Supplemental files

Document S0.

A list of captions for the Supplemental files.

Document S1.

A complete and unabridged description of the materials and methods used in this study.

Document S2.

A stand-alone, HTML webpage reporting the conservation and synteny of the HiVir gene cluster within the genomes of this study. These results were produced by CBLASTER [Gilchrist2021a] v 1.3.19, using as the input query the HiVir gene cluster from *P. ananatis* LMG 20103, as described in Asselin et al. (2018) [38].

Figure S1.

The correlation plot, based on ANI (average nucleotide identity) values, categorizes *P. agglomoerans (Pagg)* strains according to their phylogroups. Pairwise ANI values for 187 *Pagg* strains were calculated using FastANI (v. 1.3.3) and organized through a hierarchical clustering method employing the Pearson correlation matrix. The strains are partitioned into two groups, denoted by red and blue, aligning with phylogroup distinctions (blue corresponds to Phylogroup I, while red corresponds to Phylogroup II).

Figure S2. The phylogeny of *Pantoea agglomerans*, based on the *gyrB* gene is largely congruent with the core genome phylogeny.

(**A**) An approximately-maximum-likelihood phylogenetic tree, constructed using FastTree v.2.1.11, relies on the 3111 core genes shared among 181 *P. agglomerans* strains (Figure 1). Maximum-likelihood phylogenetic trees, constructed based on partial *gyrB* gene sequences targeted by primers designed by (**B**) [45] and (**C**) [46], indicate that the Bonasera *gyrB* region

reveals a phylogeny congruent with that of the core genome. Strains not assigned to phylogroups in the core genome tree (CFBP8785, T2, T3, T4, T5, and P5) form part of phylogroups in *gyrB* trees. Phylogroups are highlighted according to the phylogroup designation in the core genome tree, with a yellow box representing Phylogroup I and red representing Phylogroup II. The association of strains with onions is indicated by the purple coloring of strain names.

Figure S3.

Extended phylogroup identification of 226 *Pantoea agglomerans* strains using maximum-likelihood phylogenetic analysis of partial *gyrB* gene as targeted by primers designed by [45]. *Tatumella ptyseos* ATCC33301^{T} was used as an outgroup, and branch lengths were hidden. Phylogroups are highlighted: Phylogroup I is marked with a yellow box, and Phylogroup II with pink.

Figure S4.

A heatmap visualization of the presence (pink) and absence (yellow) genes (n= 27,204) observed across 181 *Pagg* strains. The dendrogram drawn based on the gene presence and absence categorizes *Pagg* strains into two groups whose members are consistent with those of the Phylogroup I (indicated by black line) and Phylogroup II (red line).

Figure S5.

A combined hierarchical clustering and heatmap showing the relationship among all of the replicons (chromosomes and plasmids) studied. The distances between replicons were computed using Mashtree [59]. Values shown are 100×(1-MASH), which is analogous to percentage identity.

Figure S6.

A combined hierarchical clustering and heatmap showing the relationship among the **chromosomal** replicons studied. The distances between replicons were computed using Mashtree [59]. Values shown are 100×(1-MASH), which is analogous to percentage identity.

Figure S7.

A combined hierarchical clustering and heatmap showing the relationship among the **LPP-1** replicons studied. The distances between replicons were computed using Mashtree [59]. Values shown are 100×(1-MASH), which is analogous to percentage identity.

Figure S8.

A combined hierarchical clustering and heatmap showing the relationship among the **pAggl** replicons studied. The distances between replicons were computed using Mashtree [59]. Values shown are 100×(1-MASH), which is analogous to percentage identity.

Figure S9.

A combined hierarchical clustering and heatmap showing the relationship among the **pOnion** replicons studied. The distances between replicons were computed using Mashtree [59]. Values shown are 100×(1-MASH), which is analogous to percentage identity.

Figure S10.

A combined hierarchical clustering and heatmap showing the relationship among the **pENY** replicons studied. The distances between replicons were computed using Mashtree [59]. Values shown are 100×(1-MASH), which is analogous to percentage identity.

Figure S11.

This figure shows pAggl plaster plasmids pAR5_B and pAR1aB with core pAggl genes labeled with green boxes. Core genes for pAggl were determined using Roary based on Prokka annotations. Arrows denote CDSs from RefSeq annotations. Some genes appeared to occur in

clusters where genes had similar function or seemed to act in a single pathway. Where their annotations allowed reasonable guesses of function, gene clusters are labeled with yellow bars and expected functions.

Figure S12.

An alignment of *P. agglomerans (Pagg)* AR1a's plasmids C and D against *Pagg* T88c's plasmid C. All three plasmids are members of the pOnion plaster. These results demonstrate the sequence of T88c plasmid C can be reconstructed using sequence from AR1'a plasmids C and D. T88c plasmid C CDSs with possible roles in partition and replication are filled in black, genes with annotations suggesting involvement in recombination (e.g. transposases, integrases) are filled in purple, genes annotated as *umuC* or *umuD* are filled in green, and CDSs involved in type IV secretion are filled in pink. The figure was generated using the NCBI BLAST website, the blastn tool, the "Align two or more sequences" option, optimized for somewhat similar sequences. T88c plasmid C sequence was used as the query, and AR1a plasmids C and D as subject sequences. The graphical representation of the alignment was pasted below a linear representation of the plasmid generated in SeqBuilder 17.3 (DNAStar, Madison, WI) using pgap annotations.

Table S1.

A comprehensive list of the primers utilized in this study for gene detection, sequencing, and genetic manipulations.

Table S2.

The 618 *Pantoea* strains that are used in this manuscript and the tables in which they are mentioned.

Table S3.

A list of the 100 public *Pantoea* genomes employed in this study for genome analysis, detailing strain information including source, year, and place of isolation, and their respective GenBank accession numbers.

Table S4.

A list of 501 *Pantoea* strains tested for red onion scale necrosis assay (RSN), and *P agglomerans*-specific and *Pantoea*-HiVir PCR assays.

Table S5.

QUAST quality assessment results for the 87 *Pantoea agglomerans* genomes sequenced in this study.

Table S6.

A table displaying the BUSCO genome assembly and annotation completeness scores for 81 genomes sequenced in this study.

Table S7.

Pan-replicon Analysis Results of Chromosome Replicons

A list of all of the gene groups identified by Roary [48] for replicons belonging to the "chromosome" plaster. Each row reports a single gene group. The "Gene" and "Annotation" values report the Prokka [47] predicted gene name and gene product for the genes in the group. In the case that multiple different gene names or products are predicted, all predictions are listed. The "Count" and "Pers" columns report the number and percentage of replicons in the plaster group containing members of the genome group. The remaining columns indicate the presence of the gene group in individual replicons.

Table S8.

Pan-replicon Analysis Results of LPP-1 Replicons

A list of all of the gene groups identified by Roary [48] for replicons belonging to the “LPP-1” plaster. Each row reports a single gene group. The “Gene” and “Annotation” values report the Prokka [47] predicted gene name and gene product for the genes in the group. In the case that multiple different gene names or products are predicted, all predictions are listed. The “Count” and “Pers” columns report the number and percentage of replicons in the plaster group containing members of the genome group. The remaining columns indicate the presence of the gene group in individual replicons.

Table S9.

Pan-replicon Analysis Results of pAggl Replicons

A list of all of the gene groups identified by Roary [48] for replicons belonging to the “pAggl” plaster. Each row reports a single gene group. The “Gene” and “Annotation” values report the Prokka [47] predicted gene name and gene product for the genes in the group. In the case that multiple different gene names or products are predicted, all predictions are listed. The “Count” and “Pers” columns report the number and percentage of replicons in the plaster group containing members of the genome group. The remaining columns indicate the presence of the gene group in individual replicons. The ortholog groups whose members include genes from the HiVir gene cluster are shared blue.

Table S10.

Pan-replicon Analysis Results of pOnion Replicons

A list of all of the gene groups identified by Roary [48] for replicons belonging to the “pOnion” plaster. Each row reports a single gene group. The “Gene” and “Annotation” values report the Prokka [47] predicted gene name and gene product for the genes in the group. In the case that multiple different gene names or products are predicted, all predictions are listed. The “Count”

and “Pers” columns report the number and percentage of replicons in the plaster group containing members of the genome group. The remaining columns indicate the presence of the gene group in individual replicons. The ortholog groups whose members include genes from the cop, HiVir, and alt gene clusters are shared red, blue, and green, respectively.

Table S11.

Pan-replicon Analysis Results of pENY Replicons

A list of all of the gene groups identified by Roary [48] for replicons belonging to the “pENY” plaster. Each row reports a single gene group. The “Gene” and “Annotation” values report the Prokka [47] predicted gene name and gene product for the genes in the group. In the case that multiple different gene names or products are predicted, all predictions are listed. The “Count” and “Pers” columns report the number and percentage of replicons in the plaster group containing members of the genome group. The remaining columns indicate the presence of the gene group in individual replicons.

Table S12.

A pairwise average nucleotide identity (ANI) comparison table of 187 *Pantoea agglomerans* genomes.

Table S13.

Type strains of *Pantoea* species and their respective GenBank accession numbers from which the *gyrB* gene was extracted for phylogenetic analysis.

Table S14.

Metadata information utilized for annotating the 181 *Pantoea agglomerans* strains that appear in Figure 1.

Table S15.

A pairwise nucleotide identity comparison of HiVir gene clusters among *Pantoea* species.

Table S16.

A compilation of 125 *Pantoea agglomerans* strains subjected to whole genome sequencing (WGS) and *gyrB* gene sequencing in this study. Closed, complete genomes were generated for 19 strains, draft genomes were made for 68 strains, and *gyrB* gene sequences were made for 38 strains.

Table S17.

This table lists the strains whose genomes were used in the plaster analysis.

Initially, 38 closed, complete genomes were collected into a pool of *P. agglomerans (Pagg)* genomes. The table lists, for each strain, its “phylogroup” and GenBank “accession” identifier.

In order to confirm these taxonomic identifications, we computed the ANI scores of each strain compared to the *Pagg* type strain (here denoted strain FDAARGOS1447). We found that the ANI scores of 3 strains (33.1, FL1, and HJS002), as found in the “fastani” column, fell below the typical species cutoff of 95% and were removed from the pool. We used BUSCO to evaluate the integrity of the remaining 35 assemblies; we found that that “busco_C” and “busco_D” scores ranged from 98.7%-99.5% and 0.2%-2%, respectively, indicating the quality of these assemblies ranged from good to excellent.

The “new” column records whether the genome assembly is new to this study. The “used” column records whether or not the genome was ultimately used as input for the plaster analysis.

Table S18.

A list of replicons used in the plaster analysis and their statistics. The pool of 35 genomes listed in Table S17 consisted of 35 chromosomes plus 125 plasmids for a total of 160 replicons. The chromosomes ranged in size from 3,978,822bp to 4,410,564; the plasmids ranged in size from 2,550bp to 613,013bp.

Table S19.

A list of all the plaster assignments by our plaster analysis to the 160 replicons listed in Table S18. Our analysis identified 31 distinct plasmid clusters, or "plasters", 9 of which were non-trivial (>1 member). We assigned previously published names to 3 of the non-trivial plaster plasters ("chromosome", "LPP-1", and "pPATH") and new names to 3 others ("pAggl", "pOnion", and "pENY").

Table S20.

List of genes in the pAggl core genome in the context of the *Pantoea agglomerans* AR1a pAggl plasmid. Core pAggl genes were called based on Prokka annotations using Roary. Where the coordinates for the annotations did not match exactly (e.g. Prokka and RefSeq listed the gene with different start sites), the coordinates for RefSeq are listed to be consistent with locus name, GO terms and EC numbers.

Table S21.

List of genes in the pAggl core genome in the context of the *Pantoea agglomerans* AR5 pAggl plasmid. Core pAggl genes were called based on Prokka annotations using Roary. Where the coordinates for the annotations did not match exactly (e.g. Prokka and RefSeq listed the gene with different start sites), the coordinates for RefSeq are listed to be consistent with locus name, GO terms and EC numbers.

Table S22.

A list of 123 *Pantoea agglomerans* strains was utilized in the copper tolerance assays. The presence of the *copC* gene was screened in these strains using the *copC* detection primers designed in this study. Copper tolerance was assessed based on the level of colony growth on CYE agar supplemented with varying concentrations of copper, with scores assigned as follows: 0 for no growth, 1 for non-confluent growth, and 2 for confluent growth.

Table S23.

Predicted mobilizable and conjugative elements

Mobilizable and conjugative elements were predicted by MacSyFinder v. 2.1.2 [56, 57] using the CONJScan``/Plasmids v. 2.0.1 models [58]. For each replicon, the number of predicted instances of each submodel is shown. The submodels with predicted include, MOB (mobilizable elements), T4SS_typeF and T4SS_typeG (functional conjugative T4SS's), and dCONJ_typeG and dCONJ_typeT (decayed conjugative T4SS's).

Table S24.

Details of the *Pantoea agglomerans* strains used in conjugation experiments. The replicons found in each strain are listed with some relevant features. TXXScan and CONJScan results using MacSyFinder are listed for the strains. Strain derivatives (rifampicin resistant or altered to carry a kanamycin resistance cassette) are presumed to have identical results to wild-type strains.

Table S25.

Results from preliminary conjugation experiments using *Pantoea agglomerans (Pagg)* strain CB1 harboring a version of the pCB1C plasmid with a kanamycin-resistance cassette as donor and four non-pathogenic strains of *Pagg* as potential recipients. For each mating pair, three colonies, or fewer if three were not available, were screened via PCR with primers designed to amplify the donor strain (CB1 3125-F/3125-R), the recipient strains (ImpA F1/R2), or from within the HiVir cluster (Pag 3283 F1/R1), which is present on pCB1C but not in any of the wild-type recipient strains. The results are color coded. Green means the colonies gave the expected PCR result. Red means PCR indicated rifampicin-resistant donor colonies breaking through. Yellow means the PCR result was ambiguous. Putative transconjugants were selected for genome sequencing from conjugations with Pagg CB1 pCBC-AKan colony 30 as the donor. Putative transconjugants were recovered from conjugations with AR8b Rp^r, MMD61212-C Rp^r, and FC61912-B Rp^r as recipients, but not with J22c Rp^r.

Dataset S1.

The consolidated and reconciled annotations from all of the WGS strains used in this study. In this study, Prokka was used, independently, by both the UGA and USDA authors to reannotate genomes for input to Roary [48]. For genomes available from NCBI, both GenBank and RefSeq annotations are available. Hence, for each genome, multiple sets of gene annotations are available. For each WGS strain, we have provided a table listing which locus tags are equivalent in the available annotation sets. That is, each row of each table lists the equivalent locus tags from a genome's available annotation sets. Locus tags are considered equivalent if they refer to CDS features whose amino acid sequences are equal.