BMC Genomics

RESEARCH

Open Access

Unraveling the genetic traits and functional diversity of the pan-genome in *Pantoea dispersa*

Shiyao He[1], Qi Ding[1], Wenting Wu[1], Yun Zhang[1], Yan Kang[1], Yang Meng[1], Sirui Zhu[1] and Jinyuan Wu[2*]

Abstract

Background Medical devices are crucial in modern healthcare, but commonly used clinical tools such as cotton swabs can be easily contaminated by microorganisms (such as *Pantoea*), becoming vectors for pathogens and leading to patient infections or more severe outcomes. Despite the dual nature of the *Pantoea* that has garnered significant attention, research investigating *Pantoea dispersa* (*P. dispersa*) remains limited. This study conducted a pan-genome analysis of three isolates and 57 *P. dispersa* strains from NCBI to investigate their evolutionary relationships, population structure, and functional characteristics.

Results Whole-genome analysis revealed high genomic diversity among 60 strains of *P. dispersa*, identifying 6,791 orthologous gene clusters (OGs), with core genes accounting for 45.1% and accessory genes accounting for 54.9%. Additionally, 2,185 gene clusters were not annotated in the reference genome. Further analysis demonstrated that 782 gene clusters were annotated as 406 VFs that were unevenly distributed among different strains and primarily associated with nutritional or metabolic factors, motility, and immune modulation. This study also identified four VFs genes related to the type III secretion system (T3SS) and observed some VFs present only in specific genetic clusters. In the analysis of antibiotic resistance genes (ARGs), 12 ARGs were identified, with nine being highly conserved across all isolates, and resistance mechanisms primarily involved antibiotic efflux and antibiotic target alteration. Secondary metabolite analysis identified 289 gene clusters, with 23 matching known gene clusters, while the rest were new discoveries, including arylpolyene, NRPS, and terpene types. These results reveal the complex virulence factors (VFs) and secondary metabolite genes in *P. dispersa*, providing significant insights into its genetic diversity and biological significance.

Conclusion This study provides the first pan-genome framework for *P. dispersa*, along with a map of its VFs, ARGs, and secondary metabolite gene clusters. This study provides a deep insight into the genetic diversity and potential biological significance of *P. dispersa*, offering valuable references for leveraging its unique strain characteristics and metabolic capabilities in industrial production and clinical therapy.

Keywords *Pantoea dispersa*, Genomic diversity, Pan-genome, Antibiotic resistance genes, Virulence factors, Secondary metabolite gene clusters

*Correspondence:
Jinyuan Wu
2312125704@qq.com

[1]Jiangxi Medical Device Testing Center, Nanchang, Jiangxi Province 330029, PR China
[2]College of Bioscience and Bioengineering, Jiangxi Agricultural University, Nanchang, Jiangxi Province 330045, PR China

Introduction

Medical devices are an essential aspect of modern medical practice [1], yet contamination with microorganisms can elevate the risk of clinical infections and consequently lead to potentially serious adverse outcomes. Instruments that are not thoroughly sterilized or have been contaminated may transmit pathogens, causing secondary infections, local abscesses, or systemic infections, and residues from microorganisms can also trigger immune responses [2, 3]. As a commonly used medical consumable, cotton swabs are indispensable for wound cleaning, disinfection, medication application, and sample collection. However, owing to their material characteristics, swabs may inherently carry microorganisms from the raw materials, such as those belonging to *Pantoea* and *Bacillus* if proper sterilization is not ensured during manufacturing. This could result in the swabs serving as vectors for pathogens, thus posing a significant threat to patient health [3].

Pantoea belongs to the family Enterobacteriaceae and is widely present on plant surfaces [4], seeds [5], soil [6], water [7], and in wounds, blood [8], and urine [9] of animals and humans. Currently, *Pantoea* includes at least seven known species capable of causing plant diseases, such as *P. agglomerans*, *P. ananatis*, *P. citrea*, *P. dispersa*, *P. punctata*, *P. stewartii* and *P. terra* [10]. Some of these species such as *P. ananatis*, can infect eucalyptus, corn, and rice while *P. dispersa* is often associated with seed infections. The pathogenicity of *Pantoea* in plants is typically linked to the type III secretion system (T3SS). In certain cases, *Pantoea* can also cause infections in humans, including wound infections, pneumonia, and urinary tract infections [11], and in extreme cases, can lead to sepsis [12–14]. *P. agglomerans* has received considerable attention due to repeated infections. Additionally, Fabio Rezzonico et al. used multi-locus phylogenetic analysis and fluorescent amplified fragment length polymorphism (fAFLP) fingerprinting techniques to compare genotype and phenotype markers between plant-derived strains and clinical isolates, aiming to identify characteristic markers distinguishing plant strains from clinical strains [15]. Despite the development of sequencing technologies, which enable the acquisition of microbial genomic information without the need for culturing and provide a robust foundation for genome-level analysis of strains [16], the precise mechanisms by which *Pantoea* as human pathogens remain elusive [10].

Due to their diverse biosynthesis [17, 18] and biodegradation abilities [19, 20], *Pantoea* exhibits significant value in agriculture, environmental remediation, and clinical applications [21, 22]. For example, *P. agglomerans* has been demonstrated to inhibit the growth of fungal pathogens such as *Fusarium graminearum* by secreting a substance that targets the lipid rafts of the fungi, effectively reducing plant diseases caused by the latter [23]. Additionally, *P. dispersa* is widely present in rice seeds and can promote plant growth by producing indole-3-acetic acid (IAA) and other plant hormones [24], indicating its great potential for agricultural applications. Low-molecular-weight lipopolysaccharide IP-PA1 isolated from *P. agglomerans* has exhibited broad therapeutic properties [25] and holds great potential in the prevention and treatment of diseases in humans and animals. Although *Pantoea* is increasingly recognized for its dual beneficial and harmful characteristics, existing research largely focuses on specific species such as *P. stewartii* [26] and primarily centers on the analysis of single strains [27]. However, comprehensive functional studies of the entire population remain limited, and the systematic evolutionary history of the population remains unknown.

This study sequenced three strains of *P. dispersa* isolated from a batch of microbiologically contaminated cotton swabs and analyzed their genomic characteristics. In conjunction with existing data from 57 strains of *P. dispersa* in the National Center for Biotechnology Information (NCBI) database, we analyzed the phylogenetic relationships and population structure characteristics of *P. dispersa* strains and constructed its pan-genome framework. Based on this, we further investigated the distribution characteristics and functional potential of virulence factors (VFs), antibiotic resistance genes (ARGs), and secondary metabolite gene clusters in *P. dispersa*. This study explores the genetic diversity and potential biological significance of *P. dispersa*. By analyzing its secondary metabolite production capabilities, VFs, and ARGs, we provide a theoretical foundation for the application of its secondary metabolite synthesis abilities in industrial production. Additionally, based on its antibiotic resistance and VF characteristics, we offer valuable references for the clinical treatment of *P. dispersa* infections.

Methods

Isolation and cultivation of *P. dispersa*

In this study, the strains we utilized were isolated from three swabs that failed sterility testing. These swabs were identified during a routine quality supervision and inspection of marketed medical devices conducted by the Jiangxi Province Medical Device Testing Center. The batch of non-conforming swabs was supplied by a manufacturer specializing in healthcare materials. Briefly, the entire swab was placed into sterile glass tubes containing 50 mL of fluid thioglycollate medium and soybean-casein digest broth medium separately and these were incubated at 33 °C for 3 days. Then, 70 µL of the turbid culture was spread on tryptic soy agar plates and incubated under the same conditions for 3 days. Next, single colonies were picked from the agar plates and streaked

onto the soybean-casein digest medium until pure colonies were obtained (Additional file 1). Ultimately, we successfully isolated three strains of *P. dispersa* from swabs contained in three different packages.

DNA extraction and whole genome sequencing

The three isolated strains were enriched in 50 ml sterile centrifuge tubes for 3 days, respectively. Then, 10 mL of cultured tryptic soy fluids was centrifuged at 3,000 rpm for 10 min. The cells were washed twice with sterile 0.9% NaCl solution. The washed cells were used for subsequent DNA extraction. Genomic DNA extraction and library preparation were performed using the metagenomic DNA library preparation kit (Hangzhou Biotechnology Co., Ltd., catalog number MD001) and an NGSmaster automated instrument. After library construction, the quality of the libraries was assessed using a Qubit® 3.0 fluorometer, and the effective concentrations of the libraries were determined via real-time quantitative PCR (qRT-PCR). The libraries were subsequently pooled and normalized, and high-throughput sequencing was conducted using the Illumina MiSeq platform (MiSeq Reagent Kit v2) to generate paired-end sequencing reads of 150 bp length for each isolate.

Genome assembly and annotation

The original sequencing data for each strain were processed using Trimmomatic (version 0.39) [28] with parameters set as LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, and MINLEN:50 (Additional file 2). Subsequently, genome assembly was performed using SPAdes (version 3.15.5) [29] with k-mer parameters set to 33, 57, 77, 97 and 127. The assembly results were evaluated using QUAST (version 5.2.0) [30], and gene prediction was conducted using GeneMark.hmm (version 3.42) [31]. For each strain's assembly results, BLAST (version 2.15.0) was used to align against the NCBI 16 S ribosomal RNA database and NT database to obtain species annotation information. To ensure the precision of the species identification, we additionally employed GTDB-Tk (version 2.4.0) for annotating the isolates (Additional file 3).

Whole-genome phylogenetic analysis

To construct the phylogenetic tree of the three isolates along with other strains of *P. dispersa* reported previously, we retrieved 57 high-quality whole genomes with GCF identifiers from the NCBI genome database (Additional file 4).

We used bcftools (version 1.20) [32] to filter out variant sites with an alternate allele frequency lower than 0.25 and those with more than 20% sample absence, to construct a phylogenetic tree based on single nucleotide variants (SNVs) [33, 34]. We then concatenated SNPs from each isolate for input into RAxML (version 8.2.12) [35] to construct a maximum likelihood tree (Additional file 5). The three was visualized using iTOL (version 6.9.1). We used FastANI (version 1.34) [36] to perform average nucleotide identity (ANI) analysis on the whole-genome sequences of 60 *P. dispersa* strains, obtaining the pairwise whole-genome average nucleotide identity among the species.

Population structure and diversity analysis

The population structure of 60 strains of *P. dispersa* was studied using STRUCTURE (version 2.3.4) [37] by determining the optimal number of clusters (k) corresponding to the lowest Bayesian information criterion (BIC). Genetic diversity parameters, including nucleotide diversity (π), fixation index (Fst), and Tajima's D statistic, were calculated using VCFtools (version 0.1.16) within non-overlapping windows of 10,000 bp to analyze diversity within each genetic cluster. Nucleotide diversity (π) measures the level of polymorphism at different nucleotide positions within the genomes of a group of individuals, and the genome-wide fixation index (Fst) is an indicator of population differentiation. Fst ranges from 0 to 1, where Fst = 0 indicates genetic similarity between two populations, and Fst = 1 indicates significant genetic differences between populations. Tajima's D test allows us to determine whether a population is evolving neutrally or is under natural selection.

Pan-genome analysis

Orthologous gene clusters (OGs) were identified using OrthoFinder (version 2.5.2)(Additional file 6) [38]. OGs present in all strains (100%) were defined as core genes, those present in over 95% of the strains were defined as soft core genes, those present in 5–95% of the strains were defined as shared genes, and OGs observed in fewer than 5% of the strains were defined as cloud genes. Given that each orthologous group (OG) contains multiple homologous gene sequences from different strains, to facilitate subsequent analyses and construct a functional map of the *P. dispersa* pan-genome, we selected the longest protein sequence from each OG as the representative sequence. This selection was performed using a custom Bash script (Additional file 1). These representative sequences were subsequently annotated against various databases, including VFDB and CARD, to elucidate their functional characteristics.

Functional annotation

Using the Diamond (version 2.1.9), representative sequences of each OG were aligned against the VFDB setA database (version 2024.8.30, parameters set to: --evalue 1e-5 --min-score 60 --block-size 40.0) [39] to identify VFs within the pan-genome of *P. dispersa*. The online RGI tool (version 6.0.3) was utilized to align

each OG against the CARD database (version 3.3.0) with parameters set to "Perfect and Strict hits only" and "Exclude nudge" to determine ARGs. Additionally, the whole genome data of 60 *P. dispersa* strains were annotated using antiSMASH (version 7.1.0) [40] to identify secondary metabolite biosynthesis gene clusters (BGC). To facilitate subsequent analysis, the BiG-SCAPE (version 1.1.5) [41] tool was used to construct gene cluster families (GCFs) with a distance cutoff value of 0.2 (cutoffs 0.2). The longest sequence BGC in each GCF was selected as the representative of that GCF, and the genes from these GCFs were mapped onto the pan-genome framework of *P. dispersa* to generate a landscape of secondary metabolites.

Results

De Novo assembly of *P. dispersa*

In this study, three strains of *P. dispersa* were isolated and sequenced, and genomic data from 57 previously sequenced strains of *P. dispersa* were also collected from the NCBI database, resulting in a total of 60 strains of *P. dispersa* with genomic information (Additional file 8). To obtain the draft-genomes of three novel strains, we performed *de novo* assembly. Their average genome sizes were 4.71 Mb, 4.43 Mb, and 4.37 Mb, respectively, with scaffold counts (length more than 200 bp) of 52, 37, and 115 and N50 values of 236,062 bp, 423,018 bp, and 72,138 bp, respectively (Additional file 8). The genome sizes of the 57 strains collected from NCBI ranged from 4.12 Mb to 5.08 Mb, with scaffold counts ranging from 1 to 178 and N50 values of between 36.96 kb and 4.12 Mb (Additional file 4). Overall, the three newly isolated strains of *P. dispersa* did not exhibit significant differences in genome size or scaffold count compared to those of previously sequenced strains of *P. dispersa*.

To conduct genomic studies, we selected the *P. dispersa* strain *Lsch* (GCF_019890955) as the reference genome, as it is annotated as the reference genome in the NCBI database. This genome is 4.66 Mb in size, containing 1 chromosome (3.83 Mb) and 2 plasmids (0.66 Mb and 0.17 Mb), with a GC content of 57.5%. The comparative genomics circular map illustrates the whole-genome diversity of 60 strains of *P. dispersa*. In comparison to the reference genome *Lsch*, gene deletions were notably more frequent at positions located at 600 kb, 700 kb, 1,300 kb, 1,500 kb, 2,900 kb, and 3,400 kb. It should be noted that the chromosomal sequence of the reference genome is depicted from 1 to 3,920 kb, whereas the sequences for plasmid 1 and plasmid 2 are presented from 3,920 to 4,096 kb and from 4,096 to 4,770 kb, respectively. Particularly, the region of plasmid sequences from 4,000 kb to 4,800 kb exhibits the most significant deletion phenomenon, making it the area with the greatest variation across the entire *P. dispersa* genome (Fig. 1A). Then, we conducted SNP site detection on 59 strains of *P. dispersa* and identifying a total of 284,207 SNP sites. On average, each strain contained 51,447 SNPs (ranging from 43,266 to 55,663), including 4,423 heterozygous SNPs and 47,023 homozygous SNPs. Additionally, there were approximately 690 SNPs per 1,000 bp among strains. These data reveal rich genetic diversity within the population of *P. dispersa*.

Population diversity of 60 strains of *P. dispersa*

Through population structure analysis, we identified four genetic clusters among the 60 strains of *P. dispersa* based on the CV error (Fig. 1B). The largest genetic cluster P3 included 32 isolates, the second-largest cluster P4 included 14 isolates, and the remaining two clusters each contained 7 isolates (Fig. 1C). To better understand

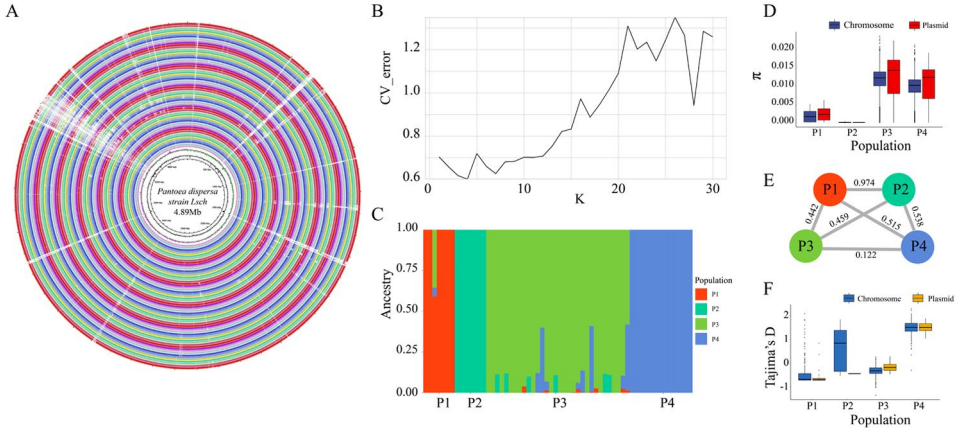

Fig. 1 Comparative genomics and diversity analysis of 60 *P. dispersa* strains. **(A)** Circos plot of comparative genomics for 60 strains of *P. dispersa* (using *P. dispersa* strain *Lsch* [GCF_019890955] as the reference genome). **(B)** CV error plot from population structure analysis exhibiting the lowest CV error when k = 4. **(C)** Population structure analysis identified four groups labeled as P1 to P4. **(D)** Box plots indicating the distribution of nucleotide diversity (π) measurements calculated in 10 kb windows for P1, P2, P3 and P4. **(E)** Fst values between the four populations, where gray links between groups indicate the Fst values between the two groups. **(F)** Box plots indicating the distribution of Tajima's D measurements calculated in 10 kb windows for P1, P2, P3 and P4

the differences between these genetic clusters, we analyzed their whole-genome nucleotide diversity (π). The results revealed that genetic cluster P3 possessed the highest average nucleotide diversity ($\pi = 1.18 \times 10^{-2}$), and this is followed by genetic cluster P4 ($\pi = 9.89 \times 10^{-3}$), whereas the nucleotide diversity of genetic cluster P1 is $\pi = 1.83 \times 10^{-3}$. Genetic cluster P2 exhibited the lowest nucleotide diversity ($\pi = 7.57 \times 10^{-5}$). Additionally, among these populations, plasmids exhibit the highest genetic diversity, followed by chromosomes (Fig. 1D). Notably, genetic cluster P2 exhibited generally lower π values and a concentrated distribution, indicating that the phylogenetic relationships within this cluster were the closest.

To assess the degree of differentiation between genetic clusters, we calculated the Fst and determined that the Fst value between genetic clusters P1 and P2 was the highest (0.974), indicating that these two genetic clusters were highly differentiated. Next, the Fst values between P4 and P1or P2 were similar (Fst = 0.515, 0.538). The fixation indices between genetic cluster P3 and P1or P2 were both less than 0.5 (Fst = 0.442, 0.459), and the lowest Fst value was observed between P3 and P4 (Fst = 0.122) (Fig. 1E). Meanwhile, we calculated the whole-genome Tajima's D diversity index for each genetic cluster to examine the abundance of rare alleles within each cluster. According to our analysis, the Tajima's D values for genetic clusters P1 and P2 were both greater than −0.5, indicating that P1 and P2 are primarily evolving under neutral conditions but are also influenced by partial selective evolution. The Tajima's D value for genetic cluster P2 was 0.46, suggesting a deficit of rare alleles in this cluster. The Tajima's D value for genetic cluster P4 was 1.50, indicating the absence of rare alleles and possibly leading to population contraction (Fig. 1F).

Based on the results of SNP analysis, we aligned 28,704 SNP variant sites for whole-genome phylogenetic analysis. It was revealed that among the three strains of isolated *P. dispersa*, strain B01 shares a closer phylogenetic relationship with GCF_037150095, while B02 and B03 are genetically closest to each other and cluster together with GCF_001477155 and GCF_001476735 (Fig. 2A). By labeling the genetic clusters, we determined that the P3 and P4 genetic clusters were not completely separated, and this is consistent with the Fst value of 0.122 between the two clusters. The results of the ANI analysis were consistent with the phylogenetic analysis, further validating the existence of the four genetic clusters (Fig. 2B). It is worth noting that the ANI values within the P2 genetic cluster approach 100%, and this is indicative of a high level of genetic similarity among the strains. This observation also raises the possibility that these strains could be of clonal origin.

Construction of the *P. dispersa* pan-genome

We constructed a pan-genome of *P. dispersa* (PdisPan) based on the 60 strains and identified 6,791 OGs with 3,063 core gene clusters accounting for 45.10% of the entire pan-genome, with 503 soft-core gene clusters (>95%) making up 7.41%, 2,500 shell gene clusters (5%~95%) representing 36.81%, and 725 cloud genes (<5%) comprising 10.68% (Fig. 3A). We also compared these gene clusters with the reference genome *Lsch* of *P. dispersa* and observed that *Lsch* contains 4,606 annotated gene models. Additionally, all *P. dispersa* strains

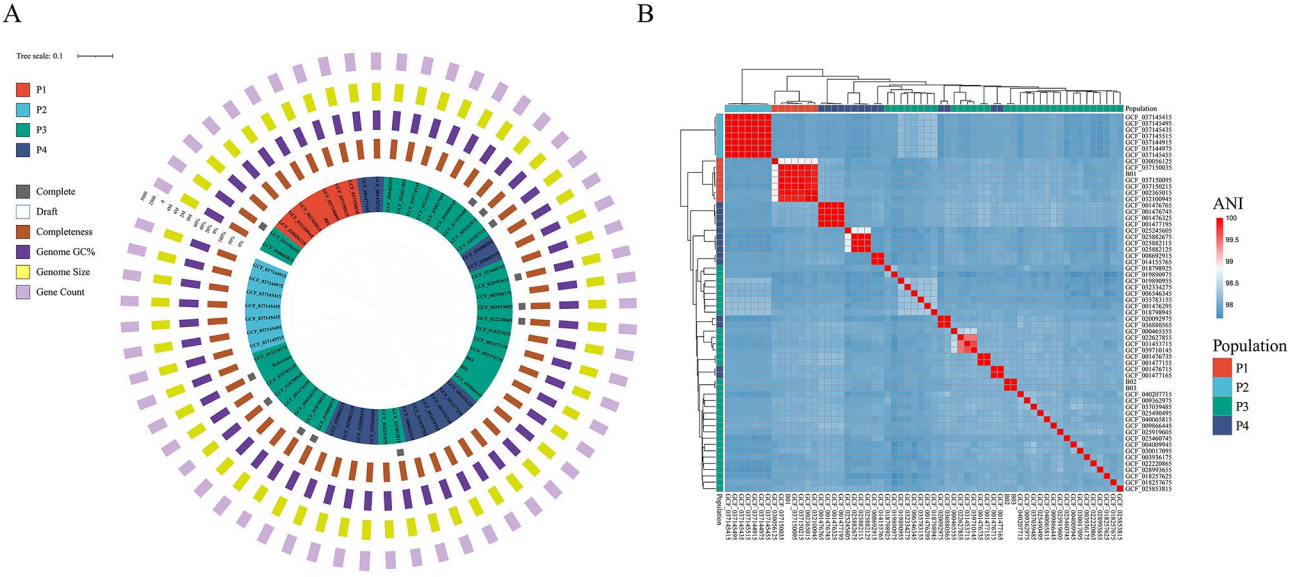

Fig. 2 Phylogenetic analysis of *P. dispersa* strains. **(A)** Whole-genome phylogenetic tree based on SNP analysis of 60 strains of *P. dispersa*. The four genetic clusters are color-coded, and outer rings from inside out represent: genome status (Draft/Complete), genome completeness, GC content, genome size, predicted number of genes. **(B)** Heatmap indicating the results of ANI analysis, with the four genetic clusters marked by color on the side of the heatmap

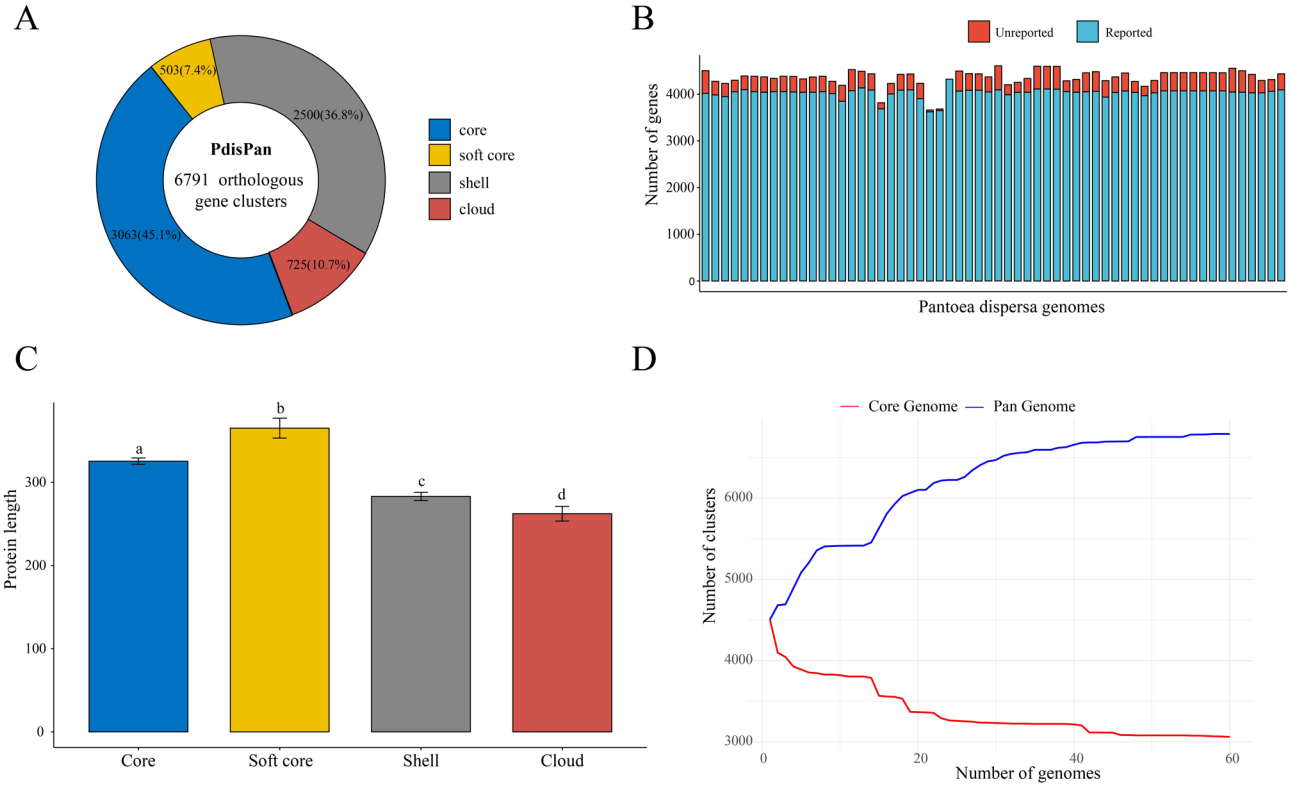

Fig. 3 An overview of the pan-genome of *P. dispersa*. **(A)** Total number of OGs in the PdisPan and the proportions of core and accessory genes. **(B)** Stacked bar chart indicating the proportion of non-reference gene clusters in PdisPan compared to the current reference genome *Lsch*. **(C)** Comparison of protein sequence lengths in the core and accessory genomes (wilcox.test, $p < 0.05$). **(D)** The changes in the number of pan-genomic and core genes in OGs as the number of genomes increases. Blue and red represent pan-genomic and core genes, respectively

harbor genes not reported in the reference genome *Lsch*, albeit to varying degrees (Fig. 3B). This indicates that the PdisPan encompasses a greater diversity of genomic fragments that are absent in the current single reference genome.

Furthermore, we observed significant differences in the lengths of protein sequences between the core and accessory genomes. The average protein length in the core genome was 325 amino acids, ranging from 29 to 4,343 amino acids. For the soft-core genome, the average protein length was 365 amino acids. Proteins in the shell genome averaged 283 amino acids, and those in the cloud genes averaged 262 amino acids (Fig. 3C). Thus, conserved proteins tend to be longer. Overall, the average protein length decreases from the soft-core to the core, and then to the shell genes and cloud genes. Additionally, the open nature of this pan-genome map indicates that an increase in the number of *P. dispersa* genomes could lead to the expansion of the pan-genome or core genes (Fig. 3D). This suggests that the current number of *P. dispersa* that has been studied is still insufficient.

Virulence factor distribution and functional annotation in PdisPan

Understanding VFs can provide deep insights into the detailed mechanisms by which pathogens invade hosts, evade the host immune system, and cause disease. Therefore, we conducted further analysis of the VFs in the PdisPan, and observed that 782 OGs were collectively annotated to 406 VFs. The distribution of these VFs was uneven (Fig. 4A), and VFs exhibited different distribution characteristics across the genetic clusters, with genetic cluster P3 having the richest VF content (404 VFs), over 80% of which are shared among all genetic clusters, although 10% of the VFs were observed only in P3 and P4 (Fig. 4B). Apart from P1 and P2, there were significant differences in the number of VFs across different genetic clusters (wilcox.test, $p < 0.05$) (Fig. 4C). Through annotation, we determined that a large number of VFs in the PdisPan's OGs were related to nutritional/metabolic factors, motility, and immune modulation (Fig. 4D).

The largest gene cluster, OG0000001 (163), was annotated as *fhaB* and was almost universally present in all isolates (58/60), aiding bacterial adhesion to respiratory epithelial cells. The second highest was OG0000003 (142), annotated as *fliC* related to flagellar motility, similarly present in almost all isolates (59/60). Numerous

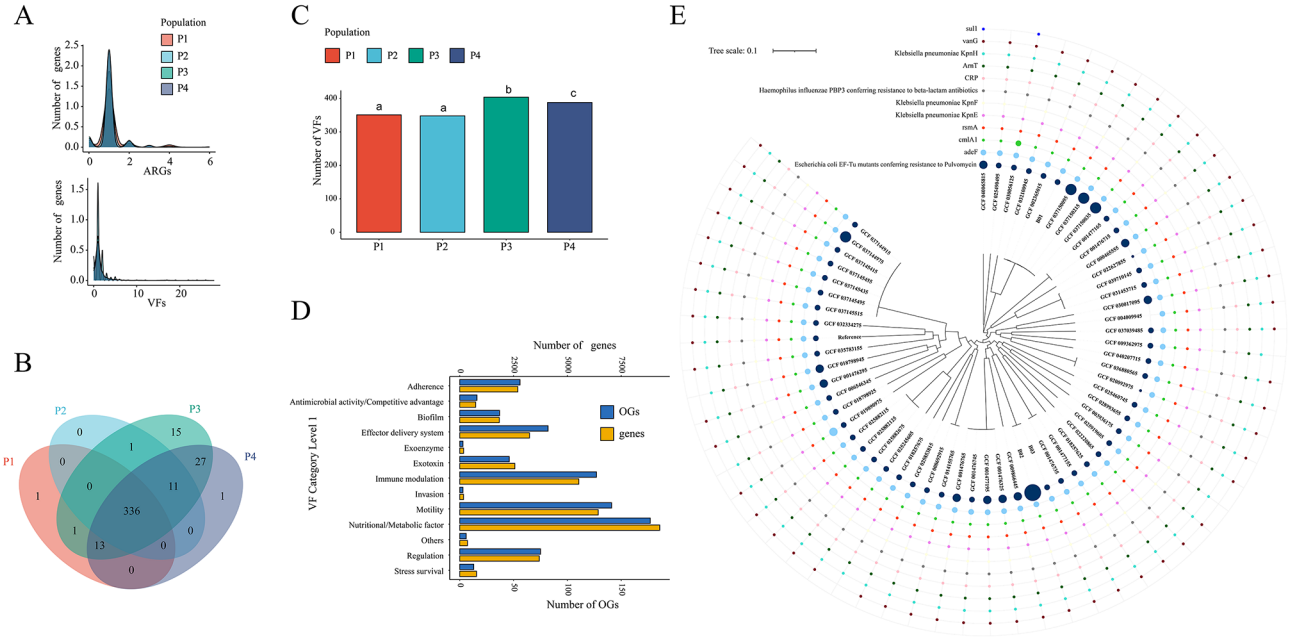

Fig. 4 Summary of CARD and VFDB annotations in the PdisPan. **(A)** Density curve chart indicating the distribution of genes annotated as VFs and ARGs in the four genetic clusters. **(B)** Venn diagram illustrating the number of unique and shared VFs among the four genetic clusters. **(C)** The number of VFs in the four genetic clusters. **(D)** The distribution of VFs at Level 1 in the PdisPan, with blue representing the number of OGs and yellow indicating the number of genes. **(E)** Whole-genome phylogenetic tree based on SNP analysis of 60 strains of *P. dispersa*. The outer layer displaying the distribution of 12 classes of ARGs in *P. dispersa*, where the presence or absence of circles represents whether ARGs exist in the strain genomes, and the size of the circles represents the number of ARG orthologous genes in the strains

genes were annotated as *hitC* (1,551), *fbpC* (1,405), *flmH* (1,053) and *ptxR* (1,013) that are associated with iron transport and flagellar motility (Additional file 9). Regarding effector delivery systems, we identified a total of 4 VF genes related to T3SS, including *etgA*, *bscC*, *bcrD* and *spa47*. While *etgA* was not widely distributed, appearing in only 38 isolates, the other three VF genes were nearly ubiquitous. We also identified the gene *fimZ* that was related to T3SS and may play a role in uptake independent of T3SS1 and was detected in most isolates (45/60). Additionally, among the three strains of *P. dispersa* isolated in this study, no new VFs were observed in B01, while eight VFs were present exclusively in the B02 and B03 isolates. These VFs were categorized under immune modulation, primarily functioning to prevent phagocytosis and protect bacteria from the host's innate immune response.

Strains of *P. dispersa* carrying ARGs

Antimicrobial resistance has been considered one of the primary threats to global health. In our subsequent analysis, we explored the distribution of ARGs within the PdisPan and identified the presence of 12 ARGs from 13 OGs (Additional file 10). Among them, two ARGs (*vanG*, Escherichia coli EF-Tu mutants conferring resistance to Pulvomycin) were present in most isolates (present in 55 and 59 isolates, respectively), the ARG *sul1* was observed in only two out of 60 strains. The remaining nine ARGs are prevalent in all strains, indicating that these genes are highly conserved across different strains of *P. dispersa* (Fig. 4E). Moreover, the abilities of *P. dispersa* to resist antibiotics primarily include antibiotic efflux and antibiotic target alteration. These mechanisms confer resistance to a variety of antibiotics, including fluoroquinolones, cephalosporins, macrolides, peptides, tetracyclines and aminoglycosides. The antibiotic target replacement mechanism that confers resistance to sulfonamide antibiotics, was observed only in the strain GCF_040065815 and GCF_030056125. Additionally, in the newly isolated strain B02, four genes belong to ARO 3,003,369, all of which are part of the orthologous gene cluster OG0000004.

The BGCs and ecological functions in *P. dispersa*

To explore the potential for secondary metabolite synthesis in the strains and to gain a deeper understanding of their biosynthetic pathways, potential biological activities, and ecological functions, we used antiSMASH to predict BGCs in the 60 strains of *P. dispersa* (Additional file 11). A total of 289 BGCs were identified, of which 23 BGCs could be aligned with known BGCs, and 4 BGCs exhibited complete identity (100% similarity) to known BGCs. These BGCs are primarily involved in the production of carotenoids, icosalide A/icosalide B,

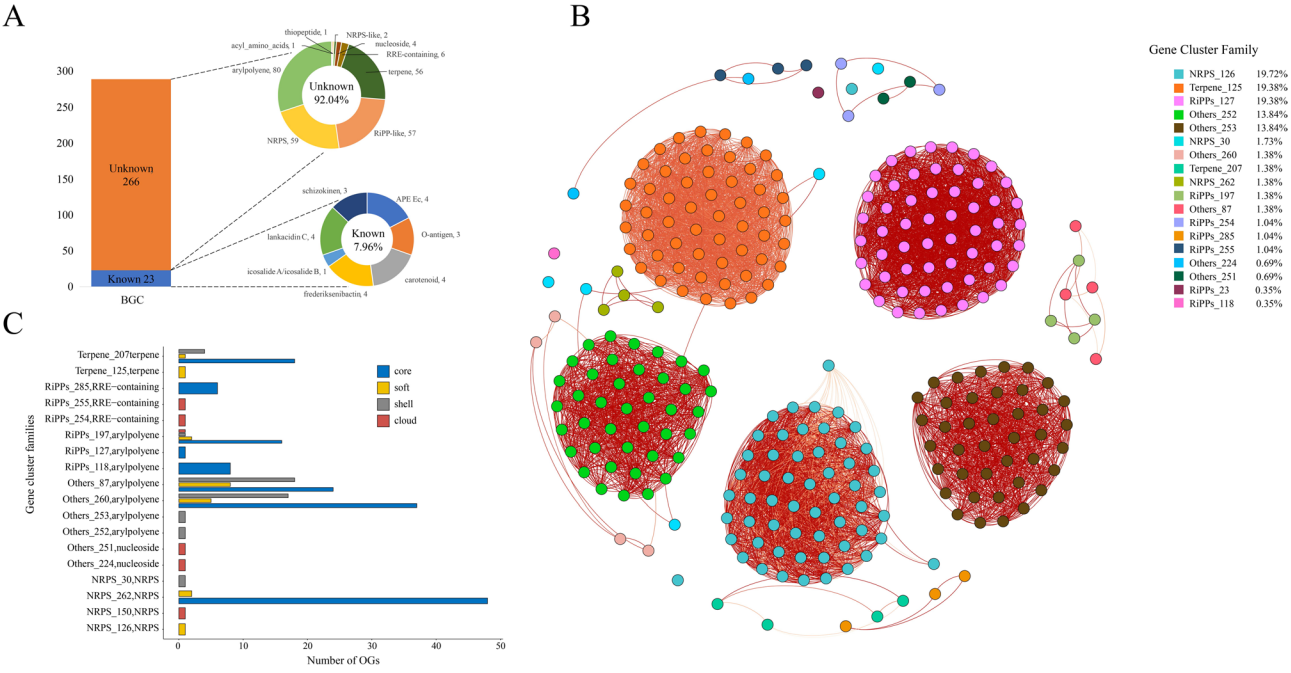

Fig. 5 Annotation of BGCs and clustering of GCFs in *P. dispersa* strains. **(A)** Overview of the antiSMASH annotation results for 60 strains of *P. dispersa*, including 23 annotations similar to known BGCs. **(B)** Network graph of secondary metabolite gene cluster family analysis of 60 strains of *P. dispersa*. Each node represents a secondary metabolite gene cluster, and the 18 GCFs are marked by the color of the nodes. The connections between nodes represent the distance between two BGCs, with deeper colors indicating closer distances. **(C)** The number of OGs covered by 18 GCFs, including core genes, soft core genes, shell genes, and cloud genes

frederiksenibactin, schizokinen, O-antigen, lankacidin C, and APE Ec. Among the remaining 266 BGCs, 30% were predicted to be arylpolyenes. BGCs predicted to be non-ribosomal peptide synthetase (NRPS), terpene, and ribosomally synthesized and post-translationally modified peptides (RiPP-like) types each accounted for approximately 20%. Less than 6% of the BGCs were predicted to contain RRE, nucleosides, NRPS-like, thiopeptide, and acyl_amino_acids types (Fig. 5A).

Cluster analysis of all BGCs resulted in a total of 18 GCFs. Among these, 86.16% of the BGCs were classified into five major GCFs, including NRPS_126, Terpene_125, RiPPs_127, Others_252, and Others_253 (Fig. 5B). This finding indicates that strains of *P. dispersa* are commonly involved in the synthesis of these five types of secondary metabolites. Notably, the network connections within RiPPs_127 are the densest, indicating a high level of homology among the gene clusters associated with RiPPs_127 in *P. dispersa*. The high degree of homology may reflect the evolutionary conservation and functional importance of these gene clusters. Additionally, the genes involved in GCFs were classified into 227 OGs within the PdisPan, including 158 core genes, 20 soft core genes, 43 shell genes, and 6 cloud genes (Fig. 5C). This result further confirms the high degree of conservation of most BGCs in *P. dispersa*.

Arylpolyenes are aromatic compounds with multiple conjugated double bonds, often possessing antibacterial and other bioactivities [42]. In the constructed pan-genome, 61 OGs were annotated as arylpolyenes, with most being core genes and soft-core genes (42/61). NRP-metallophores are peptides produced by NRPS with metal-binding capabilities. These compounds typically act as siderophores in nature, helping microorganisms acquire iron from the environment [43]. We identified 50 OGs associated with NRP-metallophores, with 48 being core genes and 2 being soft core genes. Acyl_amino_acids are a class of important endogenous signaling molecules that connect an amino acid to the acyl part of a long-chain fatty acid via an amide bond [44]. These molecules exert various physiological functions in organisms such as signaling. In the PdisPan, 50 OGs were associated with acyl_amino_acids, with 24 being core genes, 8 being soft core genes and 18 being shell genes, indicating that this gene cluster is relatively evenly distributed in the core and accessory genomes of *P. dispersa*.

Discussion

Pantoea, with an optimal growth temperature of 25 °C, is widely present on plants, fruits, vegetables, soil and in animal or human feces [45]. Numerous studies have revealed that bacteria in this genus possess dual characteristics. For example, *P. dispersa* can enhance soil

nutrient status by solubilizing phosphate [46] and can also cause sepsis in newborns [47]. According to Yang et al. [48], *P. dispersa* possesses a large number of genes involved in diverse metabolic pathways, such as those for carbohydrates, lipids, and nucleic acids. Additionally, they found that most strains carry gene sequences associated with vancomycin resistance. In the present study, we identified multiple ARGs within the genome of *P. dispersa*, indicating a possible enhancement in this bacterium's antibiotic resistance profile. To substantiate this finding and explore its implications, further experiments are warranted to rigorously assess and confirm the degree of this increased resistance. Given that current studies examining *Pantoea* are primarily limited to individual strains, analyzing their genetic evolution and functional diversity from a pan-genomic perspective could provide deeper insights into their potential applications in agriculture, industry, and healthcare. Therefore, using three isolated strains and 57 publicly available sequences, we have initially constructed a pan-genome of *P. dispersa*. The genome sizes of our newly isolated *P. dispersa* strains exhibit no significant difference compared to those of 57 strains. This suggests that the genome size of *P. dispersa* is relatively stable, however, we detected a large number of SNP sites using *Lsch* as the reference genome, that reflect the rich genetic variation within *P. dispersa*. Additionally, PdisPan contains 6,791 OGs, with 3,063 core genes and 3,728 accessory genes. Accessory genes are the variable part of the pan-genome, and the higher their proportion, the greater is the diversity among individuals [49]. Additionally, accessory genes include genes coding for supplementary or modified biochemical functions that may be useful in environments beyond basic survival such as adapting to new environments and antibiotic resistance [50]. The proportion of accessory genes in PdisPan is as high as 54.9%, indicating rapid genomic evolution as a key factor in the widespread distribution of *P. dispersa* in nature.

The pan-genome can be classified into two types: closed and open. A closed pan-genome is characterized by the stabilization of core and accessory gene numbers following the inclusion of an optimal number of genomes. Once this point is reached, the introduction of additional genomes does not result in substantial alterations to the counts of core and accessory genes [51, 52]. Conversely, if the addition of individual genomes leads to a significant increase in the number of core and accessory genes, it is referred to as an open pan-genome [53]. Based on the existing 60 strains of *P. dispersa*, the PdisPan has not yet reached closure. This is similar to the pan-genome constructed using 81 strains of *P. ananatis* that also belongs to an open pan-genome framework [54]. This result suggests that our understanding of the *Pantoea*, including *P. dispersa* and *P. ananatis*, is still insufficient. To better elucidate the human pathogenic factors of *Pantoea* and fully leverage its application potential in agriculture, industry and medicine, more strains of *Pantoea* need to be studied. Moreover, an important goal of constructing PdisPan was to evaluate the potential pathogenicity of *P. dispersa*. In PdisPan, we identified 406 VFs, and the distribution of these VFs among all isolates was uneven. Due to the limited existing research on *P. dispersa*, no VFs have been specifically identified for this species to date. While annotations from the VFDB may not comprehensively or accurately reflect the pathogenic potential of *P. dispersa*, they offer initial insights that are crucial for forming hypotheses about its virulence mechanisms. These annotations serve as a foundation for guiding further experimental validation and deeper investigation into the pathogenic potential of *P. dispersa*. The *fhaB* gene encodes Filamentous Hemagglutinin (FHA), a large surface protein with multiple functions, including promoting bacterial adherence to respiratory epithelial cells, inhibiting phagocytosis, and interacting with host cell signaling [55]. The OGs corresponding to the *fhaB* gene were present in almost all isolates (58/60), indicating that *P. dispersa* may possess a general potential to infect the upper respiratory tract. In severe cases, *P. dispersa* infection can lead to sepsis that is the most common cause of death [56]. Although the pathogenic factor of *P. dispersa* in plants is generally the T3SS, the pathogenic factor for humans remains unknown. In PdisPan, we identified four VFs related to the T3SS, including *etgA*, *bscC*, *bcrD* and *spa47*, with the latter three VFs present in all isolates. In *Bordetella bronchiseptica*, *bscC* and *bcrD* help to adjust the host immune response through the T3SS, facilitating long-term persistence and persistent infection within the trachea [57]. According to a case of neonatal sepsis caused by *Pantoea* reported by Hani et al., the initial infection occurred in the respiratory tract, suggesting a possible pathogenic route for serious infections caused by *Pantoea* [58]. Additionally, the *flmH* gene was identified in all isolates, and *flmH* is crucial for bacterial motility, adherence to surfaces, and invasion of host cells, processes that may also influence the host immune response by inducing pro-inflammatory reactions. The ARG analysis identified two ARGs (*vanG* and *EF-Tu mutants*) that are widely distributed among *P. dispersa*, potentially indicating widespread resistance to elfamycin and glycopeptide antibiotics. The *sul1* gene confers resistance to sulfonamide antibiotics by inhibiting the folic acid synthesis pathway, which is essential for bacterial growth. Sulfonamide antibiotics are widely used to treat various bacterial infections, including urinary tract infections, respiratory tract infections, and gastrointestinal infections. In this study, we identified the *sul1* gene within the pan-genome of *P. dispersa*, suggesting that this strain may exhibit resistance to sulfonamide antibiotics.

This finding is crucial for understanding its survival strategies in the environment and the potential challenges in clinical treatment. Notably, only two out of 60 strains analyzed harbored the *sul1* gene. This low frequency suggests that the possibility of horizontal gene transfer events and underscores the need for further research to fully understand the antibiotic resistance characteristics of *P. dispersa*. Although our current analysis sheds light on the antibiotic resistance landscape of *P. dispersa*, confirming these resistances necessitates further experimental studies.

P. dispersa can also serve as an important repository of secondary metabolites. Our research reveals that the PdisPan contains a total of 289 secondary metabolite gene clusters that can be classified into 18 GCFs. Notably, among these gene clusters, a large number of core genes and soft-core genes are related to arylpolyene and NRP-metallophore. Studies have indicated that arylpolyenes can inhibit or kill microorganisms by disrupting cell membrane stability, demonstrating broad-spectrum antimicrobial activity [59]. NRP-metallophores play a crucial role in microbial competition by efficiently capturing iron ions from the environment, providing a competitive advantage to the microbes producing these compounds. Additionally, certain NRP-metallophores also possess antibacterial activity, inhibiting the growth of competitors [43]. Therefore, the rich repertoire of secondary metabolites possessed by *P. dispersa* offers promising prospects for applications in industry, agriculture, and medicine. In summary, we constructed the pan-genome framework of *P. dispersa* for the first time by integrating culturomics and genomics methods, and we provide a comprehensive analysis of its genetic diversity, phylogenetic relationships, pan-genome characteristics, VFs, ARGs and the distribution of secondary metabolites. This research not only deepens the understanding of the biological characteristics of *P. dispersa* but also offers valuable insights for the treatment of *P. dispersa* infections and the subsequent development and application.

Conclusion

In conclusion, this study conducted an in-depth exploration of the pan-genome of *P. dispersa* by *de novo* assembly of three newly isolated strains and combining them with the genomic data of 57 existing strains. The results provide a detailed revelation of the rich information regarding *P. dispersa* in terms of genetic diversity, phylogenetic relationships, pan-genome characteristics, VFs, ARGs and secondary metabolites. This study lays a solid foundation and provides valuable resources for future investigating *P. dispersa*.

Abbreviations

P. dispersa	*Pantoea dispersa*
NCBI	National Center for Biotechnology Information
OGs	Orthologous gene clusters
T3SS	Type III secretion system
VFs	Virulence factors
ARGs	Antibiotic resistance genes
fAFLP	Fluorescent amplified fragment length polymorphism
IAA	Indole-3-acetic acid
qRT-PCR	Real-time quantitative PCR
SNVs	Single nucleotide variants
ANI	Average nucleotide identity
BIC	Bayesian information criterion
BGC	Biosynthetic gene cluster
GCFs	Gene cluster families
PdisPan	pan-genome of P. dispersa
NRPS	Non-ribosomal peptide synthases
RiPP-like	Ribosomally synthesized and post-translationally modified peptides
FHA	Filamentous hemagglutinin

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11625-7.

Supplementary Material 1: Additional file 7: File S1. Bash script for extracting the longest sequence from each OG as representative. This material provides a bash script designed to process the Orthogroup_Sequences folder output by OrthoFinder. The script extracts the longest sequence from each OG to serve as the representative sequence for that group

Supplementary Material 2: Additional file 1: Figure S1. A workflow for isolating cotton swab strains. Additional file 5: Figure S2. Phylogenetic tree of 60 ***P. dispersa*** strains constructed using RAxML based on aligned SNP sequences with bootstrap support. Additional file 6: Figure S3. Species tree of 60 ***P. dispersa*** strains constructed using OrthoFinder based on homologous genes

Supplementary Material 3: Additional file 2: Table S1. Statistical analysis of sequencing dataAdditional file 3: Table S2. The species annotation results based on GTDB-Tk. Additional file 4: Table S3. Summary of the complete genome information for 57 strains of ***P. dispersa*** downloaded from NCBI. Additional file 8: Table S4. The overall description of the complete genomes of 60 strains of ***P.dispersa***. Additional file 9: Table S5. Results of annotating the PdisPan to the VFDB core database. Additional file 10: Table S6. Results of annotating antibiotic resistance genes in the PdisPan using the RGI tool. Additional file 11: Table S7. Results of annotating secondary metabolic product gene clusters in 60 strains of ***P. dispersa*** using antiSMASH

Acknowledgements

We are grateful to Dr. Caiqing Zhu from the Jiangxi Medical Device Testing Center for the support provided by the "Ganjiang Hai Zhi" Talent Program for this research. We also thank the colleagues at the Jiangxi Medical Device Testing Center for their efforts in sample collection.

Author contributions

S.H.: performed the experiments, analyzed the data, wrote and revised the manuscript; Q.D., W.W., Y.Z., Y.K., Y.M. and S.Z.: performed the experiments; J.W.: conceived and designed the experiments, and revised the manuscript. All authors read and approved the final manuscript.

Funding

This project was supported by Jiangxi Provincial Department of Education (GJJ2200401).

Data availability

The whole-genome sequences three strains of P. dispersa were submitted to Genome Sequence Archive (GSA) with accession numbers PRJCA032788.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 22 November 2024 / Accepted: 22 April 2025
Published online: 02 May 2025

References

1. Challoner DR, Vodra WW. Medical devices and health–creating a new regulatory framework for moderate-risk devices. N Engl J Med. 2011;365(11):977–9.
2. Lestari T, Ryll S, Kramer A. Microbial contamination of manually reprocessed, ready to use ECG lead wire in intensive care units. Gms Hygiene Infect Control. 2013;8(1).
3. Qin S, Li R, Fang YL. Evaluating aseptic presentation of different medical device packaging configurations. Biomedical Instrum Technol /. 2023;57(3):87–97.
4. Kulkarni GB, Nayak AS, Sajjan SS, Oblesha, Karegoudar TB. Indole-3-acetic acid biosynthetic pathway and aromatic amino acid aminotransferase activities in Pantoea dispersa strain GPK. Lett Appl Microbiol. 2013;56(5):340–7.
5. Yu-Qian S, Yan Z. Growth and lipid accumulation promotion of Chlorella by endophytic *Pantoea* Sp. from rice seeds. Sci Agric Sin. 2016;49(8):1429–42.
6. Jiang L, Jeong JC, Lee JS, Park JM, Lee J. Potential of Pantoea dispersa as an effective biocontrol agent for black rot in sweet potato. Sci Rep. 2019;9(1).
7. Brady C, Cleenwerck I, Venter S, Vancanneyt M, Swings J, Coutinho T. Phylogeny and identification of Pantoea species associated with plants, humans and the natural environment based on multilocus sequence analysis (MLSA). Syst Appl Microbiol. 2008;31(6–8):447–60.
8. Habsah H, Zeehaida M, Rostenberghe HV, Noraida R, Pauzi WIW, Fatimah I, Rosliza AR, Sharimah NYN, Maimunah H. An outbreak of Pantoea spp. In a neonatal Intensive care unit secondary to contaminated parenteral nutrition. J Hosp Infect. 2005;61(3):213–8.
9. Singh BR, Agri H, Balusamy D, Jayakumar V. Bacterial profile and comparative antimicrobial efficacy of fresh urine of cows, buffaloes and humans. 传染病研究(英文); 2022.
10. Kirzinger MWB, Nadarasah G, Stavrinides J. Insights into Cross-Kingdom plant pathogenic Bacteria. Genes. 2011;2(4):980–97.
11. Büyükcam A, Tuncer Ö, Gür D, Sancak B, Ceyhan M, Cengiz AB, Kara A. Clinical and Microbiological characteristics of Pantoea agglomerans infection in children. J Infect Public Health. 2018;11(3):304–9.
12. Uche A. Pantoea agglomerans bacteremia in a 65-year-old man with acute myeloid leukemia: case report and review. South Med J. 2008;101(1):102–3.
13. Cheng A, Liu CY, Tsai HY, Hsu MS, Yang CJ, Huang YT, Liao CH, Hsueh PR. Bacteremia caused by Pantoea agglomerans at a medical center in Taiwan, 2000–2010. J Microbiol Immunol Infect. 2013;46(3):187–94.
14. Andion M, Lassaletta, Alvaro, Gonzalez, Maria J, Fernandez-Munoz. Hermogenes, Madero: Pantoea agglomerans Bacteremia in a Child With Acute Lymphoblastic Leukemia During Induction Therapy. Journal of Pediatric Hematology/oncology Official Journal of the American Society of Pediatric Hematology/oncology 2015.
15. Rezzonico F, Smits TH, Montesinos E, Frey JE, Duffy B. Genotypic comparison of *Pantoea agglomerans* plant and clinical strains. BMC Microbiol. 2009;9:204.
16. Wani AK, Akhtar N, Naqash N, Chopra C, Singh R, Kumar V, Kumar S, Mulla SI, Américo-Pinheiro JHP. Bioprospecting culturable and unculturable microbial consortia through metagenomics for bioremediation. Clean Chem Eng. 2022;2:100017.
17. Torres S, Campos V, León C, Rodríguez-Llamazares R, González M. Biosynthesis of selenium nanoparticles by *Pantoea agglomerans* and their antioxidant activity. J Nanopart Res. 2012;14:1263.
18. Albermann C. High versus low level expression of the lycopene biosynthesis genes from Pantoea ananatis in Escherichia coli. Biotechnol Lett. 2011;33(2):313–9.
19. Xu Y, Zhao J, Huang H, Guo X, Li X, Zou W, Li W, Zhang C, Huang M. Biodegradation of phthalate esters by Pantoea dispersa BJQ0007 isolated from Baijiu. J Food Compos Anal. 2022;105:104201.
20. Mayanglambam T, Sharma P, Singh DK, L A, Mishra JM, Rawat A. R: Biodegradation of quinalphos by gram negative bacteria *Pantoea agglomerans* and Acinetobacter Sp. dcm5A. Environ Conserv J. 2023;24(2):373–9.
21. Martim DB, Barbosa-Tessmann IP. Correction to: two novel acetylesterases from *Pantoea dispersa*: recombinant expression, purification, and characterization. Appl Biochem Biotechnol. 2019;189(4):1338–40.
22. Birch LQ. Characterization of the highly efficient sucrose isomerase from Pantoea dispersa UQ68J and cloning of the sucrose isomerase gene. Appl Environ Microbiol. 2005;71(3):1581–90.
23. Xu S, Liu Y-X, Cernava T, Wang H, Zhou Y, Xia T, Cao S, Berg G, Shen X-X, Wen Z, et al. Fusarium fruiting body Microbiome member Pantoea agglomerans inhibits fungal pathogenesis by targeting lipid rafts. Nat Microbiol. 2022;7(6):831–43.
24. Guo H, He S, Wang X, Zhang J, Zhang X. Phylogenetic diversity and plant growth-promoting characteristics of endophytic < italic > pantoea spp. In rice seeds. Acta Microbiol Sinica. 2019;59(12):2285–95.
25. Hebishima T, Matsumoto Y, Watanabe G, Soma GI, Kohchi C, Taya K, Hayashi Y, Hirota Y. Oral administration of immunopotentiator from Pantoea agglomerans 1 (IP-PA1) improves the survival of B16 Melanoma-Inoculated model mice. Exp Anim. 2011;60(2):101–9.
26. Agarwal G, Gitaitis RD, Dutta B. Pan-genome of novel Pantoea stewartii subsp. Indologenes reveal genes involved in onion pathogenicity and evidence of lateral gene transfer. Microorganisms. 2021;9(8):1761.
27. Zheng L, Wang S, Zhang XGGLWP. Pantoea jilinensis D25 enhances tomato salt tolerance via altering antioxidant responses and soil microbial community structure. Environ Res Sect A. 2024;243(Feb):117846117841–117846117816.
28. Bolger AM, Marc L, Bjoern U. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
29. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
30. Alexey G, Vladislav S, Nikolay V, Glenn T. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5.
31. Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 1998;26(4):1107–15.
32. Vagheesh N, Petr D, Aylwyn S, Yali X, Chris TS, Richard D. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics. 2016;32(11):1749–51.
33. Tabangin ME, Woo JG, Martin LJ. The effect of minor allele frequency on the likelihood of obtaining false positives. In: BMC proceedings:2009 Springer 2009:1–4.
34. Cagirici HB, Akpinar BA, Sen TZ, Budak H. Multiple variant calling pipelines in wheat whole exome sequencing. Int J Mol Sci. 2021;22(19):10400.
35. Stamatakis A. RAxML-VI-HPC: maximum Likelihood-Based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22(21):2688–90.
36. Hernández-Salmerón E, Moreno-Hagelsieb J. FastANI, Mash and dashing equally differentiate between. PeerJ. 2022;10:e13784.
37. Pritchard JK, Stephens MJ, Donnelly PJ. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–59.
38. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16(157):157.
39. Lihong C, Jian Y, Jun Y, Zhijian Y, Lilian S, Yan S, Qi J. VFDB: a reference database for bacterial virulence factors. Nuclc Acids Res 2005;33(Database issue):D325–8.
40. Medema MH, Kai B, Peter C, Victor DJ, Piotr Z, Fischbach MA, Tilmann W, Eriko T, Rainer B. AntiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011;39(Issue suppl_2):W339–46.
41. Navarro-Muoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, Santos ELCDL, Yeong M, Cruz-Morales P, Abubucker S. A computational framework to explore large-scale biosynthetic diversity. Nature chemical biology 2020;(1).
42. Schner TA, Gassel S, Osawa A, Tobias NJ, Okuno Y, Sakakibara Y, Shindo K, Sandmann G, Bode HB. Aryl polyenes, a highly abundant class of bacterial

natural products, are functionally related to antioxidative carotenoids. Chembiochem. 2016;17(3):247–53.
43. Reitz ZL, Medema MH. Genome mining strategies for metallophore discovery. Curr Opin Biotechnol. 2022;77:102757.
44. Prakash SA, Kamlekar RK. Function and therapeutic potential of N-acyl amino acids. Chem Phys Lipids. 2021;239:105114.
45. Andersson AM, Weiss N, Rainey F, Salkinoja-Salonen MS. Dust-borne bacteria in animal sheds, schools and children's day care centres. J Appl Microbiol. 2010;86(4):622–34.
46. Chen Y, Fan JB, Du L, Xu H, Zhang QH, He YQ. The application of phosphate solubilizing endophyte Pantoea dispersa triggers the microbial community in red acidic soil - ScienceDirect. Appl Soil Ecol. 2014;84:235–44.
47. Mehar V, Yadav D, Sanghvi J, Gupta N, Singh K. Pantoea dispersa: an unusual cause of neonatal sepsis. Brazilian J Infect Dis. 2013;17(6):726–8.
48. Yang WT, Yi Y, Xia B. Unveiling the duality of *Pantoea dispersa*: a mini review. Sci Total Environ. 2023;873:162320.
49. Whelan FJ, Hall RJ, McInerney JO. Evidence for selection in the abundant accessory gene content of a prokaryote pangenome. Mol Biol Evol. 2021;38(9):3697–708.
50. Kung VL, Ozer EA, Hauser AR. The accessory genome of < i > pseudomonas aeruginosa. Microbiol Mol Biol Rev. 2010;74(4):621–41.
51. Zhou J, Ren H, Hu M, Zhou J, Yue J. Characterization of Burkholderia cepacia complex core genome and the underlying recombination and positive selection. Front Genet. 2020;11:506.
52. Nathamuni S, Jangam AK, Katneni VK, Selvaraj A, Koyadan VK. Insights on genomic diversity of *Vibrio* spp. through Pan-genome analysis. Ann Microbiol. 2019;69:1547–55.
53. Bosi E, Fani R, Fondi M. Defining orthologs and pangenome size metrics. Methods Mol Biol. 2015;1231(1231):191–202.
54. Agarwal G, Choudhary D, Stice SP, Myers BK, Gitaitis RD, Venter SN, Kvitko BH, Dutta B. Pan-Genome-Wide analysis of *Pantoea ananatis* identified genes linked to pathogenicity in onion. Front Microbiol. 2021;12:684756.
55. Higgs R, Higgins SC, Ross PJ, Mills KHG. Immunity to the respiratory pathogen Bordetella pertussis. Mucosal Immunol. 2012;5(5):485.
56. Ruan XL, Qin X, Li M. Nosocomial bloodstream infection pathogen Pantoea dispersa: a case report and literature review. J Hosp Infect. 2022;127:77–82.
57. Goto M, Abe A, Hanawa T, Suzuki M, Kuwae A. Bcr4 is a chaperone for the inner rod protein in the < i > bordetella type III secretion system. Microbiol Spectr. 2022;10(5):e01443–01422.
58. Hani S, Tahiri FE, Lalaoui A, Bennaoui F, Soraa N, Slitine NEI, Maoulainine FMR. Pantoea SPP: A new nosocomial infection in the neonatal intensive care unit. Open J Pediatr. 2023;13(2):181–8.
59. Johnston I, Osborn LJ, Mcmanus EA, Kadam A, Schultz KB, Ahern PP, Brown JM, Claesen J. Identification of essential genes for Escherichia coli Aryl polyene biosynthesis and function in biofilm formation. Cold Spring Harbor Laboratory; 2020.

Publisher's note