Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrative genomic and network analyses map horizontal gene transfer dynamics in Pantoea agglomerans

Muhammad Sulman
Agriculture and Agri-Food Canada
Antonet M. Svircev
Agriculture and Agri-Food Canada
James T. Tambong
Agriculture and Agri-Food Canada
Walid Ellouze
oualid.ellouz@AGR.GC.CA

Agriculture and Agri-Food Canada

Research Article

Keywords: Horizontal gene transfer, gene-sharing network, antimicrobial resistance, virulence factors

Posted Date: July 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8361091/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Pantoea agglomerans is a Gram-negative bacterium with remarkable ecological versatility, functioning as a plant symbiont, opportunistic pathogen, and environmental epiphyte. This adaptability is strongly influenced by horizontal gene transfer (HGT), yet its sources, vehicles, and key partners involved in these gene exchanges remain unclear.

Results

We conducted a multi-faceted genomic analysis of 165 *P. agglomerans* genomes that resulted in the construction of a pangenome that revealing an "open" structure driven by the continuous acquisition of shell and cloud genes enriched in mobile genetic element (MGE)-related functions. Comparative analysis of antimicrobial resistance (AMR) and virulence genes across chromosomes, plasmids, and prophages showed that plasmids, particularly non-ubiquitous ones, were the primary vehicles for horizontal AMR gene acquisition, while prophages contributed more modestly. Phylogenetic analysis of key gene clusters, including Pantocin A and the Type III Secretion System (T3SS), revealed both horizontal acquisition from other species and stable vertical inheritance within *P. agglomerans*. A large-scale gene-sharing network analysis across *Enterobacterales* identified *Pantoea vagans* as the most influential gene donor to *P. agglomerans*, with network structure revealing strong ecological niche-based clustering.

Conclusions

This study maps the landscape of HGT in *P. agglomerans* and highlights the combined roles of MGEs and ecological context in shaping genome evolution in this adaptable species. These findings provide deeper insight into the evolutionary forces driving diversification within *P. agglomerans* and establish a framework for studying HGT dynamics across other environmentally flexible bacteria.

Background

Pantoea agglomerans (formerly *Erwinia herbicola*) is a Gram-negative bacterium in the family Enterobacteriaceae, known for its wide distribution and diverse functions. This bacterium has been isolated from plants, soil, water, insects, clinical samples [1, 2], and spacecraft surfaces [3]. In these variable environments, the species performs different roles. Strains may promote plant growth [4–6], protect against pathogens by producing antimicrobial compounds [7–10], and act as an opportunistic pathogen of both plants [11] and humans. This diversity of roles may arise partially from changes in gene expression in response to different environmental conditions, but also by the differences in gene content between strains [12, 13]. *Pantoea* spp. may carry genes that support agricultural applications, whereas others contain traits that result in pathogenesis.

This raises a central question: how does *P. agglomerans* acquire such diverse genes, and what sources and genetic vehicles mediate this transfer? To address this question, we must determine where the relevant genes reside, how they move, and from which donors they originate. Identifying whether genes are carried on prophages, plasmids, integrative and conjugative elements, transposons, or the chromosome clarifies both their genomic location and the mechanisms by which they are transferred [14, 15]. In *P. agglomerans*, the source of genes and their common receipts has not been addressed comprehensively. To date, there exists a lack of data that examines all mobilizable genes within *P. agglomerans*. Subsequently,, we lack a broader picture of how genes that ultimately occur in *P. agglomerans* circulate among bacteria. Without such a map, it cannot be determined who are the most common gene-sharing partners for *P. agglomerans*.

Prior work has largely focused on individual biosynthetic clusters such as pantocin A, also known as herbicolin O [16, 17], PNP-2 [10], PNP-3 [18], and herbicolin A [19, 20], describing their local genomic context in *P. agglomerans* and cataloguing which other taxa produce analogous metabolites. When the presence of these clusters is placed in a phylogenetic framework they indicate repeated horizontal transfers, resulting in patchy and incongruent distributions that rarely fit strict vertical inheritance.

In this study, a multi-part analysis was conducted to identify bacterial species and mobile genetic elements that significantly contribute to bulk gene exchange. A *P. agglomerans* pangenome was first constructed, and genes were categorized as persistent

(present in nearly all genomes), shell (shared by an intermediate fraction), or cloud (found in few isolates) [21]. Functional groups were then assigned to each category. Finally, virulence and antimicrobial resistance genes were mapped to genomic compartments, including the chromosome, prophage regions, Large *Pantoea* plasmids 1 and 2 (LPP-1 and LPP-2) [12], and other plasmids.

In order to investigate the sources of horizontal gene transfer in *P. agglomerans*, two complementary analyses were conducted. First, initiate searches for previously described variably present gene clusters of interest in *P. agglomerans* and related *Enterobacterales*; specifically antibiotic biosynthetic clusters and Type III secretion systems. Putative horizontal transfer events were inferred based on phylogenetic incongruence between gene and species trees. Second, design and build large-scale gene-sharing networks: one for all non-persistent genes and one restricted to non-persistent virulence genes, across *Enterobacterales*. These networks will quantify the frequency and direction of gene exchanges involving *P. agglomerans*, and identify the bacterial species most strongly associated with its horizontal gene-sharing patterns.

Results

Pangenome partitions vary in COG composition

The genomic diversity of 165 *P. agglomerans* strains was assessed through pangenome analysis, which identified a total of 19,368 gene families. The rarefaction curve for the total pangenome did not reach a plateau, indicating an "open" pangenome structure (Fig. 1A). The persistent genome stabilized rapidly, comprising 3782 gene families conserved across nearly all strains. In contrast, the accessory genome, composed of the shell (2249 genes) and cloud genes (13,337 genes), continued to expand with the inclusion of additional genomes.

Within a typical *P. agglomerans* genome, persistent genes account for 83.0 ± 5.9% of the total repertoire, whereas the shell and cloud fractions represent 12.24 ± 4.26% and 4.22 ± 5.14%, respectively. Functional annotation revealed significant partition-specific enrichments (Tukey's HSD, $p < 0.05$). Persistent genes were enriched in categories related to energy production and conversion (C); amino acid, carbohydrate, coenzyme, and lipid metabolism (E, G, H, I); translation and ribosome biogenesis (J); post-translational modification and chaperones (O); inorganic-ion transport (P); and signal transduction (T). The shell genome showed enrichment in transcription (K), genes of unknown function (S), and intracellular trafficking or vesicular transport (U). The cloud genome was enriched in replication and repair (L), cell-envelope biogenesis (M), motility (N), general function prediction only (R), defense mechanisms (V), and the mobilome, including prophages and transposons (X). The COG composition of the three pangenome components is illustrated in Fig. 1B.

Genomic compartments of AMR genes

Across all *Pantoea agglomerans* genomes, 9 antibiotic resistance genes and 12 metal resistance genes, 8 of which belong to the *mer* operon. Among these, *oqxB*, which confers quinolone resistance, was consistently present and classified as persistent. Similarly, *fief* (iron/zinc resistance) and other iron and zinc resistance genes were detected in nearly all strains, typically located on the chromosome. These genes represent conserved core functions which are not generally mobilized through horizontal gene transfer, hence, they were excluded from subsequent analyses.

Across the 30 complete genomes (Fig. 2A), the *bla* (β-lactamase) gene was frequently observed and almost invariably located on LPP-1. The *arsC* gene, encoding an arsenate reductase, was likewise predominantly found on LPP-1, though it occasionally detected within viral elements or non-ubiquitous plasmids. In contrast, LPP-2 contained no resistance genes. The *vat* gene, conferring streptogramin resistance, was located on the chromosome in all but one genome, while *fos* (fosfomycin resistance) was exclusively chromosomal. The silver resistance genes *silA* and *silC* were restricted to non-ubiquitous plasmids.

Patterns observed across all 165 strains (where plasmid subtypes were not distinguished; Fig. 2C) were broadly similar, with a few additional, sporadically distributed genes. The streptomycin resistance genes *strA* and *strB* were consistently plasmid-encoded. Kanamycin resistance genes *aphA* and *aphA-2* were each detected once, *aphA* on the chromosome and *aphA-2* on a plasmid. The entire *mer* operon was plasmid-encoded. Notably, every resistance gene except *aphA* was identified at least once on plasmids. On the other hand, viral elements contained only common AMR genes belonging to the shell compartment of the pangenome. The distribution of individual resistance genes across strains is shown in Fig. 2D.

Prevalence and localization of virulence genes

Searches of the VFDB across all 30 genomes identified 2,841 virulence factor genes; of these, 2,383 were part of the persistent gene compartment and were therefore excluded from subsequent analyses. Among the remaining non-persistent virulence factors, the majority were located on the chromosome. Notably, Type III secretion system genes were more frequently associated with plasmids, whereas no non-persistent virulence factors were detected on LPP-1 or LPP-2 (Fig. 2B). Type VI secretion system genes were identified within prophage regions, and only two Type IV effectors were detected across the 30 genomes. A summary of these results is presented in Fig. 2D.

Prevalence of antibiosis and secretion system gene clusters

Gene cluster prevalence varied greatly across isolates (Fig. 2D). Analysis of the 165 genomes revealed that a complete Type III Secretion System (T3SS) was encoded in a notable portion of the isolates (65/165), with an additional three strains carrying nearly complete T3SS clusters. The complete PNP-3 cluster was detected in 31 genomes, while incomplete copies of the cluster were also frequently observed, typically lacking the *pnp3c* gene. Other gene clusters of interest were less common: complete clusters for pantocin A and AGA phenazine were found in 13 and 10 genomes, respectively. The cluster for dapdiamide E was present in 6 genomes, and that for antibiotic B025670 in five. Herbicolin A was exceptionally rare, identified in only two *P. agglomerans* isolates: ZJU23 and 4.

Ubiquity of gene clusters in the order

Ubiquity of gene clusters in the *Enterobacterales* order
Several biosynthetic gene clusters were extremely rare within the *Enterobacterales* dataset. The complete B025670 antibiotic cluster was detected only in *P. agglomerans* 21TX0196, with a single partial match identified in *Pantoea dispersa* JGM112, which contained only 4 of the 14 genes. Herbicolin A cluster was entirely absent, with no homologous proteins detected in any genome. The dapdiamide E cluster was found in only one genome, *P. vagans* C9-1, which encoded 7 of 9 proteins but lacked *ddaE* and *ddaF*.

In contrast, several clusters were more widespread outside of *P. agglomerans*. *Hrc* genes of the T3SS, no genome beyond *P. agglomerans* contained a complete cluster. However, near-complete clusters were frequently observed in *Kosakonia cowanii*, *P. stewartii*, and *Erwinia amylovora*, which typically encoded 7 or 8 of the 9 genes. These consistently lacked *hrcT*, and *E. amylovora* also lacked *hrcU*. Similarly, the PNP-3 cluster was observed beyond the *Pantoea* genus. The full six-gene PNP-2 cluster was detected in 4 of 25 *P. agglomerans* genomes and in 2 of 3 *Rouxiella sylvae* genomes (Yersiniaceae). Two additional Yersiniaceae species, *Serratia nevei* and *Yersinia enterocolitica*, carried 2 of the 6 genes, specifically *pnp2e* and *pnp2f*.

Pantocin A was the most widely distributed cluster outside of *Pantoea*. The complete 4-gene cluster was detected in *P. agglomerans* (4 of 25 genomes), *P. ananatis* (1 of 25), *P. stewartii* (2 of 25), *P. vagans* (3 of 16), and *S. nevei* (1 of 25). Three of the four genes were also present in *Dickeya chrysanthemi*, while *D. dianthicola* carried only one. Specifically, *D. chrysanthemi* and *D. dianthicola* consistently possessed *paaP*, *paaA*, and *paaC*, but lacked *paaB*.

P. agglomerans prophages are relatively unknown

Across 165 genome assemblies, 184 near-complete or complete prophages were identified. Among these 73 were assigned to the family *Peduoviridae*. vConTACT3 did not resolve any named subfamily or genus-level classifications, and three unnamed putative genera encompassed 54% of the prophages (Fig. 3).

Phylogenetics of gene clusters in the *Enterobacterales* database

The placement of *P. agglomerans* varies across the 5 locus trees, PNP-2 (Fig. 4A), PNP-3 (Fig. 4B), pantocin A (Fig. 4C), AGA phenazine (Fig. 4D), and *Hrc* T3SS (Fig. 4E). In the pantocin A tree, *P. agglomerans* is non-monophyletic: Pa9-1 clusters with *P. ananatis* PNA 14 – 1, whereas a distinct, well-supported clade (ultrafast-bootstrap ≥ 70) contains the remaining *P. agglomerans* together with *P. vagans* and *P. stewartii*. PNP-2 and PNP-3 *P. agglomerans* sequences form single, strongly supported monophyletic clades; however, in PNP-2, strain 21TX0196 falls outside the principal *P. agglomerans* cluster. The AGA-phenazine tree (Fig. 4D) includes a single *P. agglomerans* sequence on a comparatively long branch. In the *Hrc* (T3SS) tree (Fig. 4E), *P. agglomerans* is

monophyletic but does not cluster with other *Pantoea* spp., instead it groups within a larger clade primarily composed of *Kosakonia cowanii*.

Phylogenetic analyses were not performed for the B025670, herbicolin A, or dapdiamide E biosynthetic gene clusters because each was detected in only two genomes within the *Enterobacterales* database.

Gene-sharing and virulence networks

In the non-persistent gene-sharing network (Fig. 5A), *P. vagans* exhibited the highest eigenvector centrality, surpassing even *P. agglomerans*. The strongest weighted edge for *P. agglomerans* connected it to *P. vagans*, despite multilocus sequence analyses typically identifying *P. eucalypti* as its closest relative. Overall, *Pantoea* spp. tended to occupy the most central positions in the network. Notably, several non-*Pantoea* taxa, including *Duffyella gerundensis*, *Erwinia rhapontici*, and *Erwinia billingiae*, also displayed high eigenvector centrality, clustering closely with multiple *Pantoea* species and ranking among the most central nodes. The single strongest edge in the network was observed between *P. septica* and *P. brenneri*.

The non-persistent virulence gene-sharing network showed a broadly similar topology (Fig. 5B). In this case, however, *P. agglomerans* exhibited the highest eigenvector centrality, followed closely by *P. vagans*. The edge connecting these two species carried the greatest weight in the network.

Discussion

The *Pantoea agglomerans* pangenome growth is driven by MGEs

P. agglomerans has an open pangenome, meaning that sequencing additional strains continues to reveal new gene families, reflecting a high capacity for horizontal gene transfer. This contrasts with species that have relatively closed pangenomes, such as *Chlamydia trachomatis* [22], *Yersinia pestis* [23], and *Bacillus anthracis* [24], where adding genomes yields few new genes. *P. agglomerans* frequently acquires genes from plasmids, phages, and other conjugative elements.

Consistent with this, we observe that shell genes are enriched for COG categories K, S, and U, and cloud genes for L, M, N, R, V, and X. Category K often includes mobile-element regulators (e.g., phage repressors and plasmid-borne transcription factors) that modulate both host and cargo gene expression [25]. COG S is common on phages and plasmids and is dominated by hypothetical proteins. COG U includes secretion and conjugation systems typical of plasmids and integrative conjugative elements and is comparatively rare in phages [25]. In the cloud, enrichment of replication and repair (L), cell envelope (M), motility (N), defense mechanisms (V), and mobilome functions (X) indicates frequent horizontal acquisition [25].

Together, these patterns show that the shell and cloud partitions largely reflect genes associated with bacteriophages, plasmids, and other mobile genetic elements. Once more strains are sampled, new mobile elements will be discovered, resulting in the introduction of additional gene families which will drive the continued expansion of the *P. agglomerans* pangenome.

The AMR gene lifestyle

AMR genes show clear compartmental preferences. In our data set, *arsC* and *bla* occurred most frequently on LPP-1, whereas *vat* and *fos* were predominantly chromosomal. The chromosome and LPP-1 are both vertically inherited, this pattern likely contributes to the widespread prevalence and stable maintenance of these four genes within the species. Concurrently, each of these genes additionally appear on non-ubiquitous plasmids and within viral elements; this is consistent with carriage on transposons that can move between compartments (chromosome, conserved and non-ubiquitous plasmids, and phage). Phages and non-ubiquitous plasmids are key vectors of horizontal gene transfer, such mobility provides a route for these genes to re-enter *P. agglomerans* clades after loss, thereby sustaining their circulation across the species. Notably, *vat, fos*, *bla* and *arsC* have all been found within transposons [26–29].

On the other hand, the remaining (rarer) AMR genes show little evidence of maintenance through vertical inheritance. The gene presence-absence tree in Fig. 2.D shows sparse occurrences of these genes, suggesting that they were likely acquired from other species. Plasmids appear to be the primary vehicle through which the majority of these rarer AMR genes are exchanged between

different bacteria. Apart from *aphA*, every single AMR gene was found at least once on plasmids, reinforcing their role as primary vehicles of horizontal transfer.

Prophages and their contribution to HGT

The dominance of three unnamed prophage lineages, together with vConTACT3's limited genus-level assignments for these elements, suggests that the prophage diversity in *P. agglomerans* remains under-resolved, reflecting gaps in current viral taxonomies and reference databases.

Lysogeny is usually a prerequisite for prophage-mediated gene transfer [15], prophages, like those detected by geNomad within bacterial genomes, should theoretically be the primary contributors to horizontal gene transfer. The recurrent detection of three prophage genera across multiple *P. agglomerans* strains implies that these phages are likely among the strongest mediators of phage-based horizontal gene transfer in the species.

We observed partially intact T6SS gene clusters within prophages in two genomes. This pattern suggests episodic prophage carriage of secretion modules, consistent with previous reports that prophages can occasionally mobilize virulence or interbacterial interaction systems [30, 31]. Such cargo could significantly alter competitive dynamics or host interactions when present.

By contrast, AMR genes were uncommon within prophages, consistent with broader evidence that phages rarely encode AMR determinants [25, 32], whereas plasmids serve as the primary vehicles of AMR dissemination. We did not identify any AMR genes unique to, or predominantly associated with viral elements. Accordingly, in *P. agglomerans*, plasmids and other mobile elements appear to be the dominant sources of AMR acquisition, particularly for "novel" AMR genes (i.e., those not previously circulating within *P. agglomerans*). When AMR genes did occur within viral elements, they were typically part of the pangenome's shell compartment, suggesting that prophages may nevertheless contribute to the maintenance of certain AMR genes in the population.

The mobility of different gene clusters

Pantocin A is non-monophyletic within *P. agglomerans*, for example Pa9-1 clusters with *P. ananatis*, which is consistent with cross-species horizontal transfer of this biosynthetic locus. In contrast, the *P. agglomerans* clades for the T3SS (*Hrc* genes), PNP-2, and PNP-3 are monophyletic, indicating predominantly vertical inheritance with only limited within-species exchange. The single *P. agglomerans* AGA-phenazine lineage lies on a long branch relative to its *Parvobacterium parvum* counterpart, suggesting substantial divergence. Because AGA-phenazine is rare in our dataset, we cannot yet infer the breadth of hosts to which this genomic island can mobilize.

Shared ecological niche plays a strong role in gene-sharing

In Fig. 5, both networks suggest that *P. vagans* may contribute disproportionately to the expansion of the *P. agglomerans* pangenome. In the non-persistent gene-sharing network, *P. vagans* has the highest centrality, and the strongest *P. agglomerans* cross-species edge connects to *P. vagans*, surpassing even *P. eucalypti*, its closest neighbor on MLSA [33, 34]. Additionally, several plant-associated non-*Pantoea* taxa ranked higher in centrality than multiple *Pantoea* spp. in both networks, underscoring the influence of shared ecological niches on gene exchange. Likewise, a tight triad links *P. septica*, *P. brenneri*, and *P. piersonii*, all commonly isolated from human-associated environments and implicated as opportunistic pathogens.

Conclusion

This study set out to unravel the complex web of horizontal gene transfer that shapes the functional diversity of *P. agglomerans*. By integrating pangenome analysis, genomic compartmentalization, phylogenetic inference, and large-scale network modeling, we constructed a comprehensive map of the genetic vehicles and exchange partners that support this species' remarkable adaptability.

Our findings reveal that *P. agglomerans* possesses a dynamic, "open" pangenome, continuously supplemented by genes from mobile genetic elements. Plasmids emerged as the principal vectors of antimicrobial resistance acquisition, clarifying how this opportunistic pathogen acquires clinically and agriculturally relevant traits. While prophages play a lesser role in AMR dissemination, they contribute to the mobilome by occasionally carrying other functional modules, such as secretion systems.

Gene-sharing network analysis identified *P. vagans*, a common plant-associated bacterium, as the most influential gene-sharing partner of *P. agglomerans*, surpassing even its closest evolutionary relatives. This finding highlights the phytosphere as a genetic

marketplace where co-habiting bacteria frequently exchange genes, driving the evolution of traits linked to virulence, antibiosis, and environmental resilience.

Future work could refine and extend these insights by broadening the taxonomic scope beyond *Enterobacterales* and provide a more complete view of the gene-sharing network available to *P. agglomerans*. Methodologically, the network could be refined by employing more robust criteria than mPID to define horizontal transfer events. A dedicated analysis of smaller mobile elements, such as insertion sequences and integrative elements, also represents a critical frontier, though it is currently limited by challenges in their reliable automated detection. Pursuing these advancements would provide higher-resolution model into the molecular and ecological forces shaping this species' evolutionary trajectory.

Methods

Pangenome construction

Pangenome was constructed with 165 strains of *P. agglomerans* (Table 1). Inclusion criteria required genomes to have completeness above 90% as assessed by NCBI's CheckM2 v.1.1.0 [35] tests. Additionally, all genomes were verified to share at least 95% average nucleotide identity (ANI) with the reference strain *P. agglomerans* FDAARGOS 1447, determined using skani v.0.2.2 [36]. Metagenomic sequences were excluded from the analysis.

Table 1
Metadata for the 165 *Pantoea agglomerans* strains included in the pangenome analysis.

Assembly Accession	Strain	Assembly Length (bp)	Assembly Level	Scaffold N50 (bp)	Scaffolds	Sequencing Tech	Completeness
GCA_019048385.1	FDAARGOS 1447	4,692,018	Complete Genome	3,999,686	3	Pacbio; Illumina	98.33
GCA_943184895.1	DAPP-PG734	5,396,424	Complete Genome	4,410,564	5	Illumina MiSeq and MinION	98.61
GCA_014841465.2	CFBP8774	5,464,831	Contig	4,331,596	7	PacBio Sequel	98.33
GCA_014839085.2	CFBP8785	5,044,125	Contig	4,328,092	4	PacBio Sequel	98.01
GCA_031888365.1	SUH1	5,157,404	Complete Genome	4,286,451	4	Illumina MiSeq; Oxford Nanopore MinION	99
GCA_014839525.2	CFBP13593	5,130,088	Contig	4,224,369	4	PacBio Sequel	99
GCA_031887785.1	T88c	5,095,245	Complete Genome	4,201,259	5	Illumina MiSeq; Oxford Nanopore MinION	99
GCA_031890185.1	BH6c	5,118,957	Complete Genome	4,186,627	5	Illumina MiSeq; Oxford Nanopore MinION	99
GCA_031891025.1	AJ2b	5,136,510	Complete Genome	4,185,742	6	Illumina MiSeq; Oxford Nanopore MinION	99
GCA_001709315.1	C410P1	5,115,241	Complete Genome	4,182,028	4	PacBio	94.56
GCA_017474165.1	AR1a	5,077,082	Complete Genome	4,169,790	5	Hybrid (Illumina MiSeq/PacBio RSII)	98.68
GCA_021559955.1	ZJU23	5,061,267	Complete Genome	4,151,813	4	PacBio	98.95
GCA_029016125.1	CB1	5,048,993	Complete Genome	4,145,235	4	Oxford Nanopore MinION; Illumina MiSeq	99
GCA_003704305.1	TH81	4,984,468	Complete Genome	4,128,817	4	Oxford Nanopore MiniION; Illumina MiSeq	97.66
GCA_031890505.1	AR5	4,967,622	Complete Genome	4,106,996	4	Illumina MiSeq; Oxford Nanopore MinION	99
GCA_022966105.1	Pa58	4,861,200	Complete Genome	4,100,605	3	PacBio	98.18

Assembly Accession	Strain	Assembly Length (bp)	Assembly Level	Scaffold N50 (bp)	Scaffolds	Sequencing Tech	Completeness
GCA_031890525.1	AR24	4,967,919	Complete Genome	4,093,174	4	Illumina MiSeq; Oxford Nanopore MinION	98.68
GCA_014839205.2	CFBP8756	5,029,763	Contig	4,091,070	5	PacBio Sequel	98.92
GCA_005233495.2	CFBP13505	4,932,464	Contig	4,077,956	4	PacBio Sequel	99
GCA_031889005.1	ROTS050421	5,046,058	Complete Genome	4,059,988	5	Illumina MiSeq; Oxford Nanopore MinION	98.65
GCA_014842695.2	CFBP13600	4,885,829	Contig	4,054,907	5	PacBio Sequel	98.65
GCA_004117135.1	CFSAN047154	5,158,055	Complete Genome	4,047,712	5	PacBio	98.33
GCA_004136415.1	CFSAN047153	5,158,055	Complete Genome	4,047,712	5	PacBio	98.49
GCA_026625085.1	AB378	4,781,164	Complete Genome	4,042,201	3	PacBio Sequel	98.63
GCA_031890205.1	AR8b	4,876,921	Complete Genome	4,038,916	4	Illumina MiSeq; Oxford Nanopore MinION	98.96
GCA_031889865.1	FC61912-B	5,168,010	Complete Genome	4,031,236	7	Illumina MiSeq; Oxford Nanopore MinION	98.65
GCA_003860325.1	L15	4,858,869	Complete Genome	4,029,228	4	Illumina MiSeq; Oxford Nanopore MiniION	98.65
GCA_036549535.1	Y03	4,768,143	Complete Genome	4,024,634	3	Nanopore PromethION	98.01
GCA_009762725.1	ASB05	4,858,648	Complete Genome	4,022,781	4	PacBio RSII; Illumina MiSeq	98.33
GCA_024799805.1	CHTF15	4,820,607	Complete Genome	4,012,069	4	PacBio Sequel; Illumina HiSeq	98.55
GCA_030710645.1	NBBC-01	4,927,193	Complete Genome	4,007,306	6	Oxford Nanopore PromethION	98.65
GCA_031889605.1	MMD61212-C	4,889,272	Complete Genome	3,978,822	4	NextSeq 500; Oxford Nanopore MinION	98.95
GCA_001662025.2	4188	5,041,798	Complete Genome	4,165,783	4	PacBio RS	98.96
GCA_001661985.2	824-1	5,034,205	Complete Genome	4,098,036	5	PacBio RS	98.06

Assembly Accession	Strain	Assembly Length (bp)	Assembly Level	Scaffold N50 (bp)	Scaffolds	Sequencing Tech	Completeness
GCA_021609925.1	PSV1-7	5,029,167	Chromosome	4,053,926	5	PacBio Sequel	98.79
GCA_010523255.1	UAEU18	4,825,350	Chromosome	4,040,629	4	Illumina NovaSeq; Oxford Nanopore MinION	99.63
GCA_001597625.1	3	4,813,581	Scaffold	4,016,073	4	454	98.58
GCA_030177195.1	9Rz4	4,977,153	Contig	3,899,485	13	Illumina NovaSeq; Oxford Nanopore GridION	98.65
GCA_000731125.1	190	5,002,566	Scaffold	3,393,498	5	454; PacBio; Illumina GAIIx	98.05
GCA_000743785.2	4	4,827,890	Scaffold	3,947,245	4	454	98.65
GCA_014354155.1	Pa21-13	4,792,898	Contig	839,060	76	Illumina MiSeq	99.28
GCA_037152335.1	AS129	4,884,712	Contig	649,680	52	Illumina NovaSeq	98.65
GCA_009759885.1	C1	4,846,162	Contig	645,999	21	Illumina MiSeq	98.65
GCA_014839425.1	CFBP13616	4,780,294	Contig	623,747	30	DNB-Seq	98.65
GCA_014354165.1	Pa17-5	4,970,051	Contig	620,738	138	Illumina MiSeq	98.65
GCA_010667925.1	T1	4,530,745	Scaffold	610,154	17	Oxford Nanopore MinION-Illumina HiSeq	97.97
GCA_010667775.1	T6	4,741,789	Scaffold	605,029	19	Oxford Nanopore MinION-Illumina HiSeq	98.35
GCA_037727885.1	G129K1-1	4,794,802	Contig	601,559	17	Illumina HiSeq	98.33
GCA_010667975.1	B3	5,229,895	Scaffold	577,213	59	Illumina HiSeq	98.65
GCA_010667995.1	R1	4,718,053	Scaffold	576,160	18	Oxford Nanopore MinION-Illumina HiSeq	98.65
GCA_000627115.1	RIT273	5,365,338	Contig	573,720	26	Illumina MiSeq	98.65
GCA_014353795.1	Pa39-7	4,965,341	Scaffold	572,330	116	Illumina MiSeq	98.33
GCA_038116615.1	B9	4,817,156	Contig	571,821	43	Illumina	98.33
GCA_014839545.1	CFBP13583	4,776,328	Contig	565,076	34	Illumina HiSeq	98.33

Assembly Accession	Strain	Assembly Length (bp)	Assembly Level	Scaffold N50 (bp)	Scaffolds	Sequencing Tech	Completeness
GCA_014838815.1	CFBP8800	5,117,730	Contig	562,885	50	DNB-Seq	98.65
GCA_039675505.1	Pa21-23	4,763,024	Scaffold	3,844,538	12	Illumina	98.33
GCA_014839125.1	CFBP8784	4,755,237	Contig	555,509	32	DNB-Seq	98.65
GCA_039671955.1	Pa21-17	4,870,609	Scaffold	3,959,095	24	Illumina	98.28
GCA_010667865.1	T5	4,739,540	Scaffold	551,962	24	Oxford Nanopore MinION-Illumina HiSeq	98.05
GCA_010667845.1	T2	4,823,700	Scaffold	551,948	30	Oxford Nanopore MinION-Illumina HiSeq	98.37
GCA_010667855.1	T3	4,714,968	Scaffold	551,948	26	Oxford Nanopore MinION-Illumina HiSeq	98.05
GCA_010667895.1	T4	4,952,873	Scaffold	551,948	31	Oxford Nanopore MinION-Illumina HiSeq	98.68
GCA_039671905.1	Pa47-1	4,852,240	Scaffold	3,962,706	22	Illumina	98.33
GCA_039672365.1	Pa21-19	5,015,122	Scaffold	4,050,663	13	Illumina	98.65
GCA_037151835.1	HN022	4,937,041	Contig	539,596	56	Illumina NovaSeq	98.65
GCA_039672095.1	Pa9-1	4,837,860	Scaffold	4,024,459	9	Illumina	98.33
GCA_039672205.1	Pa13-4	4,703,659	Scaffold	3,894,625	12	Illumina	98.65
GCA_039672115.1	Pa1-28	5,023,239	Scaffold	4,070,922	22	Illumina	98.65
GCA_039672415.1	Pa31-4	4,971,316	Scaffold	4,064,127	11	Illumina	98.63
GCA_039671785.1	Pa47-2	4,895,426	Scaffold	3,953,721	27	Illumina	98.33
GCA_010667945.1	R5	4,821,784	Scaffold	533,500	47	Illumina HiSeq	98.81
GCA_000757415.2	MP2	4,733,829	Contig	531,471	16	454	97.6
GCA_039671645.1	Pa45-2	4,910,947	Scaffold	4,092,317	13	Illumina	98.32
GCA_039671325.1	Pa46-1	5,117,288	Scaffold	4,006,042	13	Illumina	98.01
GCA_039671705.1	Pa39-21	5,029,421	Scaffold	3,906,632	12	Illumina	98.54
GCA_039671805.1	Pa39-11	4,738,359	Scaffold	3,982,121	8	Illumina	98.68
GCA_039671815.1	Pa39-H1	4,738,324	Scaffold	3,981,815	9	Illumina	98.68
GCA_010667965.1	A2	4,970,487	Scaffold	518,825	20	Oxford Nanopore MinION-Illumina HiSeq	98.33

Assembly Accession	Strain	Assembly Length (bp)	Assembly Level	Scaffold N50 (bp)	Scaffolds	Sequencing Tech	Completeness
GCA_039671925.1	Pa39-1	4,960,804	Scaffold	4,072,344	17	Illumina	98.41
GCA_039672755.1	PaJII2	4,961,037	Scaffold	4,072,469	16	Illumina	98.41
GCA_039671475.1	Pa45-1	5,342,770	Scaffold	4,085,400	35	Illumina	98.33
GCA_039671755.1	Pa39-25	5,048,676	Scaffold	4,066,683	16	Illumina	98.33
GCA_039675815.1	Pa17-3	4,834,160	Scaffold	4,032,012	9	Illumina	98.65
GCA_034044955.1	CFBP9000	4,725,785	Scaffold	582,759	25	DNBseq	98.65
GCA_002222515.1	JM1	4,794,872	Scaffold	582,415	34	Illumina HiSeq	98.46
GCA_039671745.1	Pa39-23	5,090,935	Scaffold	4,003,360	14	Illumina	98.54
GCA_014838975.1	CFBP8792	5,239,108	Contig	471,411	54	DNB-Seq	97.7
GCA_039671725.1	Pa39-22	5,041,576	Scaffold	4,041,063	42	Illumina	98.63
GCA_039671945.1	Pa31-7	5,147,932	Scaffold	4,011,578	55	Illumina	98.65
GCA_039672495.1	Pa39-26-1	4,989,609	Scaffold	3,946,205	17	Illumina	98.65
GCA_039672705.1	Pa21-221	4,832,409	Scaffold	3,989,163	12	Illumina	98.33
GCA_039671275.1	Pa39-8	5,085,198	Scaffold	4,063,457	23	Illumina	98.65
GCA_039671455.1	Pa39-20	4,937,985	Scaffold	4,009,243	9	Illumina	98.65
GCA_039671365.1	Pa39-16	5,200,073	Scaffold	4,040,603	24	Illumina	98.62
GCA_039672605.1	Pa39-WII2	4,986,545	Scaffold	4,058,084	11	Illumina	98.54
GCA_039671635.1	Pa39-27	4,986,280	Scaffold	3,967,095	12	Illumina	98.54
GCA_039671825.1	Pa39-26	4,989,744	Scaffold	3,945,621	18	Illumina	98.65
GCA_039671505.1	Pa54-1	4,913,204	Scaffold	4,047,449	15	Illumina	98.65
GCA_039671855.1	Pa54-6	4,923,277	Scaffold	3,989,513	17	Illumina	98.65
GCA_010667985.1	B1	5,225,316	Scaffold	456,717	45	Illumina HiSeq	98.65
GCA_039671515.1	Pa54-2	4,913,457	Scaffold	4,044,981	14	Illumina	98.65
GCA_039671885.1	Pa54-7	4,923,640	Scaffold	3,988,053	21	Illumina	98.65
GCA_039671345.1	Pa54-31	4,871,596	Scaffold	4,005,291	15	Illumina	98.65
GCA_039671545.1	Pa54-8	4,922,767	Scaffold	3,987,885	18	Illumina	98.65
GCA_039671555.1	Pa54-3	4,913,453	Scaffold	4,045,205	16	Illumina	98.65
GCA_039675745.1	Pa21-5	4,832,757	Scaffold	4,020,803	12	Illumina	97.99
GCA_039672275.1	Pa21-20	4,775,429	Scaffold	4,041,540	12	Illumina	98.33
GCA_039672295.1	Pa21-12	4,775,104	Scaffold	4,041,603	11	Illumina	98.33
GCA_014839115.1	CFBP13569	4,797,171	Contig	454,687	36	Illumina HiSeq	98.65
GCA_039672735.1	Pa39-13	4,899,748	Scaffold	3,976,809	36	Illumina	98.65
GCA_039671605.1	Pa54-5	4,913,486	Scaffold	4,045,766	14	Illumina	98.65

Assembly Accession	Strain	Assembly Length (bp)	Assembly Level	Scaffold N50 (bp)	Scaffolds	Sequencing Tech	Completeness
GCA_039672455.1	Pa21-18	4,858,627	Scaffold	3,949,856	17	Illumina	98.65
GCA_039672185.1	Pa13-1	4,798,939	Scaffold	3,454,880	12	Illumina	98.96
GCA_039672015.1	Pa39-2	4,916,450	Scaffold	4,003,861	7	Illumina	98.65
GCA_039672395.1	Pa31-3	5,082,569	Scaffold	4,077,468	20	Illumina	98.65
GCA_037152265.1	AS139	4,824,484	Contig	432,218	62	Illumina NovaSeq	98.96
GCA_039672165.1	Pa21-15	4,831,859	Scaffold	4,020,021	13	Illumina	97.99
GCA_039672445.1	Pa21-16	4,777,072	Scaffold	3,629,906	14	Illumina	99.28
GCA_037152495.1	AS085	4,862,563	Contig	424,945	54	Illumina NovaSeq	98.49
GCA_009668105.1	RIT710	4,740,510	Contig	413,689	34	Illumina MiSeq	98.65
GCA_039672195.1	Pa13-5	5,008,547	Scaffold	3,944,924	26	Illumina	98.65
GCA_039671625.1	PaHJ	5,076,611	Scaffold	4,053,318	16	Illumina	98.65
GCA_905332865.1	RSO7	4,829,129	Contig	394,006	128	Illumina	98.98
GCA_039671975.1	Pa54-4	4,906,478	Scaffold	4,040,489	15	Illumina	98.65
GCA_014839175.1	CFBP8783	5,124,033	Contig	378,124	90	DNB-Seq	98.33
GCA_000952095.1	GB1	4,765,050	Contig	363,336	33	IonTorrent	93.58
GCA_014353755.1	Pa39-5	5,102,516	Contig	362,563	129	Illumina MiSeq	98.65
GCA_025642955.1	RIT-PI-T	4,955,229	Contig	360,618	46	Illumina MiSeq	98.65
GCA_039672265.1	Pa17-1	4,820,308	Scaffold	3,982,307	10	Illumina	98.33
GCA_029838845.1	GD03966	4,816,530	Contig	358,281	23	Illumina HiSeq	99
GCA_034044915.1	CFBP8988	4,810,088	Scaffold	579,250	30	DNBseq	98.65
GCA_039671265.1	Pa39-10	5,201,761	Scaffold	4,083,642	21	Illumina	98.65
GCA_037151955.1	HN003	5,068,934	Contig	348,912	57	Illumina NovaSeq	98.54
GCA_039671985.1	Pa39-3	5,044,342	Scaffold	4,017,007	21	Illumina	98.57
GCA_007050955.1	K1	5,000,942	Contig	330,732	46	Illumina HiSeq	98.58
GCA_012241415.1	KM1	4,995,756	Scaffold	330,495	46	Illumina HiSeq	98.58
GCA_039671375.1	Pa39-14	5,121,331	Scaffold	4,066,776	29	Illumina	98.65
GCA_039671255.1	Pa39-29	5,204,090	Scaffold	4,065,553	28	Illumina	98.65
GCA_015711555.1	NC	4,505,360	Scaffold	323,081	21	Illumina MiSeq	97.86
GCA_039671655.1	Pa39-18	5,310,801	Scaffold	4,142,091	40	Illumina	98.65
GCA_014354135.1	Pa21-3	5,079,038	Contig	310,071	117	Illumina MiSeq	98.61

Assembly Accession	Strain	Assembly Length (bp)	Assembly Level	Scaffold N50 (bp)	Scaffolds	Sequencing Tech	Completeness
GCA_034044845.1	CFBP8989	4,766,298	Scaffold	304,439	26	DNBseq	98.65
GCA_014839315.1	CFBP13731	4,984,051	Contig	297,864	60	DNB-Seq	98.65
GCA_017348995.1	MM2021_7	4,931,425	Scaffold	285,778	56	Illumina MiSeq	98.33
GCA_039671395.1	Pa39-17	5,116,385	Scaffold	4,068,642	11	Illumina	98.96
GCA_013403235.1	FJII-L5-SW-P2	4,724,025	Contig	283,046	26	Illumina NovaSeq 6000	95.71
GCA_029623495.1	RIT-To-3	5,215,402	Contig	281,499	83	Illumina MiSeq	98.96
GCA_039675785.1	Pa17-2	4,846,176	Scaffold	3,946,569	16	Illumina	98.31
GCA_014839325.1	CFBP13709	5,248,093	Contig	264,734	138	DNB-Seq	98.17
GCA_014353865.1	E325	4,786,783	Contig	257,949	162	Illumina MiSeq	98.65
GCA_004344785.1	ES418	5,176,065	Contig	256,581	66	Illumina MiSeq	98.33
GCA_022585235.1	20TX0122	5,001,150	Scaffold	341,936	49	Illumina NovaSeq	99
GCA_014839015.1	CFBP8786	4,928,760	Contig	253,460	51	DNB-Seq	98.96
GCA_014353995.1	E325-750	4,796,980	Contig	246,836	108	Illumina MiSeq	98.65
GCA_030433795.1	Pan8	4,729,284	Contig	217,580	56	Illumina	98.33
GCA_014839265.1	CFBP13566	5,123,550	Contig	206,479	84	Illumina HiSeq	98.2
GCA_005233795.1	CFBP13516	5,026,237	Contig	186,387	54	Illumina HiSeq	99
GCA_014838905.1	CFBP8791	4,905,192	Contig	185,134	71	DNB-Seq	98.96
GCA_005233805.1	CFBP13532	4,934,344	Contig	182,863	61	Illumina HiSeq	98.89
GCA_017315165.1	ANP8	5,035,017	Scaffold	179,494	1390	Illumina HiSeq	98.82
GCA_000814075.1	LMAE-2	4,981,165	Contig	152,429	155	Illumina MiSeq	97.99
GCA_013201565.1	DOAB1048	4,795,525	Contig	147,325	52	Illumina NextSeq	98.59
GCA_002157425.2	P5	5,074,260	Scaffold	112,468	127	Illumina HiSeq	98.41
GCA_015025115.1	R6	4,671,676	Contig	26,259	339	IonTorrent	96.86
GCA_000475055.1	Tx10	4,856,993	Scaffold	586,961	22	Illumina	98.33

All genomes were annotated using Bakta v.1.8.2 with default parameters [37]. The annotated genomes were subsequently analyzed using PPanGGOLiN v.2.2.1 [21] to construct the *P. agglomerans* pangenome following the complete workflow with default parameters. Based on this pangenome analysis, genes were classified into persistent, shell, or cloud partitions as described by Gautreau et. al. Clusters of Orthologous Groups (COG) categories were assigned to all genes in each partition using COGclassifier

v.2.0.0 [38]. Each COG category, pairwise differences among partitions, were tested with Tukey's HSD.Partition assignments in PPanGGOLiN are made independently for each category. Consequently, a gene family may be assigned to more than one partition, and the cumulative proportions of the persistent, shell, and cloud genomes do not necessarily sum to 100%.

Prophage identification and classification

Viral sequences and prophages were initially identified from all genomes using geNomad v.1.8.0 [39]. The resulting viral contigs underwent a stringent quality control process using CheckV v.1.0.3 [40] to remove host contamination and assess the completeness of viral genomes. Only viral elements, including both prophage regions and partial viral sequences identified by geNomad v.1.8.0, with greater than 90% estimated completeness were retained for taxonomic classification with vConTACT3 (accessed 12 May 2025) [41] with the v230 database. Admitting lower quality and incomplete prophage sequences increased the risk of false positives, as well as misclassifications, which we judged to be more important for our viral taxonomic analysis than the risk of false negatives. However, for the purposes of studying the localization of virulent and biosynthetic genes within viral elements, all identified prophages were included.

Chromosome and plasmid classification

Only 30 of the 165 *P. agglomerans* genomes were assembled to completion (i.e., single chromosomal contig ± plasmids). In Table 1 (and in their publicly available NCBI genome data), these are designated at the assembly level "Chromosome" or "Complete Genome". These 30 non-fragmented assemblies were analyzed separately to determine the precise genomic compartment of genes of interest. In each completed assembly, the largest contig was designated as the chromosome. The second-largest contig was classified as LPP-1 based on size and the LPP-2 plasmid was identified by replicon homology. We queried the MOB-suite replicon database [42] with MMseqs2 v.15.6f452 [43] against the *P. agglomerans* DAPP-PG 734 LPP-2 reference and determined rep_807 replicon as the marker for LPP-2. Subsequently, if a strong hit to rep_807 was found, it was classified as LPP-2. All remaining plasmid contigs in these complete assemblies were categorized as non-ubiquitous plasmids.

Because only 30 genomes were complete, many loci present exclusively in fragmented assemblies could not be resolved to specific ubiquitous plasmids by contiguity. To incorporate these genomes, we used PLASMe [44] (default parameters) to classify contigs as chromosomal or plasmid. In this mode, "plasmid" denotes any extrachromosomal element; LPP-1 and LPP-2 could not be reliably distinguished, and plasmid assignments do not imply membership in either ubiquitous plasmid.

Functional gene annotation

The functional potential of the pangenome was investigated by annotating representative gene families. Genes were classified as virulent if their representative sequences showed significant homology (E-value < $1x10^{-5}$) to entries in the Virulence Factor Database (VFDB 2022, http://www.mgc.ac.cn/VFs/) [45]. Each protein was allowed a single VFDB match, selected based on the highest bitscore. Antimicrobial resistance (AMR) genes were identified using AMRFinderPlus v.4.0.19 [46]. When a representative gene was annotated with a virulence or AMR function, all members of the corresponding gene family were assigned the same classification. Genes co-localized with previously identified viral elements were designated as viral. While all identified viral elements were used in this overlapping analysis, only complete viral genomes were used in the initial taxonomic classification. This approach prevents false-negative classifications, and ensures that gene-virus co-localization patterns are comprehensively captured.

Construction of gene-sharing networks

To investigate the sources of horizontal gene transfer (HGT) in *P. agglomerans*, gene-sharing networks were constructed in which nodes represented species and edges reflected the extent of gene exchange. Conventional methods for tracking HGT are generally not designed to handle more than one gene at a time and are unsuitable for analyses involving thousands of genes. Therefore, we developed a new methodology to enable large-scale investigation of HGT within *P. agglomerans*. The genome database was first assembled to serve as the node set for network construction. The minimum percent identity (mPID) threshold was determined for each *P. agglomerans* gene family to distinguish vertical inheritance from horizontal gene transfer (HGT). Because genes acquired via HGT share a more recent ancestor, they exhibit higher sequence similarity. Thus, for each gene family, the percent identity threshold at which DIAMOND BLASTp v.2.1.9 [47] hits were most likely attributable to horizontal rather than vertical transmission was determined. Networks were subsequently constructed and analyzed following this framework. Gene families were defined according to the pangenome analysis (see pangenome construction section), while the genomes used, including those of *P. agglomerans*, were independently downloaded for this analysis.

All species within the order *Enterobacterales* were included for network construction, This taxonomic scope maximized diversity while keeping the database computationally manageable given the available memory (96 GB RAM). Genomes classified as metagenome-assembled genomes (MAGs), labeled as atypical, or originating from large multi-isolate projects were excluded. On June 19, 2025, 128,309 *Enterobacterales* genomes were available in NCBI; of these, 90,282 included predicted protein sets and were retained for further analysis. Genome quality was assessed using NCBI-provided CheckM-based estimates and only genomes with > 90% completeness and < 10% contamination were retained. Where available, average nucleotide identity (ANI) was used to verify taxonomic consistency. Genomes lacking ANI or with unresolved classifications were retained only if no taxonomic conflicts were detected. Certain species were heavily overrepresented in the database, most notably *Escherichia coli*, which accounted for more than 47,000 genomes. To mitigate this imbalance, species-level normalization was applied: species represented by fewer than three high-quality genomes were excluded, whereas those with more than 25 genomes were randomly downsampled to 25. The final database comprised of 3,565 genomes spanning the *Enterobacterales* (Supplementary Fig. 1). Genomes annotated as [*Curtobacterium plantarum*] were treated as *P. agglomerans*, because the [*C. plantarum*] reference genome shares > 95% ANI with the *P. agglomerans* reference genome FDAARGOS_1447.

To determine mPID for each gene family, all proteins within a family were aligned to the family's representative protein. Families that contained four or more members, the representative plus at least three additional proteins, standard deviation of percent identities was calculated and defined as mPID, the mean minus two standard deviations, capturing approximately 95% of *P. agglomerans* proteins within the expected range. Smaller families with fewer than four members, the standard deviation could not be estimated; therefore, to minimize the risk of misclassifying vertically inherited genes as horizontally transferred, mPID was conservatively set to 100%.

Two gene-sharing networks were constructed. The first included representatives from all non-persistent *P. agglomerans* gene families, while the second was limited to non-persistent families functionally annotated as virulence-associated. In both networks, each species was represented as a node. To define edges, representative proteins from the selected *P. agglomerans* gene families were searched against the *Enterobacterales* protein database using DIAMOND BLASTp. Hits with percent identity below the family-specific mPID threshold were filtered out. This approach was designed to detect putative HGT events: if a *P. agglomerans* gene originated in another *Enterobacterales* species, its sequence would still cluster within the same gene family under these criteria. The corresponding genome was thus inferred to encode a "*P. agglomerans*-like" variant of that gene.

For each species s and gene family f we computed $c_{s,f}$ as the total number of unique protein hits of f across all genomes of s that passed the mPID filter, and $p_{s,f}$ as the fraction of genomes in s that contain at least one protein from f. Let N_s be the number of the total number of sampled genomes for the species s. The edge weight between species a and b is defined as

$$W_{a,b} = \frac{\sum_f \left[\min\left(p_{a,f}, p_{b,f}\right) \cdot \left(c_{a,f} + c_{b,f}\right)\right]}{N_a + N_b} \qquad Equation\ 1$$

The shared-prevalence term $\varvec{min}\left(\varvec{p}_{\varvec{a},\varvec{f}}, \varvec{p}_{\varvec{b},\varvec{f}}\right)$ is inherently sampling-invariant, because each $\varvec{p}_{\varvec{s},\varvec{f}}$ is a species-level proportion computed across all genomes of species $\varvec{s}$. In contrast, the abundance term $\varvec{c}_{\varvec{a},\varvec{f}} + \varvec{c}_{\varvec{b},\varvec{f}}$ increases approximately linearly with the total number of genomes sampled $\varvec{N}_{\varvec{a}} + \varvec{N}_{\varvec{b}}$. Consequently, normalizing by $\varvec{N}_{\varvec{a}} + \varvec{N}_{\varvec{b}}$ yields an edge weight that is stable with respect to the number of genomes per species.

Gene cluster analysis and phylogeny

Gene families from the pangenome corresponding to gene clusters of interest were annotated. Only clusters that were not conserved across all species were considered of interest. These included Pantocin A, Herbicolin A, the Type III secretion system (T3SS) *Hrp*-conserved genes (*Hrc*), PNP-2, PNP-3, antibiotic B025670, and AGA phenazine. Reference *P. agglomerans* protein sequences were retrieved individually from UniProt for Pantocin A, the T3SS *Hrc* [48] genes, Dapdiamide E [49], and AGA phenazine [50] (supplementary data). The PNP-2 cluster was obtained from GenBank accession MN329808 [10], B025670 from MT711882 [51], and PNP-3 from MN807451 [7]. Herbicolin A genes were retrieved from *P. agglomerans* strain 9Rz4 (loci NO300_21850 to NO300_21800) [19].

Gene families were searched against reference cluster genes using DIAMOND BLASTp, with a minimum threshold of 70% identity and 80% coverage for both query and subject, under default sensitivity settings. These parameters were chosen to ensure that matched proteins were functionally similar while allowing for homologs detection across diverse genomes. Under these criteria, most genomes were found to possess complete gene clusters, supporting the suitability of these thresholds for downstream comparative analysis. Gene families with at least one match to a cluster gene were retained for subsequent network analysis. Cluster completeness for each *P. agglomerans* genome was determined based on the proportion of corresponding gene families detected. However, BLAST parameters were more permissive than the thresholds used by PPanGGOLiN to define gene boundaries, where a single protein could match multiple gene families, resulting in multiple possible combinations through which a genome could achieve 100% cluster completeness.

P. agglomerans gene family representatives corresponding to proteins in clusters of interest were queried against the *Enterobacterales* database assembled for the gene-sharing network analysis using DIAMOND BLASTp (≥ 80% identity and ≥ 80% coverage, matching the default within-gene family thresholds in PPanGGOLiN). These thresholds were intentionally more permissive than those used for network construction, as downstream phylogenetic inference provides greater power to distinguish vertical inheritance from horizontal transfer. The rationale was that if a given cluster is present across multiple species and *P. agglomerans* sequences cluster with non-*Pantoea* taxa in the phylogeny, this pattern would be consistent with horizontal acquisition; conversely, *P. agglomerans* monophyly would favor vertical inheritance for that gene cluster.

Hits were grouped according to their *P. agglomerans* gene families. Multiple sequence alignments were generated for each cluster of interest and the corresponding families using MAFFT v7.525 with the --auto flag option and subsequently concatenated. When genomes lacked a particular family but contained others from the same cluster, the alignment segment for the missing gene was replaced with a string of repeating “N” characters matching the expected alignment length. Phylogenetic trees were inferred using IQ-TREE v3.0.134. ModelFinder was used to select the best-fit substitution model, and node support was evaluated with 1,000 ultrafast bootstrap replicates. All trees were midpoint rooted.

Declarations

Ethics approval and consent to participate:

Not applicable.

Consent for publication:

Not applicable.

Competing interests:

The authors declare no competing interests.

Funding:

This research and the associated Open Access publication fees were funded by Agriculture and Agri-Food Canada under grant number J-002617, awarded to Ellouze and Svircev.

Author Contribution

Muhammad Sulman: Conceptualization, methodology, experimentation, investigation, data analysis and visualization, data curation, writing original draft, review and editing. Antonet M. Svircev: Conceptualization, validation, funding acquisition, project administration, review and editing. James T. Tambong: methodology, validation, review and editing. Walid Ellouze: Conceptualization, methodology, validation, funding acquisition, resources, project administration, supervision, data curation, review and editing, corresponding author.

Acknowledgement

We gratefully acknowledge the researchers and institutions who generated and deposited the Pantoea agglomerans, other Pantoea species, and Enterobacterales genome assemblies in NCBI GenBank, enabling the comparative analyses conducted in this study.

Data Availability

Genome assemblies for all 165 Pantoea agglomerans strains analyzed in this study are available at NCBI under the accession numbers listed in Table 1.

References

1. Büyükcam A, Tuncer Ö, Gür D, Sancak B, Ceyhan M, Cengiz AB, et al. Clinical and microbiological characteristics of *pantoea agglomerans* infection in children. J Infect Public Health. 2018;11:304–09. https://doi.org/https://doi.org/10.1016/j.jiph.2017.07.020.
2. Casale R, Boattini M, Bianco G, Comini S, Corcione S, Garazzino S, et al. Bloodstream infections by *pantoea* species: Clinical and microbiological findings from a fetrospective study, italy, 2018–2023. Antibiotics. 2023;12:1723. https://doi.org/https://doi.org/10.3390/antibiotics12121723.
3. Kumar RK, Singh NK, Balakrishnan S, Parker CW, Raman K, Venkateswaran K. Metabolic modeling of the international space station microbiome reveals key microbial interactions. Microbiome. 2022;10:102. https://doi.org/https://doi.org/10.1186/s40168-022-01279-y.
4. Luziatelli F, Gatti L, Ficca AG, Medori G, Silvestri C, Melini F, et al. Metabolites secreted by a plant-growth-promoting *pantoea agglomerans* strain improved rooting of *pyrus communis* l. Cv dar gazi cuttings. Front Microbiol. 2020;11:539359. https://doi.org/https://doi.org/10.3389/fmicb.2020.539359.
5. Apine OA, Jadhav JP. Optimization of medium for indole-3-acetic acid production using *pantoea agglomerans* strain pvm. J Appl Microbiol. 2011;110:1235–44. https://doi.org/https://doi.org/10.1111/j.1365-2672.2011.04976.x.
6. Luziatelli F, Ficca AG, Bonini P, Muleo R, Gatti L, Meneghini M, et al. A genetic and metabolomic perspective on the production of indole-3-acetic acid by *pantoea agglomerans* and use of their metabolites as biostimulants in plant nurseries. Front Microbiol. 2020;11:1475. https://doi.org/https://doi.org/10.3389/fmicb.2020.01475.
7. Williams AN, Stavrinides J. *Pantoea* natural product 3 is encoded by an eight-gene biosynthetic gene cluster and exhibits antimicrobial activity against multi-drug resistant *acinetobacter baumannii* and *pseudomonas aeruginosa*. Microbiological Research. 2020;234:126412 https://doi.org/https://doi.org/10.1016/j.micres.2020.126412
8. Giddens SR, Bean DC. Investigations into the in vitro antimicrobial activity and mode of action of the phenazine antibiotic d-alanylgriseoluteic acid. Int J Antimicrob Agents. 2007;29:93–7. https://doi.org/https://doi.org/10.1016/j.ijantimicag.2006.08.028.
9. Smits THM, Duffy B, Blom J, Ishimaru CA, Stockwell VO. Pantocin a, a peptide-derived antibiotic involved in biological control by plant-associated *pantoea* species. Arch Microbiol. 2019;201:713–22. https://doi.org/https://doi.org/10.1007/s00203-019-01647-7.
10. Robinson LJ, Verrett JN, Sorout N, Stavrinides J. A broad-spectrum antibacterial natural product from the cystic fibrosis isolate, *pantoea agglomerans* tx10. Microbiol Res. 2020;237:126479. https://doi.org/https://doi.org/10.1016/j.micres.2020.126479.
11. Shin GY, Asselin JA, Smith A, Aegerter B, Coutinho T, Zhao M, et al. Plasmids encode and can mobilize onion pathogenicity in *pantoea agglomerans*. ISME J. 2025;19:wraf019. https://doi.org/https://doi.org/10.1093/ismejo/wraf019.
12. Sulja A, Pothier JF, Blom J, Moretti C, Buonaurio R, Rezzonico F, et al. Comparative genomics to examine the endophytic potential of *pantoea agglomerans* dapp-pg 734. BMC Genomics. 2022;23:742. https://doi.org/https://doi.org/10.1186/s12864-022-08966-y.
13. Dutkiewicz J, Mackiewicz B, Kinga Lemieszek M, Golec M, Milanowski J. *Pantoea agglomerans*: A mysterious bacterium of evil and good. Part iii. Deleterious effects: Infections of humans, animals and plants. Ann Agric Environ Med. 2016;23:197–205. https://doi.org/https://doi.org/10.5604/12321966.1203878.

14. Pan T, Li Q. Mobile genetic elements in *klebsiella pneumoniae*. J Bacteriol. 2025;207:e0001225. https://doi.org/https://doi.org/10.1128/jb.00012-25.
15. Arnold BJ, Huang IT, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. Nat Rev Microbiol. 2022;20:206–18. https://doi.org/10.1038/s41579-021-00650-4. https://doi.org/.
16. Vanneste JL, Yu J, Cornish DA. Presence of genes homologous to those necessary for synthesis of microcin mcceh252 in strains of *pantoea agglomerans*. Acta Hortic. 2008. https://doi.org/10.17660/ActaHortic.2008.793.58. 793:391 – 96 https://doi.org/.
17. Ishimaru CA, Lansdell TA, Clardy J, Duffy B, Smits THM. The histidine-reversible antibiotic herbicolin o produced by *pantoea vagans* c9-1 is pantocin a. J Plant Pathol. 2017;99:91–7. https://doi.org/http://www.jstor.org/stable/45156724.
18. Williams AN, Stavrinides J. Genetic and evolutionary characterization of the major facilitator superfamily transporters of the antibacterial, pantoea natural product 3. Res Microbiol. 2022;173:103899. https://doi.org/https://doi.org/10.1016/j.resmic.2021.103899.
19. Matilla MA, Evans TJ, Martín J, Udaondo Z, Lomas-Martínez C, Rico-Jiménez M, et al. Herbicolin a production and its modulation by quorum sensing in a *pantoea agglomerans* rhizobacterium bioactive against a broad spectrum of plant-pathogenic fungi. Microb Biotechnol. 2023;16:1690–700. https://doi.org/https://doi.org/10.1111/1751-7915.14193.
20. Kempf H-J, Bauer PH, Schroth MN. Herbicolin a associated with crown and roots of wheat after seed treatment with *erwinia herbicola* b247. Phytopathology. 1993;83:213–16. https://doi.org/https://doi.org/10.1094/Phyto-83-213.
21. Gautreau G, Bazin A, Gachet M, Planel R, Burlot L, Dubois M, et al. Ppanggolin: Depicting microbial diversity via a partitioned pangenome graph. PLoS Comput Biol. 2020;16:e1007732. https://doi.org/https://doi.org/10.1371/journal.pcbi.1007732.
22. Versteeg B, Bruisten SM, Pannekoek Y, Jolley KA, Maiden MCJ, van der Ende A, et al. Genomic analyses of the *chlamydia trachomatis* core genome show an association between chromosomal genome, plasmid type and disease. BMC Genomics. 2018;19:130. https://doi.org/https://doi.org/10.1186/s12864-018-4522-3.
23. Eppinger M, Worsham PL, Nikolich MP, Riley DR, Sebastian Y, Mou S, et al. Genome sequence of the deep-rooted *yersinia pestis* strain angola reveals new insights into the evolution and pangenome of the plague bacterium. J Bacteriol. 2010;192:1685–99. https://doi.org/https://doi.org/10.1128/jb.01518-09.
24. Rouli L, Robert MMB, Ndiaye C, La Scola M, Raoult B. Genomic analysis of three african strains of *bacillus anthracis* demonstrates that they are part of the clonal expansion of an exclusively pathogenic bacterium. NMNI. 2014;2:161–9. https://doi.org/https://doi.org/10.1002/nmi2.62.
25. Botelho J, Cazares A, Schulenburg H. The eskape mobilome contributes to the spread of antimicrobial resistance and crispr-mediated conflict between mobile genetic elements. Nucleic Acids Res. 2023;51:236–52. https://doi.org/https://doi.org/10.1093/nar/gkac1220.
26. Partridge SR. Analysis of antibiotic resistance regions in gram-negative bacteria. FEMS Microbiol Rev. 2011;35:820–55. org/10.1111/j.1574-6976.2011.00277.x. https://doi.org/https://doi..
27. Carlin A, Shi W, Dey S, Rosen BP. The *ars* operon of *escherichia coli* confers arsenical and antimonial resistance. J Bacteriol. 1995;177:981–6. https://doi.org/https://doi.org/10.1128/jb.177.4.981-986.1995.
28. Hammerum AM, Flannagan SE, Clewell DB, Jensen LB. Indication of transposition of a mobile DNA element containing the *vat*(d) and *erm*(b) genes in *enterococcus faecium*. Antimicrob Agents Chemother. 2001;45:3223–5. https://doi.org/https://doi.org/10.1128/aac.45.11.3223-3225.2001.
29. Lee S-Y, Park Y-J, Yu JK, Jung S, Kim Y, Jeong SH, et al. Prevalence of acquired fosfomycin resistance among extended-spectrum β-lactamase-producing *escherichia coli* and *klebsiella pneumoniae* clinical isolates in korea and is26-composite transposon surrounding *fosa3*. J Antimicrob Chemother. 2012;67:2843–47. https://doi.org/https://doi.org/10.1093/jac/dks319.
30. Gu BL, She Y, Pei GK, Du Y, Yang R, Ma LX, et al. Systematic analysis of prophages carried by *porphyromonas gingivalis*. Infect Genet Evol. 2023;113:105489. https://doi.org/https://doi.org/10.1016/j.meegid.2023.105489.
31. Ou Y, Yan J, Wang Y, Chen L. Recent discovery of diverse prophages located in genomes of *vibrio* spp. And their implications for bacterial pathogenicity, environmental fitness, genome evolution, food safety, and public health. Foods. 2025;14:403. https://doi.org/https://doi.org/10.3390/foods14030403.
32. Chang TH, Pourtois JD, Haddock NL, Furukawa D, Kelly KE, Amanatullah DF, et al. Prophages are infrequently associated with antibiotic resistance in *pseudomonas aeruginosa* clinical isolates. mSphere. 2025;10:e00904–24.

https://doi.org/https://doi.org/10.1128/msphere.00904-24.

33. Duan G, Liu X, Zhang S, Chai M, Peng Z, Lin Z, et al. An emerging bacterial leaf disease in rice caused by *pantoea ananatis* and *pantoea eucalypti* in northeast china. Microorganisms. 2025;13:1376. https://doi.org/https://doi.org/10.3390/microorganisms13061376.
34. De Armas S, Galván GA, Lapaz MI, González-Barrios P, Vicente E, Pianzzola MJ, et al. Phylogeny and identification of *pantoea* species associated with bulb rot and bacterial leaf blight of onion crops in uruguay. Plant Dis. 2022;106:1216–25. https://doi.org/https://doi.org/10.1094/PDIS-06-21-1140-RE.
35. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. Checkm: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 2015 https://doi.org/https://doi.org/10.1101/gr.186072.114
36. Shaw J, Yu YW. Fast and robust metagenomic sequence comparison through sparse chaining with skani. Nat Methods. 2023;20:1661–65. https://doi.org/https://doi.org/10.1038/s41592-023-02018-3.
37. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A, Bakta. Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb Genom. 2021;7. https://doi.org/https://doi.org/10.1099/mgen.0.000685.
38. Shimoyama Y, Cogclassifier. A tool for classifying prokaryote protein sequences into cog functional category. 2022 https://doi.org/https://github.com/moshi4/COGclassifier
39. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with genomad. Nat Biotechnol. 2024;42:1303–12. https://doi.org/https://doi.org/10.1038/s41587-023-01953-y.
40. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. Checkv assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2021;39:578–85. https://doi.org/10.1038/s41587-020-00774-7. https://doi.org/.
41. Bolduc B. Vcontact3: Viral genome clustering and taxonomy tool. Bitbucket repository. 2025 https://doi.org/https://bitbucket.org/MAVERICLab/vcontact3/
42. Robertson J, Nash JHE. Mob-suite: Software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb Genom. 2018;4:e000206. https://doi.org/https://doi.org/10.1099/mgen.0.000206.
43. Steinegger M, Söding J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35:1026–28. https://doi.org/https://doi.org/10.1038/nbt.3988.
44. Tang X, Shang J, Ji Y, Sun Y, Plasme. A tool to identify plasmid contigs from short-read assemblies using transformer. Nucleic Acids Res. 2023;51:e83. https://doi.org/https://doi.org/10.1093/nar/gkad578.
45. Liu B, Zheng D, Zhou S, Chen L, Yang J, Vfdb. 2022: A general classification scheme for bacterial virulence factors. Nucleic Acids Res. 2022;50:D912-D17 https://doi.org/https://doi.org/10.1093/nar/gkab1107
46. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al. Amrfinderplus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021;11:12728. https://doi.org/https://doi.org/10.1038/s41598-021-91456-0.
47. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using diamond. Nat Methods. 2021;18:366–68. https://doi.org/10.1038/s41592-021-01101-x. https://doi.org/.
48. Bogdanove AJ, Beer SV, Bonas U, Boucher CA, Collmer A, Coplin DL, et al. Unified nomenclature for broadly conserved *hrp* genes of phytopathogenic bacteria. Mol Microbiol. 1996. https://doi.org/10.1046/j.1365-2958.1996.5731077.x. 20:681 – 83 https://doi.org/.
49. Kamber T, Lansdell TA, Stockwell VO, Ishimaru CA, Smits THM, Duffy B. Characterization of the biosynthetic operon for the antibacterial peptide herbicolin in *pantoea vagans* biocontrol strain c9-1 and incidence in *pantoea* species. Appl Environ Microbiol. 2012;78:4412–19. https://doi.org/https://doi.org/10.1128/AEM.07351-11.
50. Giddens SR, Feng Y, Mahanty HK. Characterization of a novel phenazine antibiotic gene cluster in *erwinia herbicola* eh1087. Mol Microbiol. 2002;45:769–83. org/10.1046/j.1365-2958.2002.03048.x. https://doi.org/https://doi..
51. Williams AN, Sorout N, Cameron AJ, Stavrinides J. The integration of genome mining, comparative genomics, and functional genetics for biosynthetic gene cluster identification. Front Genet. 2020;11. https://doi.org/https://doi.org/10.3389/fgene.2020.600116.

Figures

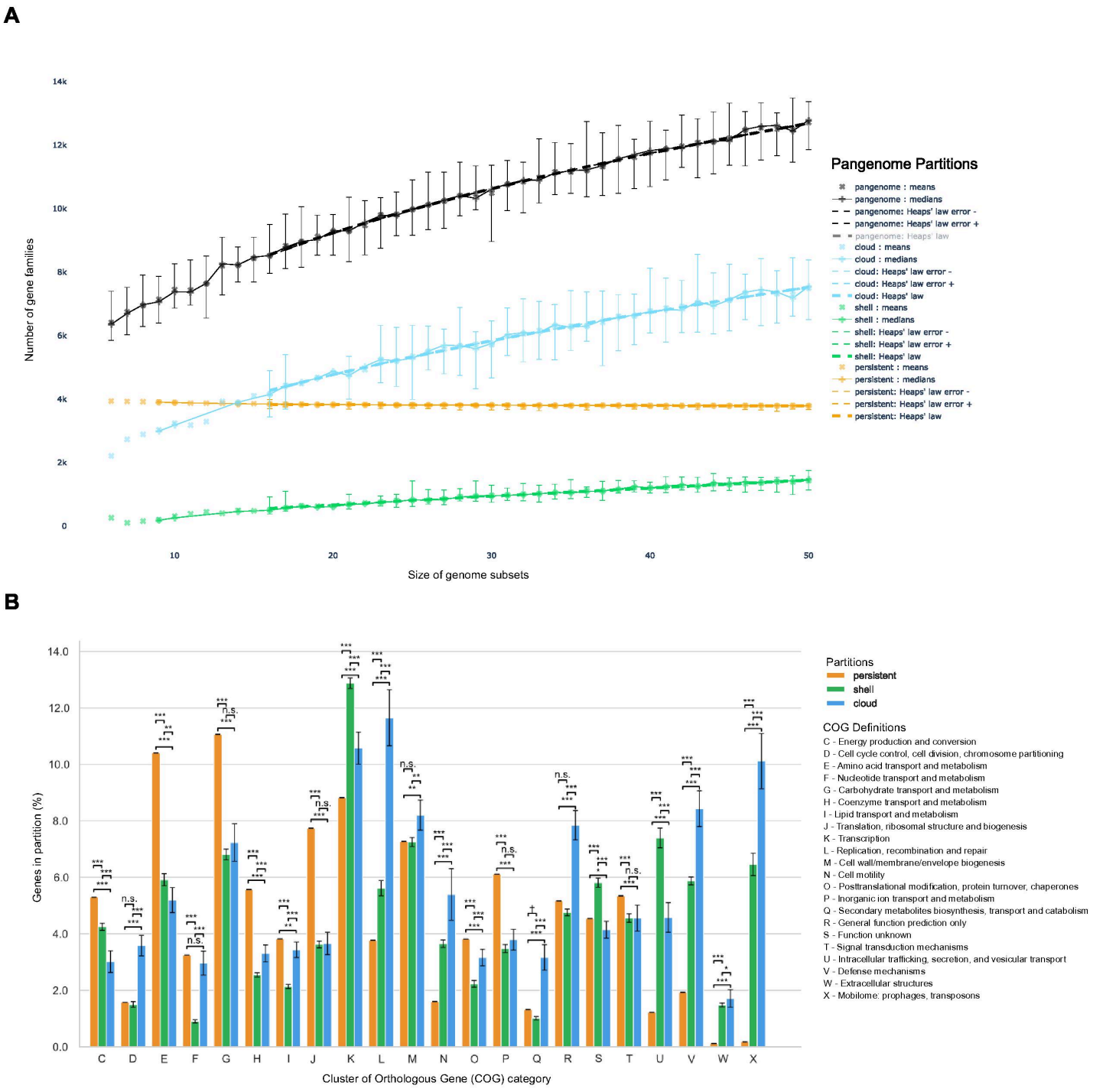


Figure 1

Summary of the *Pantoea agglomerans* pangenome and differential COG composition across its partitions. (A) Rarefaction curves showing the number of gene families as a function of the number of genomes sampled. Curves are shown for the total pangenome (black), the persistent genome (orange), the shell genome (green), and the cloud genome (light blue). (B) Bar chart showing the mean percentage of functionally annotated genes assigned to each Cluster of Orthologous Groups (COG) category, partitioned by pangenome component: persistent (core genes; orange), shell (green), and cloud (accessory genes; blue). The y-axis indicates the

percentage of genes within a given partition for each genome that fall under a specific COG category. Error bars represent 90% confidence intervals of the mean. Statistical comparisons between partitions for each COG category were performed using a pairwise Tukey's HSD test. Brackets indicate significant differences in mean percentages, with significance levels denoted as: *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$), † ($p < 0.10$), and n.s. (not significant). The key on the right provides the full description for each COG category code.

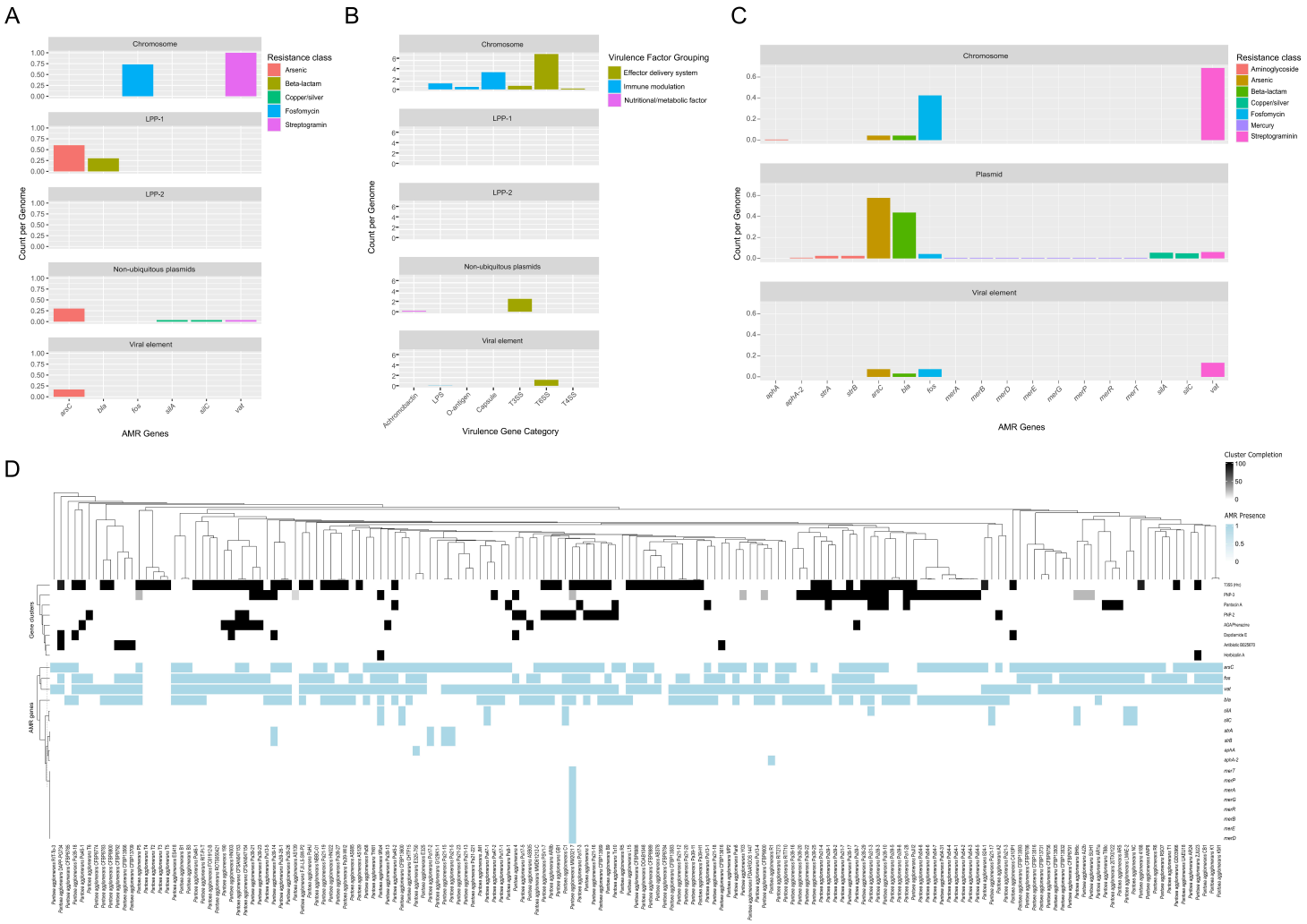


Figure 2

Genomic distribution and co-occurrence of non-persistent AMR genes, virulence factors, and gene clusters in *Pantoea agglomerans*. Panels A and B show the average abundance and distribution of antimicrobial resistance (AMR) genes (A) and virulence factors (B), across 30 complete genomes. Facets correspond to genomic components: chromosome, LPP-1, LPP-2, non-ubiquitous plasmids, and viral elements. Bars represent the mean number of genes per genome, colored by resistance class (A) or virulence factor group (B; e.g., adherence, effector delivery systems, exoenzymes). (C) Summary of AMR gene abundance across 165 genomes (complete and draft), with facets representing chromosome, plasmid, and viral elements. Colors indicate resistance classes. (D) Distribution of AMR genes and biosynthetic or functional gene clusters across genomes, ordered by hierarchical clustering of the pangenome presence–absence matrix; the top heatmap (black–white gradient) represents gene cluster presence and completeness, while the bottom heatmap (light blue) denotes AMR gene presence (blue) or absence (white).

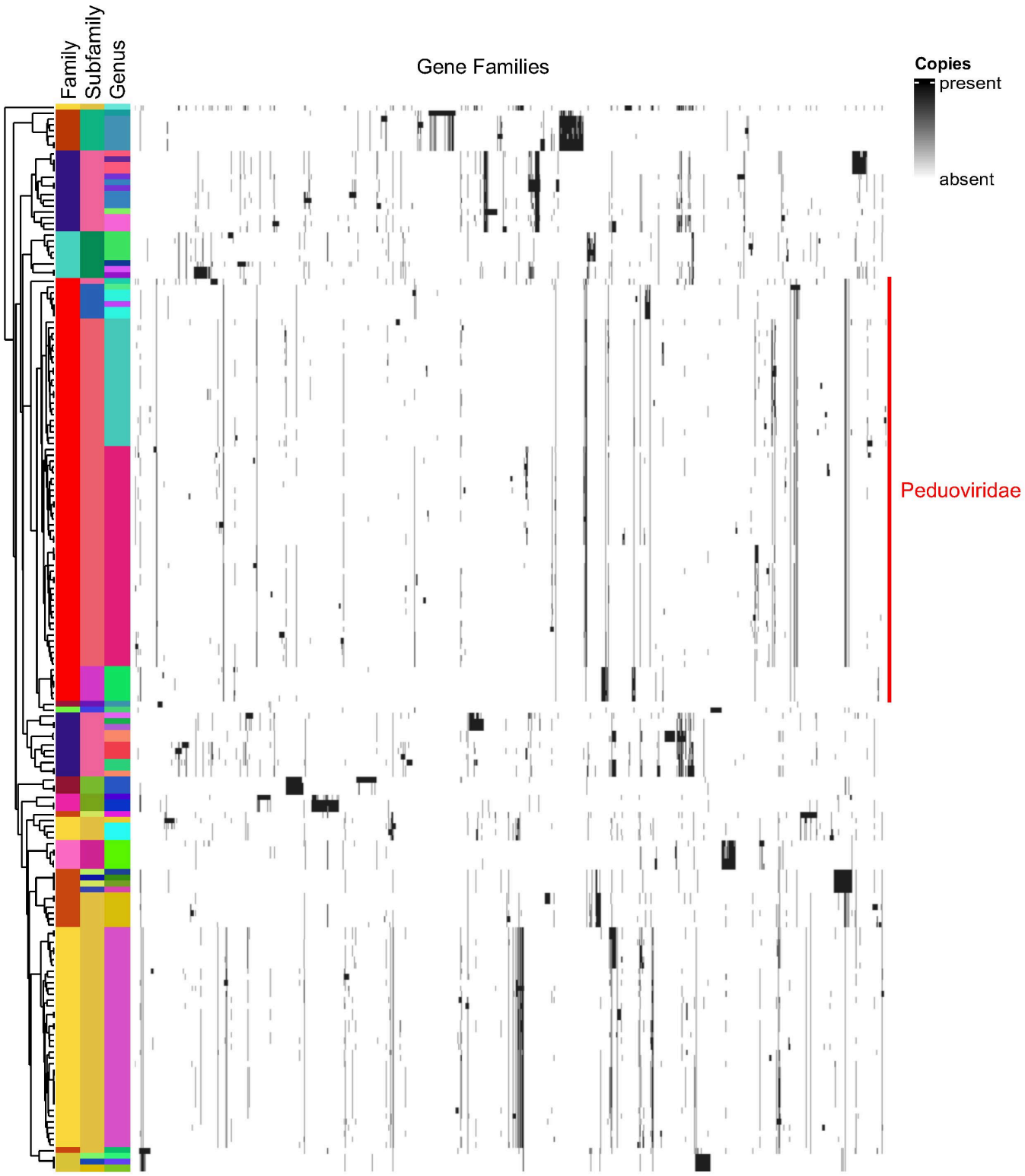


Figure 3

Gene family composition of 184 bacteriophages identified across 165 *Pantoea agglomerans* genomes. Colored tiles on the left indicate taxonomic assignments generated using vConTACT3, with each color representing a distinct taxon at the family, subfamily, or genus level. Only one named taxon, *Peduoviridae*(highlighted with a red bar), was identified, while all other phages remained unclassified at these ranks. The main heatmap shows the presence (black) or absence (white) of gene families across all identified phages, with each column corresponding to a distinct gene family. The leftmost dendrogram, generated by hierarchical clustering of the gene family presence–absence matrix, illustrates the relatedness among phages based on shared gene content.

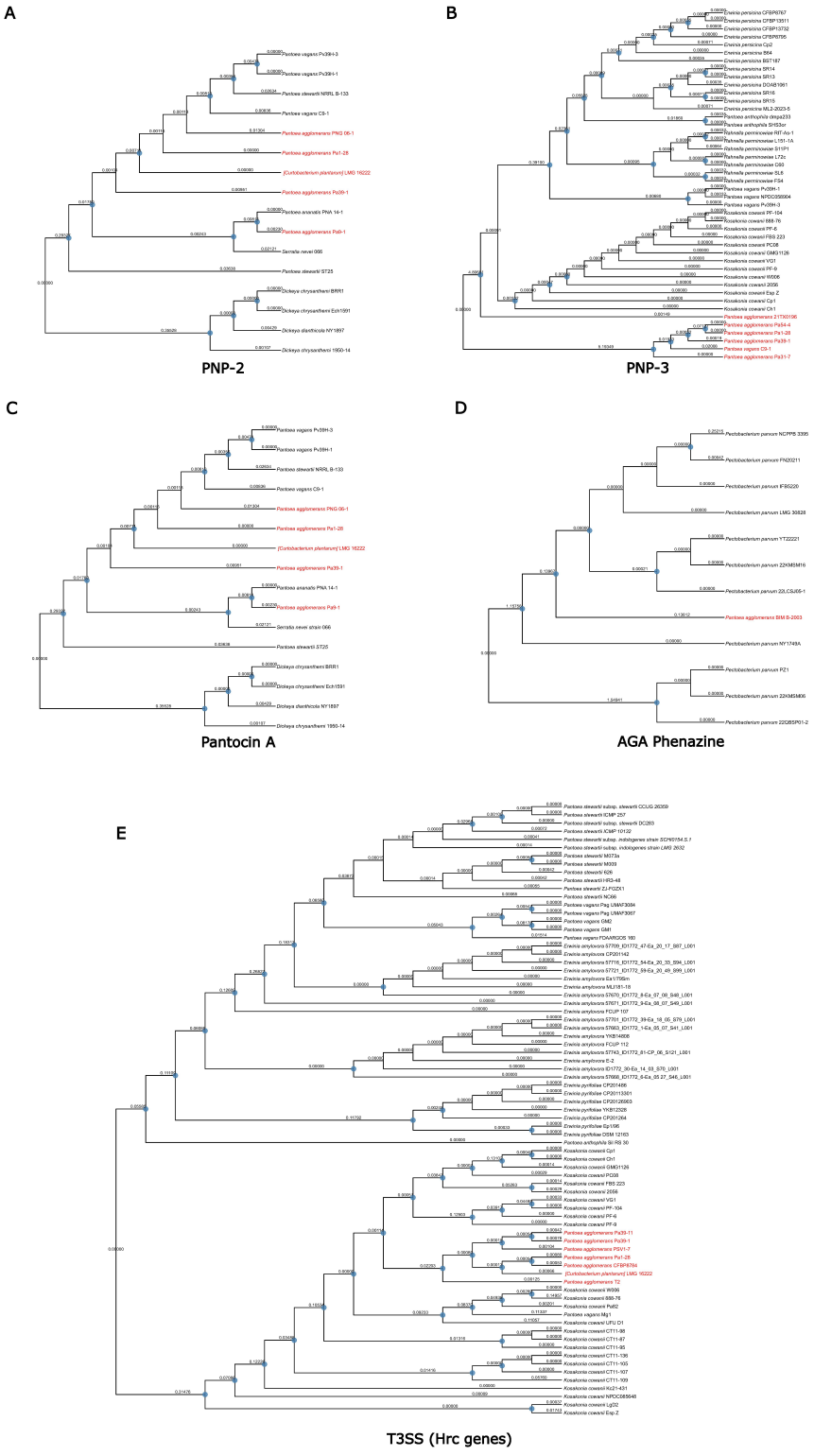


Figure 4

Phylogenetic trees of selected *Pantoea agglomerans* gene clusters. Maximum-likelihood trees showing the evolutionary relationships of *P. agglomerans* gene clusters across *Enterobacterales* species for: (A) PNP-2, (B) PNP-3, (C) Pantocin A, (D) AGA phenazine, and (E) Type III secretion system (*Hrc*genes). Branch lengths are scaled to substitutions per site, and blue circles mark nodes with ultrafast bootstrap support >70. *P. agglomerans*strains are highlighted in red.

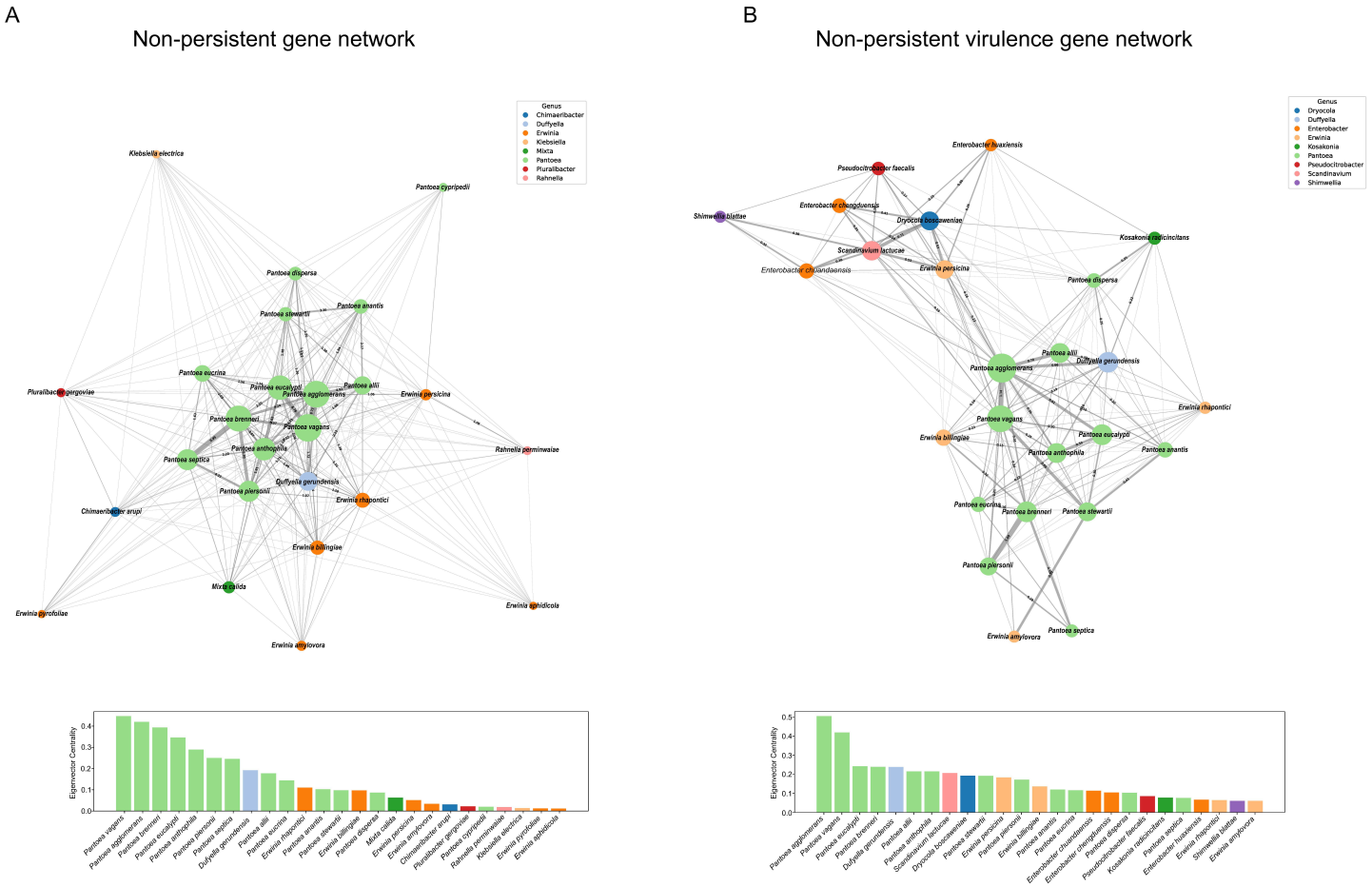


Figure 5

Gene-sharing networks of *Pantoea agglomerans* gene families with other *Enterobacterales*species. In each panel, nodes represent species, with each genus assigned a distinct color. Edge thickness reflects the number of shared gene families, and node size is proportional to eigenvector centrality. Only the top 25 species per network are shown, ranked by eigenvector centrality. The six panels correspond to networks constructed from different subsets of *P. agglomerans* genes: (A) non-persistent genes, (B) non-persistent virulence genes. Below each network, bar plots display the eigenvector centrality for each node.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryfilesSearchedClustersSequences.zip
- SupplementaryFigure1Enterobacteralesdatabase.pdf