPR-Cas systems can constrain horizontal gene transfer in *P. aeruginosa* (28–30, 49), at least for genomes belonging to the larger phylogroups.

The three phylogroups vary substantially in the distribution of AMR genes, defence systems, and virulence genes. This variation is particularly apparent in the separate analyses of RGPs and masked genomes. In particular, while the length of RGPs is substantially smaller than that of masked genomes, the absolute counts of most defence systems were higher in RGPs than in masked genomes across the three phylogroups (**Figure S9**). Curiously, representatives of the recently described set of defence systems that are part of Doron’s seminal study (38), such as Zorya, Wadjet, and Hachiman systems, were exclusively found in RGPs across the three phylogroups. In this study, the authors demonstrated that the Wadjet system provided protection against plasmid transformation in *Bacillus subtilis*, while the Zorya and Hachiman systems mediated defence against bacteriophages. These findings highlight the important role of RGPs in protecting genomes against infection by foreign DNA and their contribution to MGE-MGE conflict. Moreover, AMR and defence systems are rare in RGPs from phylogroup C, which may suggest that these strains are more often subjected to infection by foreign DNA. Assuming that there is no sampling bias across the three phylogroups, then the smaller number of phylogroup C members in public databases could thus be a consequence of the weaker arsenal of AMR and defence systems. Alternatively, phylogroup C strains may indeed be underrepresented, for example if they mainly occur in non-clinical habitats, which are usually less well sampled.

In general, our results underscore the role of ICEs/IMEs as vectors not only of AMR genes (35), but also of defence systems. Indeed, most of these systems show nonrandom clustering in defence islands and are often co-localized with mobilome genes (38, 50–52). Co-occurrence of genes alone, however, does not infer an ecological interaction between them (53). Recently, it was proposed that the accessory genome of the genus *Pseudomonas* is influenced by natural selection, showing a higher level of genetic structure than would be expected if neutral processes governed the pangenome formation (54). This suggests that coincident genes in ICEs/IMEs are more likely to act together for the benefit of the host or to ensure their own maintenance (9, 11). ICEs/IMEs, in particular, provide abundant material for the experimental study of bacterial defence systems. For example, SXT ICEs in *Vibrio cholerae*, which are also involved in AMR, consistently encode defence systems localized to a single hotspot of genetic shuffling (55). Additionally, ICEs in *Acidithiobacillia* carry type-IV CRISPR-Cas systems with remarkable evolutionary plasticity, which are often involved in MGE-MGE warfare (56). While size constraints were recently proposed as a possible justification for the low abundance of large defence systems on prophages (36), the larger size of ICEs/IMEs when compared with prophages (57) may explain the distribution of BREX and defence island system associated with restriction–modification (DISARM) (58) across our dataset (**Figure S10**). Even though the CBASS systems are not as prevalent as restriction-modification and CRISPR-Cas systems across the bacterial phylogeny (36), three types of this system were found across ICEs/IMEs from the larger phylogroups.

For our analyses, we used complete and draft genome assemblies retrieved from public databases. However, incomplete genome assemblies likely impact RGP definition, due to highly fragmented genomes, that might have inadvertently split RGPs into multiple contigs. With that in mind, we subsampled the complete genomes from our collection and used these to accurately delineate ICEs/IMEs. With the sequence similarity comparison between all pairs of ICEs/IMEs found in this study, as well as between all pairs of RGPs, we were able to explore interactions between these elements, suggesting that members of the same and of different phylogroups frequently undergo DNA shuffling events. Importantly, this network-based approach using pairwise genetic distances of alignment-free *k*-mer sequences between MGE pairs has bypassed the exclusion of non-coding elements, providing a more comprehensive picture of MGE populations and dynamics (29, 59).

To conclude, our work used a refined approach to explore phylogroup-specific and pangenome dynamics in *P. aeruginosa*. Members of phylogroup B contribute a comparatively larger number of pangenome families, have larger genomes, and have a lower prevalence of CRISPR-Cas systems. AMR and defence systems are pervasive in RGPs and ICEs/IMEs from phylogroups A and B, and these two functional groups are often significantly correlated, including both positive and negative correlations. We also observed multiple interaction events between the accessory genome content both between and within phylogroups, suggesting that recombination events are frequent. Finally, our work provides a representative set of phylogenetically diverse *P. aeruginosa* strains that can be used as a reference set for future functional analyses.

MATERIAL AND METHODS

Sequencing and hybrid assembly of the Hannover panel

Genomic DNA from 18 Hannover strains (17) were extracted using a Macherey-Nagel NucleoSpin Tissue kit, according to the standard bacteria support protocol from the manufacturer. We then used Nanodrop 1000 for quantification and control measures (260/280 and 260/230 ratios), followed by measurements in Qubit for a more precise quantification. The Agilent TapeStation tape and the FragmentAnalyzer Genomic DNA 50KB kit were used to control fragment size. Sequencing libraries were prepared using Illumina Nextera DNA flex and multiplexed SMRTBell libraries from PacBio. Libraries were sequenced on the Illumina MiSeq at 2x300bp or the PacBio Sequel II, respectively. Illumina reads were verified for quality using FastQC v0.11.9 (60) and trimmed with Trim Galore v0.6.6 (61), using the paired-end mode with default parameters and a quality Phred score cutoff of 10. Both datasets were then combined using the Unicycler v0.4.8 assembly pipeline (62). We used the default normal mode in Unicycler to build the assembly graphs of most strains, except H02, H14, H15, H18, and H19, where we used the bold mode. The assemblies were visually inspected using the assembly graph tool Bandage v0.8.1 (63).

Bacterial collection

We downloaded a total of 5468 *P. aeruginosa* genomes from RefSeq's NCBI database using PanACoTA v1.2.0 (64). After quality control to remove low-quality assemblies, 2704 were kept and 2764 genomes with more than 100 contigs were discarded (**Table S15**). Next, 713 genomes were discarded by the distance filtering step, using minimum (1e-4) and maximum (0.05) mash distance cut-offs to remove duplicates and misclassified assemblies at the species level (65), respectively. This resulted in 1991 publicly available genomes. The 18 genomes sequenced in this study from the Tümmler panel (17) passed both filtering steps, resulting in a pruned collection of 2009 genomes in total.

Pangenome and phylogenomics

We then used these genome sequences to generate a pangenome with the panrgp subcommand of PPanGGOLiN v1.1.136 (66, 67). We built a softcore-genome alignment (threshold 95%) and next used the alignment to infer a maximum likelihood tree with the

General Time Reversible model of nucleotide substitution in IQ-TREE v2.1.2 (68). To detect recombination events in our collection and account for them in phylogenetic reconstruction, we used ClonalFrameML v1.12 (69). Both trees were plotted in iTOL v6 (https://itol.embl.de/) (70). We then explored the phylogenies in order to cluster genomes according to the phylogroup. Separate pangenomes were performed for each phylogroup, using the panrgp subcommand from PPanGGOLiN. To classify core and accessory genes across genomes from different phylogroups, we used a publicly available R script (https://github.com/ghoresh11/twilight) (13). We used the gene presence/absence output from the whole collection's pangenome and the grouping of our genomes according to the phylogroup.

Identification of RGPs and ICEs/IMEs

To mask all the genomes, we used the RGPs coordinates determined by panrgp for each phylogroup as input in bedtools maskfasta v2.30.0 (71). We then used bedtools getfasta to extract the nucleotide sequence of the RGPs. All genomes were annotated with prokka v1.4.6 (72). To look for ICEs/IMEs on complete genomes, we used the genbank files created by prokka as input in the standalone-version of ICEfinder (73).

Identification of ICEs and functional categories

We retrieved the annotated proteins for the RGPs and masked genomes across the three phylogroups. We then clustered each of the six groups of proteins with MMseqs2 v13.45111 (74) and an identity cut-off of 80%. These clustered proteins were scanned for functional categories in eggNOG-mapper v2 (75), using the built-in database for clusters of orthologous groups (76). We then calculated the relative frequency of these categories by dividing the absolute counts for each category by the total number of clustered proteins found in each of the six groups. To look for CRISPR-Cas systems, we used CRISPRCasTyper v1.2.3 (77). We then used AMRFinder v3.10.18 (45) to look for AMR genes and resistance-associated point mutations. To look for virulence genes, we used the pre-downloaded database from VFDB (46) (updated on the 12-05-2021 and including 3867 virulence factors) in abricate v1.0.1 (https://github.com/tseemann/abricate). Finally, we used the protein sequenced generated by prokka as input in defense-finder v0.0.11 (36) to look for defence systems. This too searches for known defence systems in prokaryotic genomes, and includes systems with at least one experimental evidence of the defence function.

Network-based analysis of RGPs and ICEs/IMEs

To calculate the Jaccard Index between the RGPs, we used BinDash v0.2.1 (78) with *k*-mer size equal to 21 bp. We first used the sketch subcommand to reduce multiple sequences into one sketch, followed by the dist subcommand, to estimate distance (and relevant statistics) between RGPs in query sketch and RGPs in target-sketch. To calculate the Jaccard Index between the ICEs/IMEs, we also used BinDash. We used the mean() function in R to calculate the arithmetic mean of the Jaccard Index. Only Jaccard Index values equal to or above the mean were considered, and the mutation distances were used as edge attributes to plot the networks with Cytoscape v3.9.1 under the prefuse force directed layout (https://cytoscape.org/). We used the Analyzer function in Cytoscape to compute a comprehensive set of topological parameters, such as the clustering coefficient, the network density, the centralization, and the heterogeneity. To find clusters in our networks, we used the AutAnnotate and clusterMaker apps available in Cytoscape, and used the connected components as the clustering algorithm.

Statistical analysis

The correlation matrix was ordered using the hclust function in R. Multiple comparisons between the three phylogroups (e.g., genome size, GC content) were performed using the Kruskal-Wallis test. The unpaired two-sample Wilcoxon test was used in multiple comparisons between two independent groups of samples (RGPs vs. masked genomes, CRISPR-Cas positive vs negative genomes). In both tests, p-values were adjusted using the Holm–Bonferroni method. Values above 0.05 were considered as non- significant (ns). We used the following convention for symbols indicating statistical significance: * for p <= 0.05, ** for p <= 0.01, *** for p <= 0.001, and **** for p <= 0.0001.

DATA AVAILABILITY

Scripts for reproducing the analyses performed in this work are available at https://gitlab.gwdg.de/botelho/pa_pangenome. Analyses were made with a combination of shell and R 4.0.3 scripting. Sequencing performed in this project were deposited in NCBI under the Bioproject accession number PRJNA810040.

FUNDING

We acknowledge financial support from the German Science Foundation (grant SCHU 1415/12-2 to HS, funding under Germany's Excellence Strategy EXC 2167-390884018 as well as the Research Training Group 2501 TransEvo to HS and SN, and funding within the SFB 900 (TP A2) to BT), the Max-Planck Society (Max-Planck fellowship to HS), the Leibniz ScienceCampus Evolutionary Medicine of the Lung (EvoLUNG, to HS and SN), and the BMBF program Medical Infection Genomics (AZ 0315827A) to BT. This work was supported by the DFG Research Infrastructure NGS_CC (project 407495230) as part of the Next Generation Sequencing Competence Network (project 423957469). NGS was carried out at the Competence Centre for Genomic Analysis (Kiel). This research was supported in part through high-performance computing resources available at the Kiel University Computing Centre.

REFERENCES

1. Botelho,J., Grosso,F. and Peixe,L. (2019) Antibiotic resistance in *Pseudomonas aeruginosa* – Mechanisms, epidemiology and evolution. *Drug Resist. Updat.*, **44**, 100640.
2. De Oliveira,D.M.P., Forde,B.M., Kidd,T.J., Harris,P.N.A., Schembri,M.A., Beatson,S.A., Paterson,D.L. and Walker,M.J. (2020) Antimicrobial resistance in ESKAPE pathogens.

Clin. Microbiol. Rev., **33**, e00181-19.

3. Murray,C.J., Ikuta,K.S., Sharara,F., Swetschinski,L., Robles Aguilar,G., Gray,A., Han,C., Bisignano,C., Rao,P., Wool,E., *et al.* (2022) Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet*, **399**, 629–655.
4. Horcajada,J.P., Montero,M., Oliver,A., Sorlí,L., Luque,S., Gómez-Zorrilla,S., Benito,N. and Grau,S. (2019) Epidemiology and treatment of multidrug-resistant and extensively drug-resistant *Pseudomonas aeruginosa* infections. *Clin. Microbiol. Rev.*, **32**, e00031-19.
5. Tacconelli,E., Carrara,E., Savoldi,A., Harbarth,S., Mendelson,M., Monnet,D.L., Pulcini,C., Kahlmeter,G., Kluytmans,J., Carmeli,Y., *et al.* (2018) Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect. Dis.*, **18**, 318–327.
6. Koonin,E. V. and Wolf,Y.I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.*, **36**, 6688–6719.
7. Collins,R.E. and Higgs,P.G. (2012) Testing the Infinitely Many Genes Model for the Evolution of the Bacterial Core Genome and Pangenome. *Mol. Biol. Evol.*, **29**, 3413–3425.
8. Arnold,B.J., Huang,I.-T. and Hanage,W.P. (2021) Horizontal gene transfer and adaptive evolution in bacteria. *Nat. Rev. Microbiol.*, **20**, 206-218.
9. Botelho,J. and Schulenburg,H. (2020) The Role of Integrative and Conjugative Elements in Antibiotic Resistance Evolution. *Trends Microbiol.*, **29**, 8-18.
10. Partridge,S.R., Kwong,S.M., Firth,N. and Jensen,S.O. (2018) Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clin. Microbiol. Rev.*, **31**, e00088-17.
11. Rocha Id,E.P.C. and Id,D.B. (2022) Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLOS Biol.*, **20**, e3001514.
12. Pinilla-Redondo,R., Russel,J., Mayo-Muñoz,D., Shah,S.A., Garrett,R.A., Nesme,J., Madsen,J.S., Fineran,P.C. and Sørensen,S.J. (2021) CRISPR-Cas systems are widespread accessory elements across bacterial and archaeal plasmids. *Nucleic Acids Res.*, gkab859.
13. Horesh,G., Taylor-Brown,A., McGimpsey,S., Lassalle,F., Corander,J., Heinz,E. and Thomson,N.R. (2021) Different evolutionary trends form the twilight zone of the bacterial pan-genome. *Microb. Genomics*, **7**, 000670.
14. Ozer,E.A., Nnah,E., Didelot,X., Whitaker,R.J. and Hauser,A.R. (2019) The population structure of *Pseudomonas aeruginosa* is characterized by genetic isolation of *exoU*+ and *exoS*+ lineages. *Genome Biol. Evol.*, **11**, 1780-1796.
15. Trouillon,J., Imbert,L., Villard,A.-M., Vernet,T., Attrée,I. and Elsen,S. (2021) Determination of the two-component systems regulatory network reveals core and accessory regulations across *Pseudomonas aeruginosa* lineages. *Nucleic Acids Res.*, **49**, 11476-11490.
16. Trouillon,J., Han,K., Attrée,I., Lory,S. and Kook,H. (2022) The core and accessory Hfq interactomes across *Pseudomonas aeruginosa* lineages. *Nat. Commun.*, **13**, 1258.
17. Hilker,R., Munder,A., Klockgether,J., Losada,P.M., Chouvarine,P., Cramer,N., Davenport,C.F., Dethlefsen,S., Fischer,S., Peng,H., *et al.* (2015) Interclonal gradient of

virulence in the *Pseudomonas aeruginosa* pangenome from disease and environment. *Environ. Microbiol.*, **17**, 29–46.

18. Wiehlmann,L., Cramer,N. and Tümmler,B. (2015) Habitat-associated skew of clone abundance in the *Pseudomonas aeruginosa* population. *Environ. Microbiol. Rep.*, **7**, 955–960.
19. Wiehlmann,L., Wagner,G., Cramer,N., Siebert,B., Gudowius,P., Morales,G., Köhler,T., Van Delden,C., Weinel,C., Slickers,P., *et al.* (2007) Population structure of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 8101–8106.
20. Fischer,S., Dethlefsen,S., Klockgether,J. and Tümmler,B. (2020) Phenotypic and Genomic Comparison of the Two Most Common ExoU-Positive *Pseudomonas aeruginosa* Clones, PA14 and ST235. *mSystems*, **5**, e01007-20.
21. Freschi,L., Vincent,A.T., Jeukens,J., Emond-Rheault,J.-G., Kukavica-Ibrulj,I., Dupont,M.-J., Charette,S.J., Boyle,B. and Levesque,R.C. (2018) The *Pseudomonas aeruginosa* pan-genome provides new insights on its population structure, horizontal gene transfer and pathogenicity. *Genome Biol. Evol.*, **11**, 109-120.
22. Stover,C.K., Pham,X.Q., Erwin,A.L., Mizoguchi,S.D., Warrener,P., Hickey,M.J., Brinkman,F.S.L., Hufnagle,W.O., Kowalik,D.J., Lagrou,M., *et al.* (2000) Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature*, **406**, 959–964.
23. Roy,P.H., Tetu,S.G., Larouche,A., Elbourne,L., Tremblay,S., Ren,Q., Dodson,R., Harkins,D., Shay,R., Watkins,K., *et al.* (2010) Complete genome sequence of the multiresistant taxonomic outlier *Pseudomonas aeruginosa* PA7. *PLoS One*, **5**, e8842.
24. Morimoto,Y., Tohya,M., Aibibula,Z., Baba,T., Daida,H. and Kirikae,T. (2020) Re-identification of strains deposited as *Pseudomonas aeruginosa*, *Pseudomonas fluorescens* and *Pseudomonas putida* in GenBank based on whole genome sequences. *Int. J. Syst. Evol. Microbiol.*, **70**, 5958-5963.
25. Brockhurst,M.A., Harrison,E., Hall,J.P.J., Richards,T., McNally,A. and MacLean,C. (2019) The Ecology and Evolution of Pangenomes. *Curr. Biol.*, **29**, R1094–R1103.
26. Filloux,A. (2011) Protein secretion systems in *Pseudomonas aeruginosa*: An essay on diversity, evolution, and function. *Front. Microbiol.*, **2**, 155.
27. Koonin,E. V., Makarova,K.S., Wolf,Y.I. and Krupovic,M. (2019) Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat. Rev. Genet.*, **21**, 119-131.
28. Wheatley,R.M. and MacLean,R.C. (2020) CRISPR-Cas systems restrict horizontal gene transfer in *Pseudomonas aeruginosa*. *ISME J.*, **15**, 1420-1433.
29. Botelho,J., Cazares,A. and Schulenburg,H. (2022) The ESKAPE mobilome contributes to the spread of antimicrobial resistance and CRISPR-mediated conflict between mobile genetic elements. *bioRxiv*, 10.1101/2022.01.03.474784.
30. Pursey,E., Dimitriu,T., Paganelli,F.L., Westra,E.R. and Houte,S. van (2022) CRISPR-Cas is associated with fewer antibiotic resistance genes in bacterial pathogens. *Philos. Trans. R.*

Soc. B, **377**, 20200464.

31. Pinilla-Redondo,R., Mayo-Muñoz,D., Russel,J., Garrett,R.A., Randau,L., Sørensen,S.J. and Shah,S.A. (2020) Type IV CRISPR-Cas systems are highly diverse and involved in competition between plasmids. *Nucleic Acids Res.*, **48**, 2000–2012.
32. León,L.M., Park,A.E., Borges,A.L., Zhang,J.Y. and Bondy-Denomy,J. (2021) Mobile element warfare via CRISPR and anti-CRISPR in *Pseudomonas aeruginosa*. *Nucleic Acids Res.*, **49**, 2114–2125.
33. Reboud,E., Basso,P., Maillard,A.P., Huber,P. and Attrée,I. (2017) Exolysin Shapes the Virulence of *Pseudomonas aeruginosa* Clonal Outliers. *Toxins (Basel)*, **9**, 364.
34. Arora,S.K., Bangera,M., Lory,S. and Ramphal,R. (2001) A genomic island in *Pseudomonas aeruginosa* carries the determinants of flagellin glycosylation. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 9342–9347.
35. Botelho,J., Mourão,J., Roberts,A.P. and Peixe,L. (2020) Comprehensive genome data analysis establishes a triple whammy of carbapenemases, ICEs and multiple clinically relevant bacteria. *Microb. Genom.*, **6**, mgen000424.
36. Tesson,F., Herve,A., Touchon,M., d'Humieres,C., Cury,J. and Bernheim,A. (2021) Systematic and quantitative view of the antiviral arsenal of prokaryotes. *bioRxiv*, 10.1101/2021.09.02.458658.
37. Cohen,D., Melamed,S., Millman,A., Shulman,G., Oppenheimer-Shaanan,Y., Kacen,A., Doron,S., Amitai,G. and Sorek,R. (2019) Cyclic GMP–AMP signalling protects bacteria against viral infection. *Nature*, **574**, 691–695.
38. Doron,S., Melamed,S., Ofir,G., Leavitt,A., Lopatina,A., Keren,M., Amitai,G. and Sorek,R. (2018) Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, **359**, eaar4120.
39. Labrie,S.J., Samson,J.E. and Moineau,S. (2010) Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.*, **8**, 317–327.
40. Goldfarb,T., Sberro,H., Weinstock,E., Cohen,O., Doron,S., Charpak-Amikam,Y., Afik,S., Ofir,G. and Sorek,R. (2015) BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.*, **34**, 169–183.
41. Guglielmini,J., Quintais,L., Garcillán-Barcia,M.P., de la Cruz,F. and Rocha,E.P.C. (2011) The Repertoire of ICE in Prokaryotes Underscores the Unity, Diversity, and Ubiquity of Conjugation. *PLoS Genet.*, **7**, e1002222.
42. Botelho,J., Grosso,F. and Peixe,L. (2020) ICEs Are the Main Reservoirs of the Ciprofloxacin-Modifying *crpP* Gene in *Pseudomonas aeruginosa*. *Genes (Basel)*, **11**, 889.
43. Ghaly,T.M., Geoghegan,J.L., Tetu,S.G. and Gillings,M.R. (2020) The Peril and Promise of Integrons: Beyond Antibiotic Resistance. *Trends Microbiol.*, **28**, 455–464.
44. Matilla,M.A., Martín-Mora,D., Gavira,J.A. and Krell,T. (2021) *Pseudomonas aeruginosa* as a Model To Study Chemosensory Pathway Signaling. *Microbiol. Mol. Biol. Rev.*, **85**, e00151-20.
45. Feldgarden,M., Brover,V., Haft,D.H., Prasad,A.B., Slotta,D.J., Tolstoy,I., Tyson,G.H.,

Zhao,S., Hsu,C.-H., McDermott,P.F., *et al.* (2019) Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrob. Agents Chemother.*, **63**, e00483-19.

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

47. Cury,J., Oliveira,P.H., de la Cruz,F. and Rocha,E.P.C. (2018) Host Range and Genetic Plasticity Explain the Coexistence of Integrative and Extrachromosomal Mobile Genetic Elements. *Mol. Biol. Evol.*, **35**, 2230–2239.

48. Lassalle,F., Muller,D. and Nesme,X. (2015) Ecological speciation in bacteria: reverse ecology approaches reveal the adaptive part of bacterial cladogenesis. *Res. Microbiol.*, **166**, 729–741.

49. van Belkum,A., Soriaga,L.B., LaFave,M.C., Akella,S., Veyrieras,J.-B., Barbu,E.M., Shortridge,D., Blanc,B., Hannum,G., Zambardi,G., *et al.* (2015) Phylogenetic Distribution of CRISPR-Cas Systems in Antibiotic-Resistant *Pseudomonas aeruginosa*. *MBio*, **6**, e01796-15.

50. Makarova,K.S., Wolf,Y.I., Snir,S. and Koonin,E. V. (2011) Defense Islands in Bacterial and Archaeal Genomes and Prediction of Novel Defense Systems. *J. Bacteriol.*, **193**, 6039–6056.

51. Hussain,F.A., Dubert,J., Elsherbini,J., Murphy,M., VanInsberghe,D., Arevalo,P., Kauffman,K., Rodino-Janeiro,B.K., Gavin,H., Gomez,A., *et al.* (2021) Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science*, **374**, 488–492.

52. Vliet,A.H.M. van, Charity,O.J. and Reuter,M. (2021) A *Campylobacter* integrative and conjugative element with a CRISPR-Cas9 system targeting competing plasmids: a history of plasmid warfare? *Microb. Genom.*, **7**, 000729.

53. Blanchet,F.G., Cazelles,K. and Gravel,D. (2020) Co-occurrence is not evidence of ecological interactions. *Ecol. Lett.*, **23**, 1050–1063.

54. Whelan,F.J., Hall,R.J. and McInerney,J.O. (2021) Evidence for selection in the abundant accessory gene content of a prokaryote pangenome. *Mol. Biol. Evol.*, **38**, 3697-3708.

55. LeGault,K.N., Hays,S.G., Angermeyer,A., McKitterick,A.C., Johura,F., Sultana,M., Ahmed,T., Alam,M. and Seed,K.D. (2021) Temporal shifts in antibiotic resistance elements govern phage-pathogen conflicts. *Science*, **373**, eabg2166.

56. Moya-Beltrán,A., Makarova,K.S., Acuña,L.G., Wolf,Y.I., Covarrubias,P.C., Shmakov,S.A., Silva,C., Tolstoy,I., Johnson,D.B., Koonin,E. V., *et al.* (2021) Evolution of Type IV CRISPR-Cas Systems: Insights from CRISPR Loci in Integrative Conjugative Elements of *Acidithiobacillia*. *CRISPR J.*, **4**, 656-672.

57. Cury,J., Touchon,M. and Rocha,E.P.C. (2017) Integrative and conjugative elements and their hosts: composition, distribution and organization. *Nucleic Acids Res.*, **45**, 8943–8956.

58. Ofir,G., Melamed,S., Sberro,H., Mukamel,Z., Silverman,S., Yaakov,G., Doron,S. and

Sorek,R. (2017) DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat. Microbiol.*, **3**, 90–98.

59. Acman,M., van Dorp,L., Santini,J.M. and Balloux,F. (2020) Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.*, **11**, 2452.

60. Andrews,S. FastQC A Quality Control tool for High Throughput Sequence Data. *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*.

61. Babraham Bioinformatics - Trim Galore!, https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

62. Wick,R.R., Judd,L.M., Gorrie,C.L. and Holt,K.E. (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Comput. Biol.*, **13**, e1005595.

63. Wick,R.R., Schultz,M.B., Zobel,J. and Holt,K.E. (2015) Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics*, **31**, 3350–3352.

64. Perrin,A. and Rocha,E.P.C. (2021) PanACoTA: a modular tool for massive microbial comparative genomics. *NAR Genom. Bioinform.*, **3**, lqaa106-

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

66. Gautreau,G., Bazin,A., Gachet,M., Planel,R., Burlot,L., Dubois,M., Perrin,A., Médigue,C., Calteau,A., Cruveiller,S., *et al.* (2020) PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLOS Comput. Biol.*, **16**, e1007732.

67. Bazin,A., Gautreau,G., Médigue,C., Vallenet,D. and Calteau,A. (2020) panRGP: a pangenome-based method to predict genomic islands and explore their diversity. *Bioinformatics*, **36**, i651–i658.

68. Minh,B.Q., Schmidt,H.A., Chernomor,O., Schrempf,D., Woodhams,M.D., Von Haeseler,A., Lanfear,R. and Teeling,E. (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.*, **37**, 1530–1534.

69. Didelot,X. and Wilson,D.J. (2015) ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLOS Comput. Biol.*, **11**, e1004041.

70. Letunic,I. and Bork,P. (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, **49**, W293–W296.

71. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

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

73. Liu,M., Li,X., Xie,Y., Bi,D., Sun,J., Li,J., Tai,C., Deng,Z. and Ou,H.-Y. (2018) ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res.*, **47**, D660-D665.

74. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.

75. Cantalapiedra,C.P., Hernández-Plaza,A., Letunic,I., Bork,P. and Huerta-Cepas,J. (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.*, **38**, 5825-5829.

76. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E. V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.

77. Russel,J., Pinilla-Redondo,R., Mayo-Muñoz,D., Shah,S.A. and Sørensen,S.J. (2020) CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *Cris. J.*, **3**, 462–469.

78. Zhao,X. (2019) BinDash, software for fast genome distance estimation on a typical personal laptop. *Bioinformatics*, **35**, 671–673.

SUPPLEMENTARY FIGURES

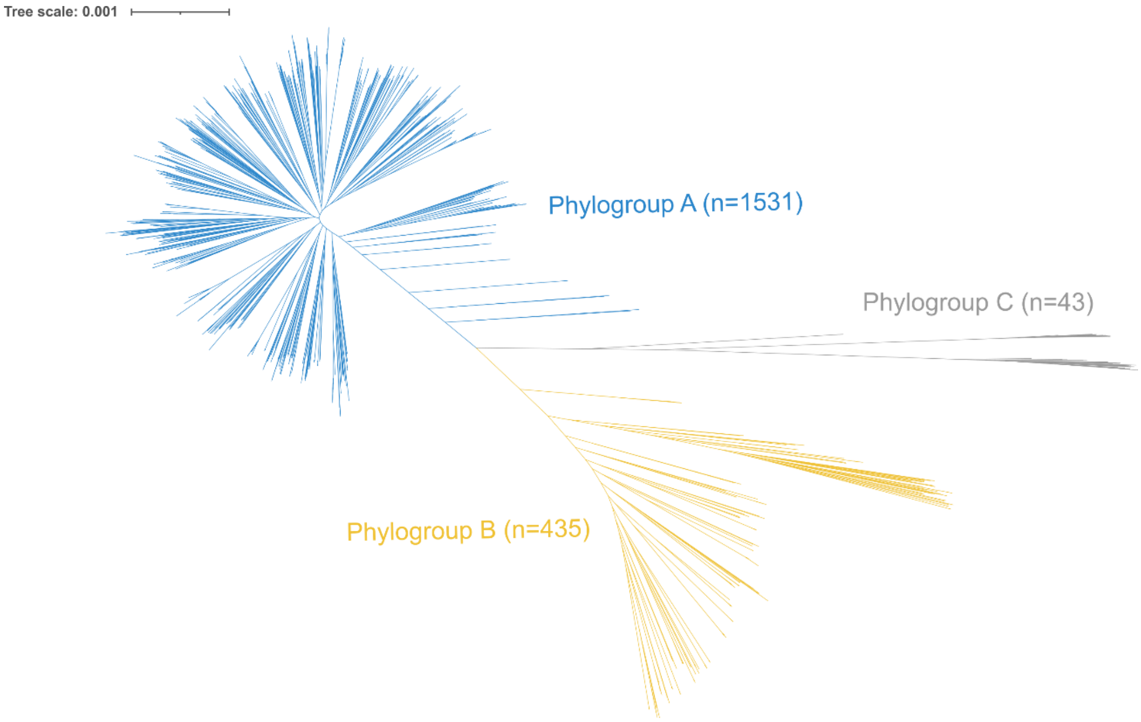

Figure S1. Maximum-likelihood tree of the softcore-genome alignment of all *P. aeruginosa* isolates used in this study (n=2009), corrected for recombination. The scale bar represents the genetic distance. Members of phylogroup A are coloured in blue, B in yellow, and C in grey.

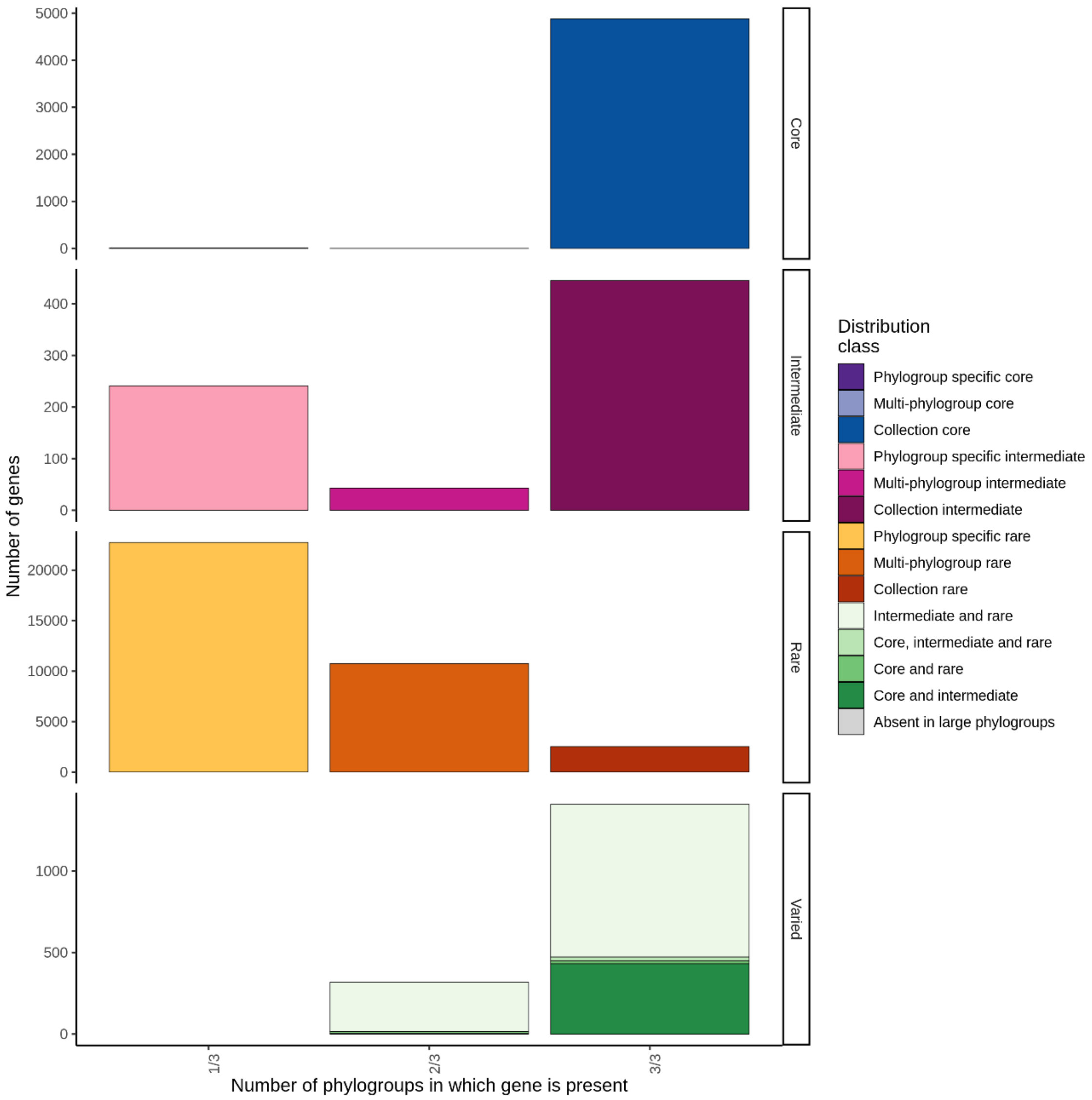

Figure S2. Barplots of the distribution of gene families into core, intermediate, rare, or varied parts of the pangenome across phylogroups. The first column shows genes that are specific to a given phylogroup, and further classified into core (≥95%), intermediate, rare (≤15%), or varied. The second column shows genes that are specific to two phylogroups, and their classification into core, intermediate, rare, or varied. The third column shows genes that are present across all three phylogroups, and their classification into core, intermediate, rare, or varied. A different colour is assigned to each classification. To create the plot, we modified the R script available in https://github.com/ghoresh11/twilight/blob/master/classify_genes.R.

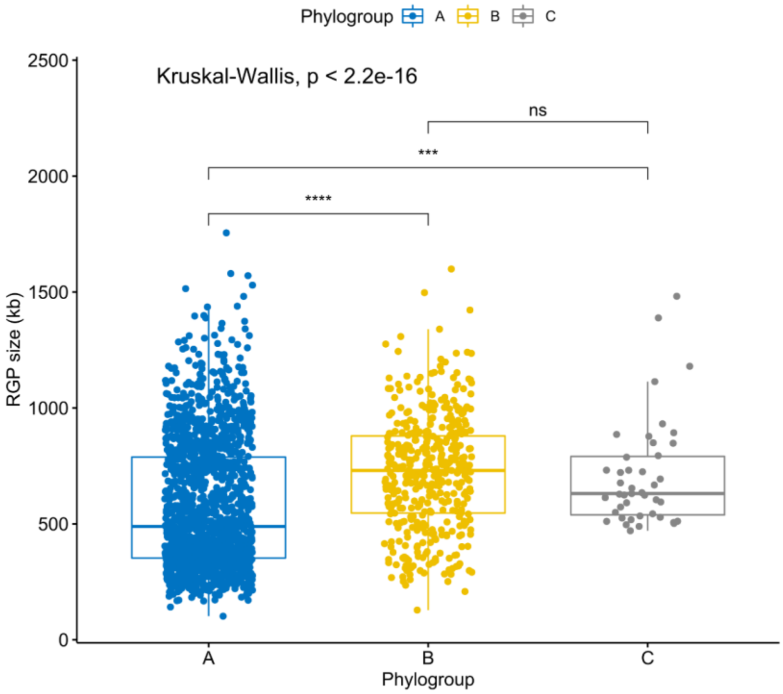

Figure S3. Boxplots showing the variation in RGP size across the three phylogroups. Values above 0.05 were considered as non-significant (ns). Stars indicate significance level: * p <= 0.05, ** p <= 0.01, *** p <= 0.001, and **** p <= 0.0001. Boxplots in blue represent phylogroup A, yellow B, and grey C.

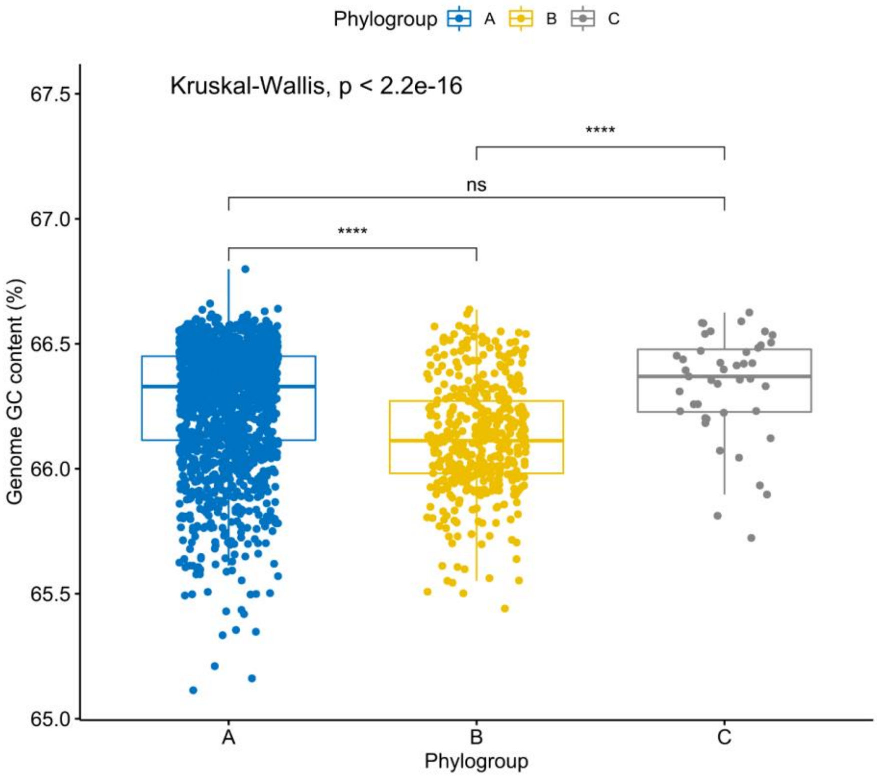

Figure S4. Boxplots showing the variation in GC content across the three phylogroups. Values above 0.05 were considered as non-significant (ns). Stars indicate significance level: * p <= 0.05, ** p <= 0.01, *** p <= 0.001, and **** p <= 0.0001. Boxplots in blue represent phylogroup A, yellow B, and grey C.

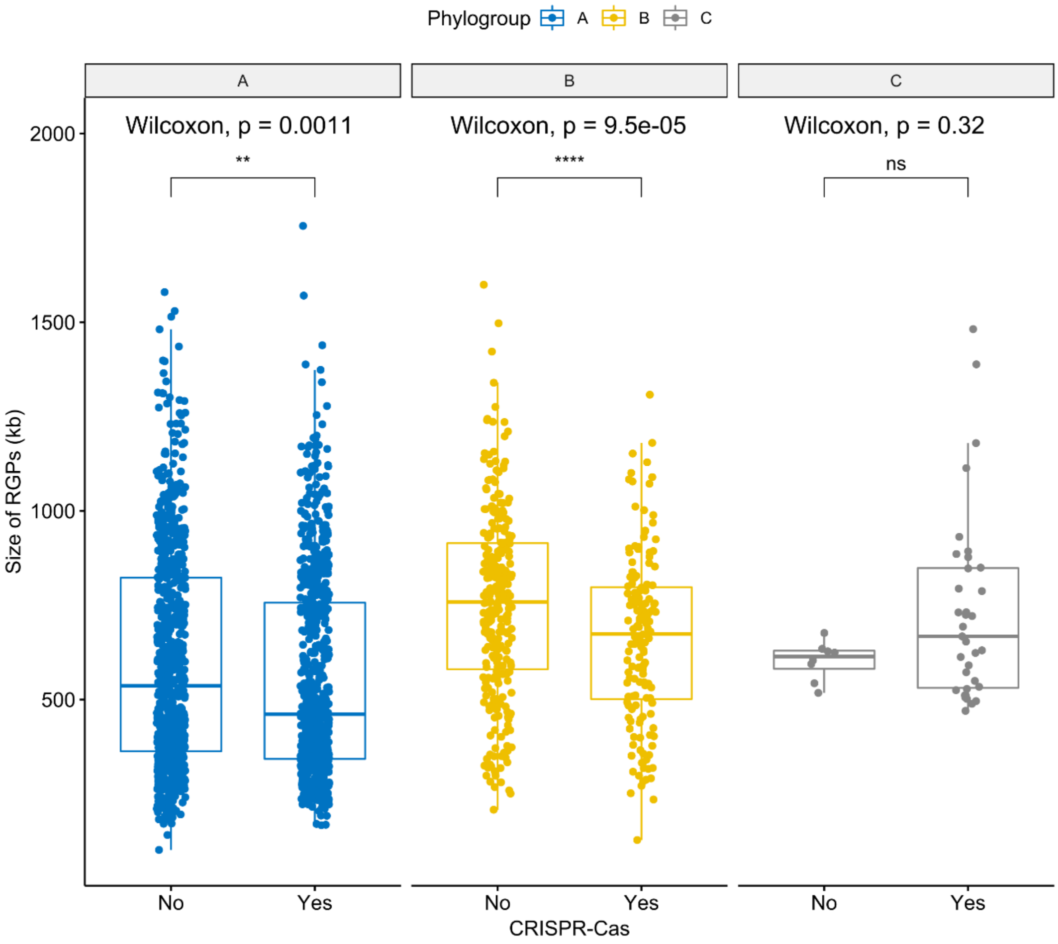

Figure S5. Boxplots representing the variation in the size of RGPs across pairs of conspecific genomes from the same phylogroup with and without CRISPR-Cas systems. Values above 0.05 were considered as non-significant (ns). Stars indicate significance level: * p <= 0.05, ** p <= 0.01, *** p <= 0.001, and **** p <= 0.0001. Boxplots in blue represent phylogroup A, yellow B, and grey C.

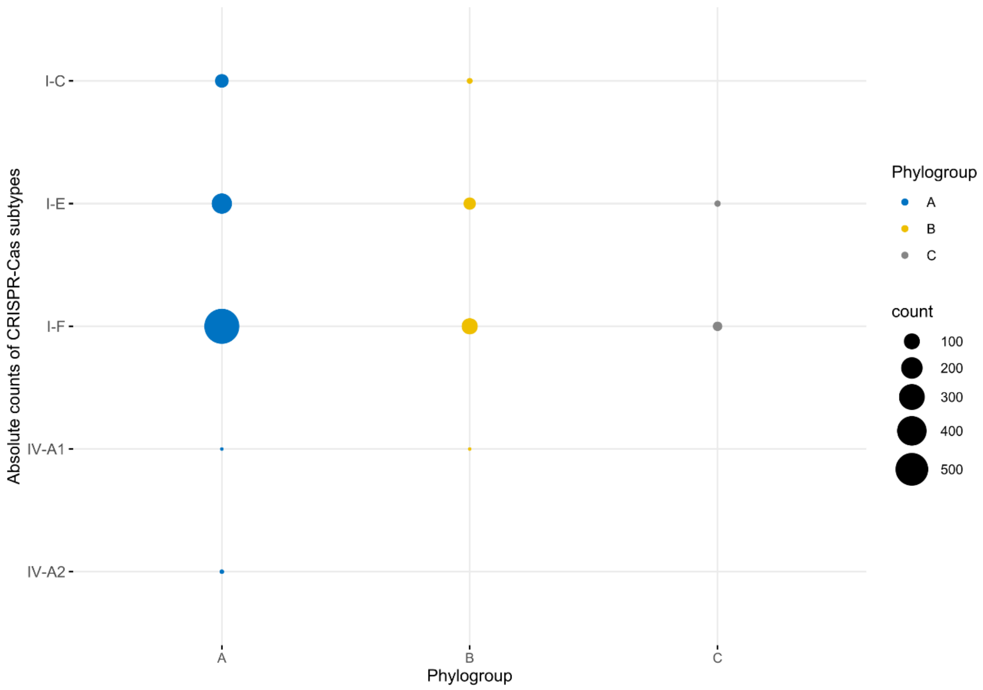

Figure S6. Absolute counts of CRISPR-Cas subtypes identified across genomes from the three phylogroups. Circles in blue represent phylogroup A, yellow B, and grey C. Circle size is proportional to the number of absolute counts.

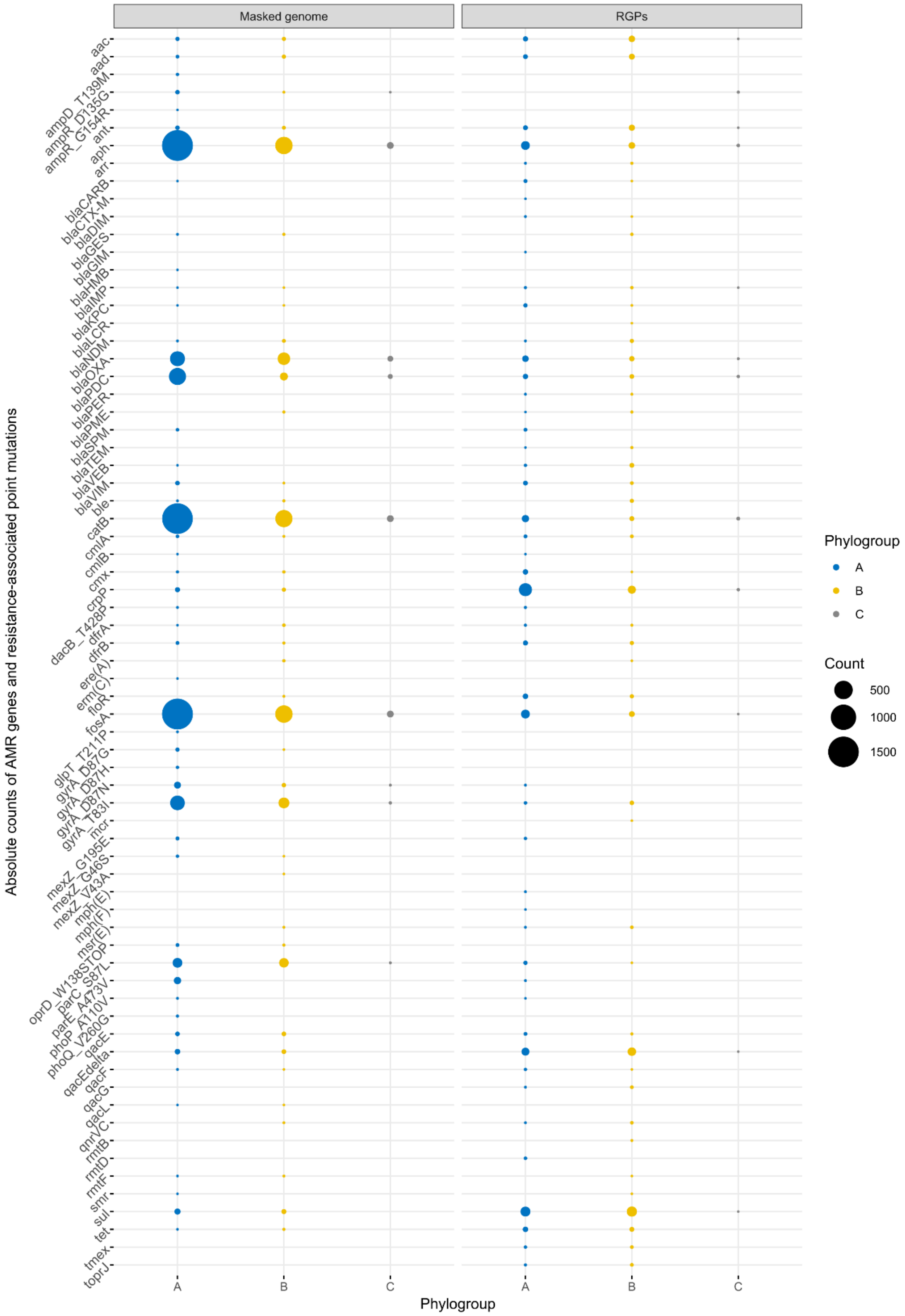
Masked genome
RGPs
Absolute counts of AMR genes and resistance-associated point mutations
aac
aad
ampD_T139M
ampR_D135G
ampR_G154R
ant
aph
arr
blaCARB
blaCTX-M
blaDIM
blaGES
blaGIM
blaHMB
blaIMP
blaKPC
blaLCR
blaNDM
blaOXA
blaPDC
blaPER
blaPME
blaSPM
blaTEM
blaVEB
blaVIM
ble
catB
cmlA
cmlB
cmx
crpP
dacB_T428P
dfrA
dfrB
ere(A)
erm(C)
floR
fosA
glpT_T211P
gyrA_D87G
gyrA_D87H
gyrA_D87N
gyrA_T83I
mcr
mexZ_G195E
mexZ_G46S
mexZ_V43A
mph(E)
mph(F)
msr(E)
oprD_W138STOP
parC_S87L
parE_A473V
phoP_A110V
phoQ_V260G
qacE
qacEdelta
qacF
qacG
qacL
qnrVC
rmtB
rmtD
rmtF
smr
sul
tet
tmex
toprJ
A
B
C
A
B
C
Phylogroup
Phylogroup
● A
● B
● C
Count
● 500
● 1000
● 1500

Figure S7. Absolute counts of AMR genes and resistance-associated point mutations across masked genomes and RGPs from the three phylogroups. Genes and mutations are part of the AMRFinder database (45). Circle size is proportional to the number of absolute counts. Circles in blue represent phylogroup A, yellow B, and grey C.

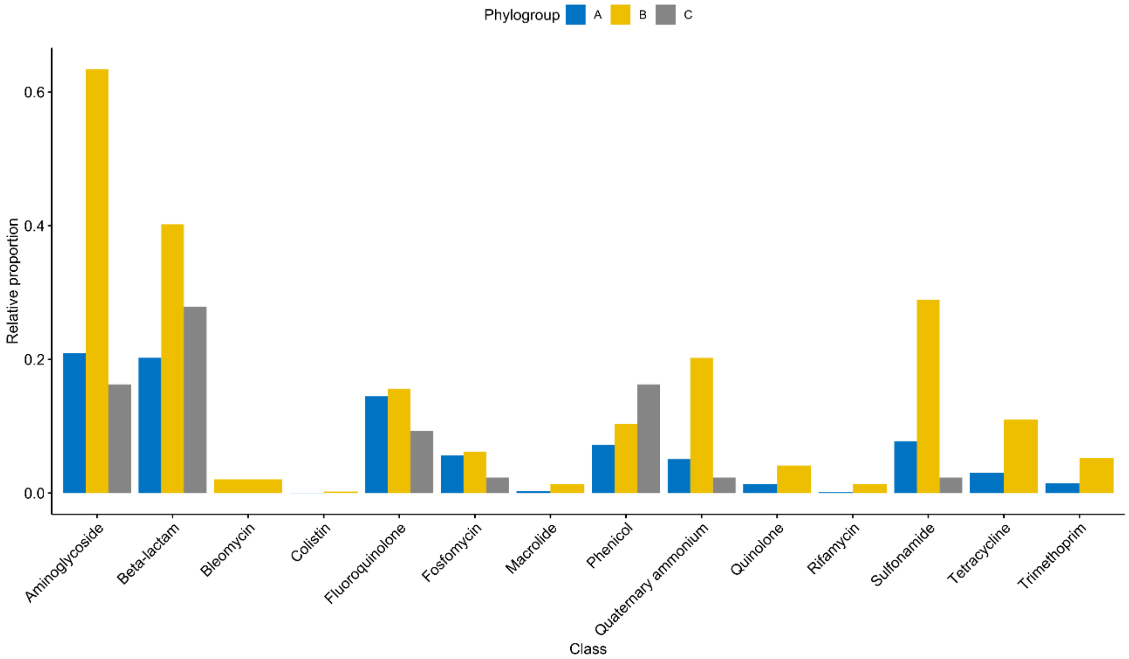

Figure S8. Barplots showing the relative proportion of genes encoding resistance to antibiotics from different classes across RGPs from the three phylogroups. Genes were normalized to the total number of genomes found in each phylogroup. Bars in blue represent phylogroup A, yellow B, and grey C.

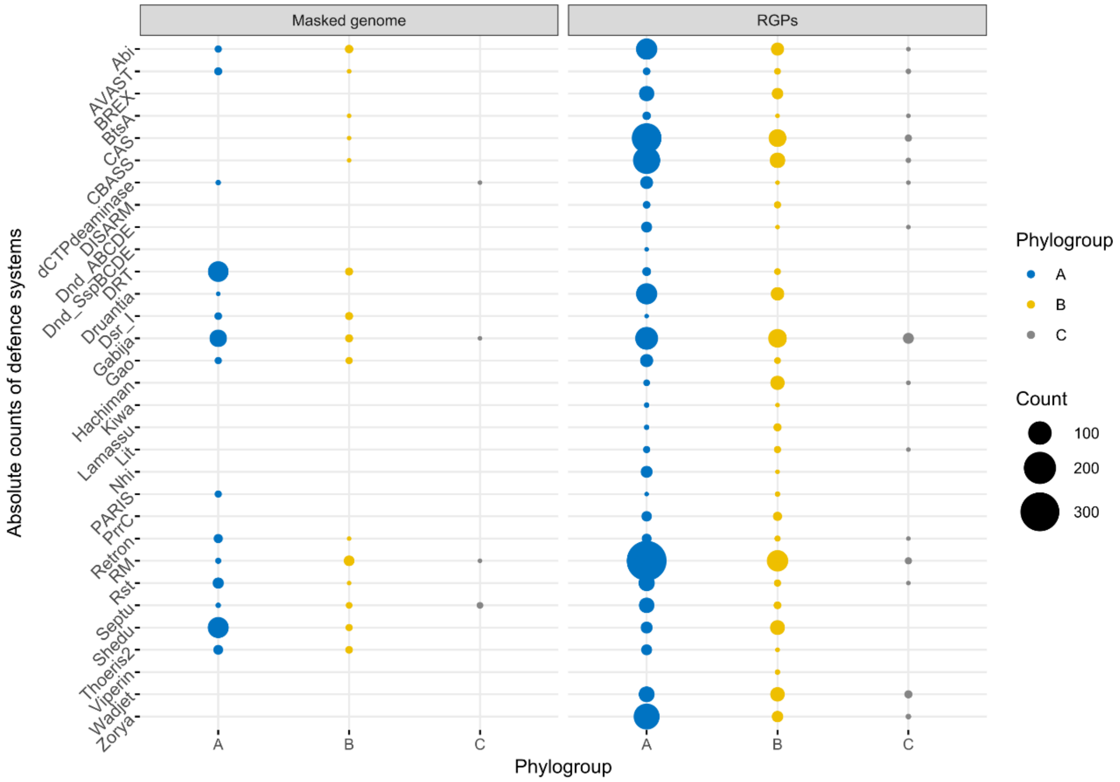

Figure S9. Absolute counts of defence systems across masked genomes and RGPs from the three phylogroups. Defence systems are part of the defense-finder database (36). Circle size is proportional to the number of absolute counts. Circles in blue represent phylogroup A, yellow B, and grey C.

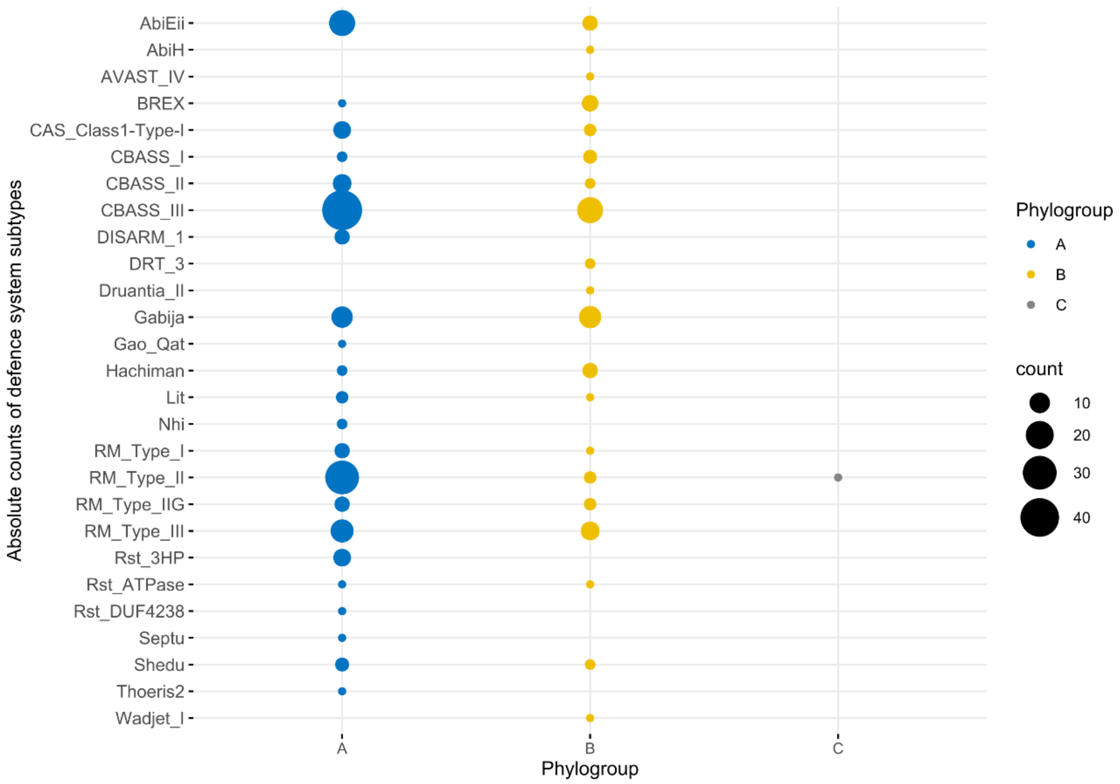

Figure S10. Absolute counts of defence systems across ICEs/IMEs from the three phylogroups. Defence systems are part of the defense-finder database (36). Circle size is proportional to the number of absolute counts. Circles in blue represent phylogroup A, yellow B, and grey C.

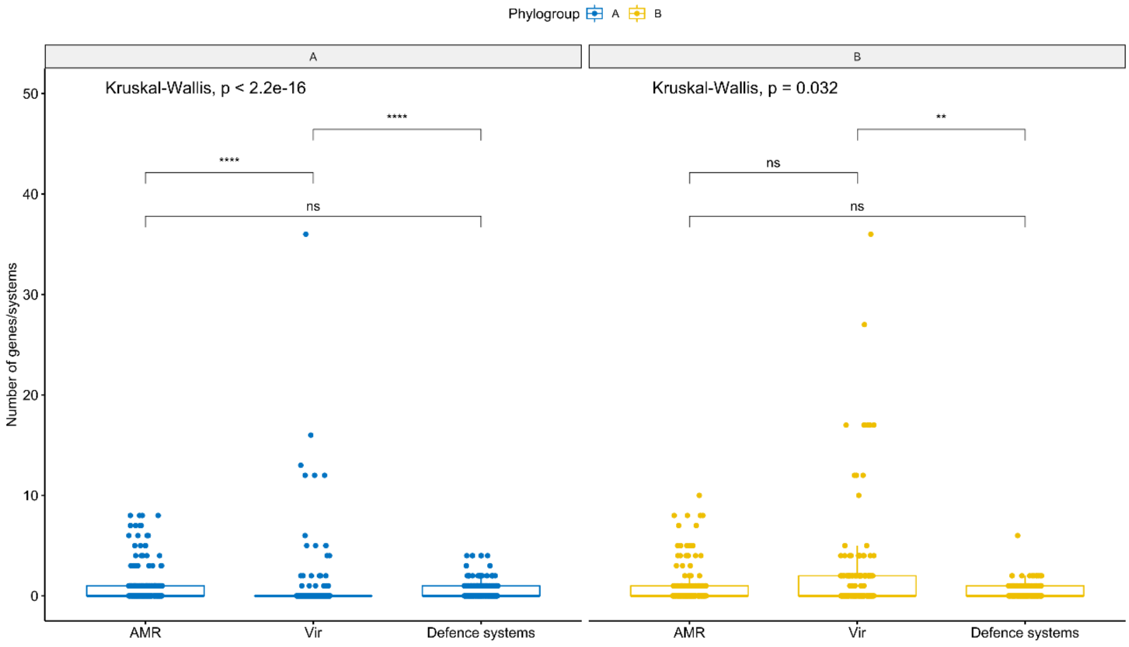

Figure S11. Boxplots representing the variation in the number of AMR genes, defence systems, and virulence genes found in ICEs/IMEs across the two larger phylogroups A and B. Values above 0.05

were considered as non-significant (ns). Stars indicate significance level: * p <= 0.05, ** p <= 0.01, *** p <= 0.001, and **** p <= 0.0001. Boxplots in blue represent phylogroup A, yellow B, and grey C.

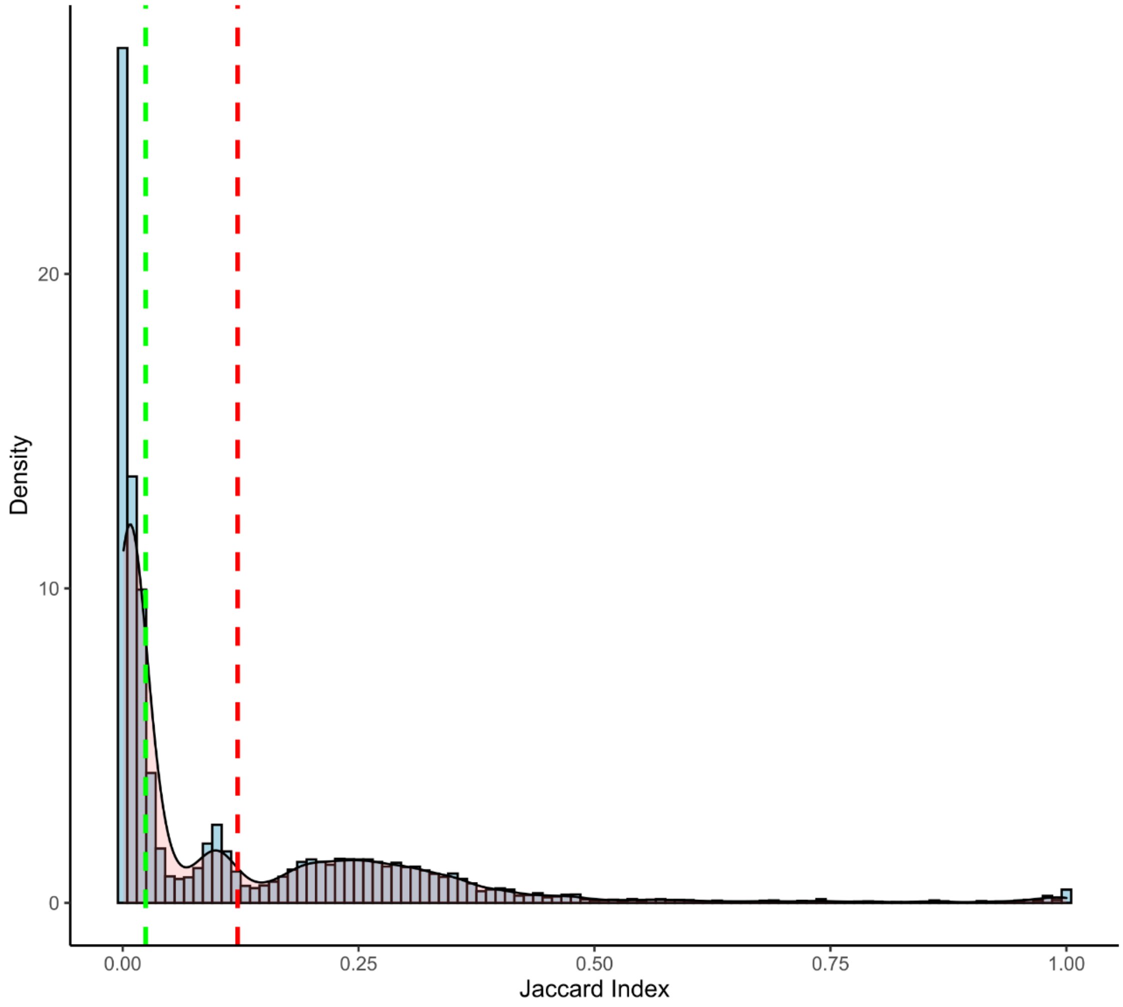

Figure S12. Histogram showing the right-skewed distribution of the Jaccard Index between all pairs of ICEs/IMEs. Median and mean values are highlighted by vertical dashed lines in green and red, respectively.

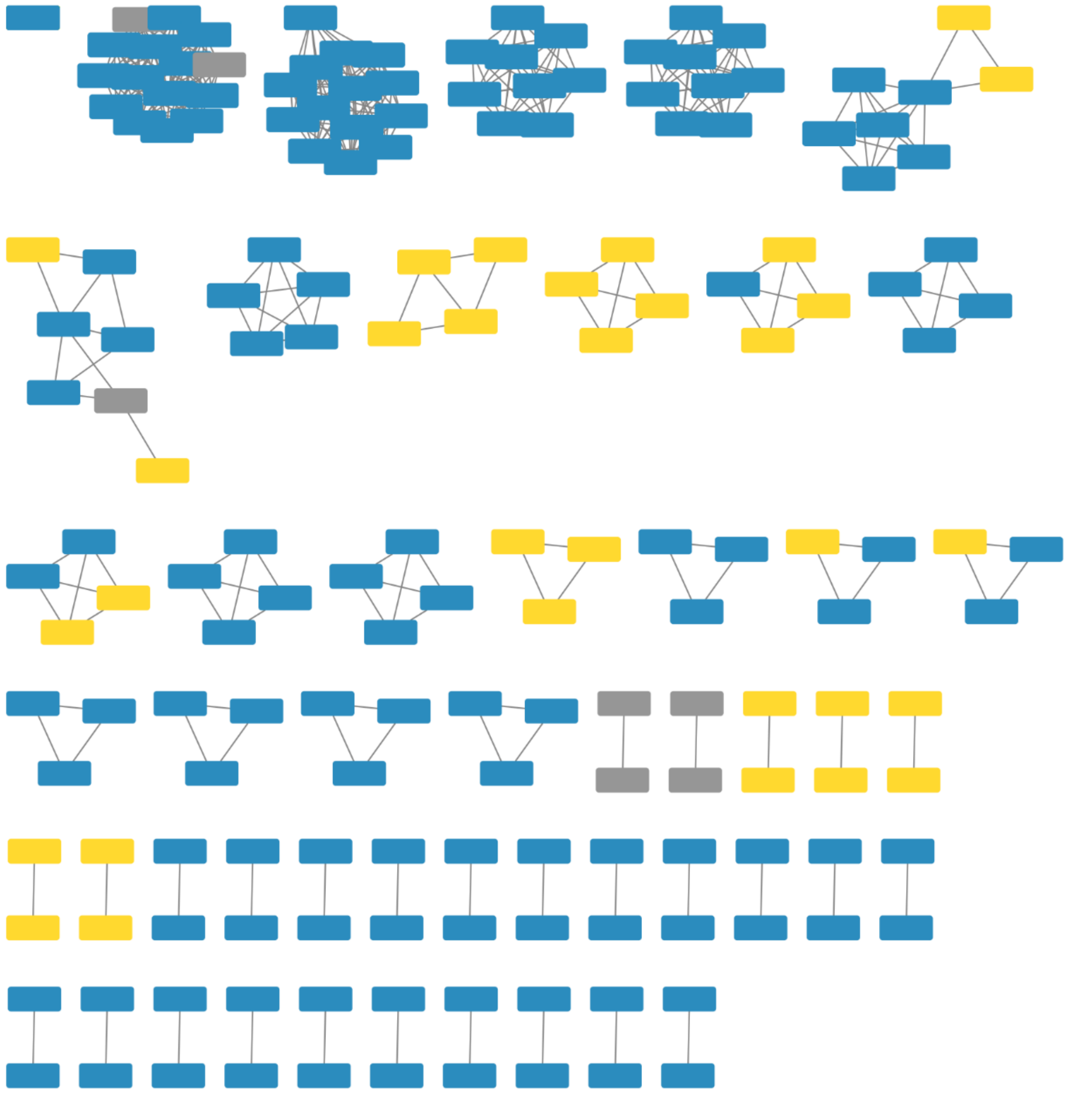

Figure S13. Network of clustered RGPs from the three phylogroups, using the mean Jaccard Index between all pairs of RGPs as a threshold. Each RGP is represented by a node, connected by green edges according to the pairwise distances between all RGPs pairs. Numbered ellipses represent RGPs that belong to the same cluster. The network has a clustering coefficient of 0.777, a density of 0.007, a centralization of 0.026, and a heterogeneity of 0.755. RGPs from phylogroup A are coloured in blue, from phylogroup B in yellow, and from phylogroup C in grey.