Molecular Biology and Evolution, 2025, **42**, 1–16
https://doi.org/10.1093/molbev/msaf273
Advance access publication 29 October 2025
Discoveries

MBE OXFORD

Evolution of the Plant-Associated *Pantoea* was Accompanied by Plasmid Domestication Events

Devani Romero Picazo ,* Florence Muccino , Paul Kwasigroch , Lisa Hartmann , Nils F. Hülter , Tal Dagan

Institute for General Microbiology, Kiel University, Am Botanischen Garten 11, 24118 Kiel, Germany

*Corresponding author: E-mail: dpicazo@ifam.uni-kiel.de.
Associate editor: Deepa Agashe

Abstract

Plasmids are important drivers of evolutionary transformations and ecological adaptation in prokaryotes. Plasmids supplying the host with beneficial functions may become domesticated and gain a stable inheritance within the host lineage. Domesticated plasmids may comprise core genes that are present in all taxon members. The origin of plasmid core genes remains poorly understood and alternative scenarios entailing gene translocation or genetic redundancy are debated. Studying plasmid evolution in the plant-associated *Pantoea*, we show that the large *Pantoea* plasmids (LPP-1 and LPP-2) are domesticated. We infer that the LPP-1 was acquired in the ancestor of plant-associated *Pantoea* species. The LPP-2 acquisition is traced to the ancestor of plant growth–promoting species. We show that both plasmids are vertically inherited and the LPP-1 replication is furthermore coordinated with chromosome replication. Both plasmids harbor core gene families at the genus (LPP-1) or species (LPP-2) level. Using phylogenomics we infer a deep divergence between plasmid and chromosomal core genes, indicating rare gene translocation between the replicons. Our results suggest that the LPP-1 and LPP-2 acquisition introduced genetic redundancy with chromosomal genes, that was followed by successive waves of differential gene loss. The domestication of both plasmids likely contributed to species divergence in *Pantoea*.

Keywords: plasmid evolution, symbiosis, *Pantoea*, pangenomes, plasmid domestication

Significance statement

Bacterial adaptation to host-associated lifestyles can be promoted by the acquisition of plasmids, which are mobile extrachromosomal replicons. Canonic examples are the symbiosis plasmids of rhizobia, where traits essential for their interaction with legumes are plasmid-encoded. Symbiosis plasmids are likely to evolve into stably inherited domesticated plasmids. Recent studies suggest that domesticated plasmids are widely found among plant-associated bacteria. Studying the origin of plasmids in the plant-associated genus *Pantoea*, we show that the large *Pantoea* plasmids (LPP-1 and LPP-2) are domesticated. Phylogenomic analyses trace back the plasmid origin to ancestral nodes in the *Pantoea* species tree. Our reconstruction of main events in LPP-1 and LPP-2 evolution indicates a preliminary massive gene gain—some likely redundant with the chromosome—followed by differential gene loss.

Introduction

Plasmids are extrachromosomal genetic elements that reside in prokaryotic organisms, where they replicate using the host replication machinery. Plasmid persistence within the host depends on their ability to replicate and segregate into daughter cells during cell division (Wein and Dagan 2020). Alternatively, plasmids can confer their host a fitness advantage over non-hosts in the population via addiction mechanisms (Naito et al. 1995), or the expression of traits beneficial to the host (eg lactic acid metabolism in *Bacillus lactis*; McKay et al. 1976). Plasmid-host adaptation over long time scales may lead to the evolution of chromids—extrachromosomal replicons of plasmid origin—that comprise essential genes and are an integral component of the host genome (Harrison et al. 2010; diCenzo and Finan 2017). Plasmids *en route* to becoming a chromid can be considered as "domesticated" in their host lineage (Hall et al. 2022). The coordination of plasmid replication with cell division, which ensures faithful plasmid segregation, and thus vertical inheritance of the plasmid, is a crucial step during plasmid domestication. In *Vibrio cholera*e, chromid replication is well coordinated with the chromosome replication via a check-point mechanism that is encoded in the chromosome (Val et al. 2016). Plasmid domestication may be furthermore accompanied by non-functionalization and loss of plasmid conjugation genes (Ares-Arroyo et al. 2023; Hanke et al. 2024). Domesticated plasmids typically confer their host traits that are beneficial in their natural habitat. Examples are nodulation and nitrogen

Received: July 18, 2025. **Revised:** October 3, 2025. **Accepted:** October 10, 2025

fixation in rhizobia (Nuti et al. 1979; Bánfalvi et al. 1981), virulence factors and biofilm formation in enterohemorrhagic *Escherichia* (Lim et al. 2010) and chitin utilization in *V. cholerae* (Meibom et al. 2004), where plasmid acquisition likely facilitated the host habitat expansion.

A main characteristic of chromids is the presence of essential and core genes that are universally present in all taxon members (Harrison et al. 2010; diCenzo and Finan 2017). The evolutionary origin of core genes in domesticated plasmids has been debated in the literature, where two alternative hypotheses have been suggested. According to the "translocation hypothesis", core genes in chromids result from the translocation of chromosomal genes to the chromid following plasmid domestication. Alternatively, the "redundancy hypothesis" posits that the evolution of core genes in chromids entails preliminary genetic redundancy between the plasmid and chromosome that is followed by differential loss (diCenzo and Finan 2017; Hall et al. 2022). According to the redundancy hypothesis, precursors of domesticated plasmids are megaplasmids that harbor homologs to chromosomal genes; if the chromosomal and plasmid genes are functionally redundant, a loss of the chromosomal homolog is expected to have a negligible effect on the host phenotype. A gradual process of redundant chromosomal gene loss is thus expected to promote the evolution of plasmid indispensability to the host. Nonetheless, evidence in support of either hypothesis remains scarce.

Members of the genus *Pantoea* are widely associated with plants (Walterson and Stavrinides 2015). Several *Pantoea* species are recognized plant pathogens, including *P. ananatis*, that infects mono- and dicotyledonous plants (Coutinho and Venter 2009; Stice et al. 2020), the maize pathogen *P. stewartii* (Stewart 1897) and the onion pathogen *P. allii* (Brady et al. 2011). In contrast, strains among the species *P. vagans*, *P. agglomerans*, *P. eucalypti*, *P. eucrina*, and *P. dispersa* are generally recognized as commensals (Castagno et al. 2011; Kamber et al. 2012; Dutkiewicz et al. 2016; Singh et al. 2021; Lekired et al. 2023). Recent evidence suggests that commensal *P. agglomerans* is vertically inherited in wheat (Sanz-Puente et al. 2025), yet pathogenic *P. agglomerans* have been reported in other species (Barash and Manulis-Sasson 2009). Plant growth–promoting *Pantoea* harbor traits such as nitrogen fixation (Ruppel and Merbach 1997; Singh et al. 2021), phosphate solubilization (Castagno et al. 2011), production of the plant hormone indole-3-acetic acid (Sergeeva et al. 2007), heavy metal tolerance (Lekired et al. 2023), antifungal activity (Winkelmann et al. 1980; Xu et al. 2022), and antibiotic activity (eg against the causative agent of fire blight in trees; Stockwell et al. 2011). Many *Pantoea* species harbor a megaplasmid named "Large *Pantoea* Plasmid" (LPP-1) and a medium sized plasmid named LPP-2 (De Maayer et al. 2012). Homologs of plasmid LPP-2 (termed also pAggl) are widely present in *P. agglomerans*; besides their putative contribution to sucrose utilization, not much is known about traits conferred by LPP-2 (Sulja et al. 2022; Shin et al. 2025). In contrast, curing of LPP-1 has a significant effect on *Pantoea* physiology. LPP-1 curing in the maize pathogen *P. ananatis* strain DZ-12 (termed pDZ-12), leads to a decrease in biofilm formation and decreased disease symptoms in infected maize plants (Zhao et al. 2021). Similarly, curing of LPP-1 in *P. vagans* strain C9-1 led to reduced (or loss of) ability to utilize several carbohydrate sources, e.g. maltose and salicin (Smits et al. 2010), reduced swarming motility, and reduced indole production (Klein et al. 2017). The LPP-1 thus harbors multiple traits that are relevant for the *Pantoea* endophytic lifestyle. The wide distribution of plasmids LPP-1 and LPP-2 in the genus and their conservation across *Pantoea* species suggests that they have been domesticated during the genus evolutionary history (De Maayer et al. 2012).

Here we reconstruct the evolution of LPP-1 and LPP-2 based on comparative genomics of nine *Pantoea* species. Additionally, we examine plasmid replication dynamics in *P. agglomerans* R1 that we previously isolated from wheat grains (Soluch et al. 2021). Analyzing the phylogenies of plasmid genes, we find evidence in support of LPP-1 and LPP-2 domestication in *Pantoea*.

Results

The Acquisition and Diversification of Plasmids LPP-1 and LPP-2 Coincide with *Pantoea* Speciation Events

To investigate the evolution of *Pantoea* plasmids, we clustered 462 plasmids found in 253 *Pantoea* isolates into 39 groups of homologous plasmids according to their gene content (Table S1). Homologs of LPP-1 and LPP-2 plasmids correspond to 80% of the *Pantoea* plasmids and are characterized by similar gene content (Fig. 1a). LPP-1 homologs (214 plasmids) were clustered into six distinct groups, where all plasmids harbor a replication initiation protein (RepB) that is homologous to previously annotated LPP-1 plasmids (De Maayer et al. 2012). The clustering of 114 LPP-2 homologs yielded two distinct groups. All LPP-2 plasmids harbor a homolog to the previously annotated LPP-2 RepB. The remaining 134 plasmids were clustered into 31 small groups having low gene sharing (ie connectivity) with the other plasmid groups (Fig. 1a).

To reconstruct the LPP-1 and LPP-2 evolutionary history, we first inferred the *Pantoea* species phylogeny. The species tree topology comprises clades that correspond to the nine *Pantoea* species (Fig. 1b). The inferred root of the species tree splits *P. eucrina* and *P. dispersa* from the remaining species (Fig. 1b; Note S1; Fig. S1; Table S2). The distribution of LPP-1 and LPP-2 clusters across the *Pantoea* species tree suggests the co-divergence of both plasmids with the *Pantoea* species (Fig. 1b). The prevalence of LPP-1 adds support to the view that the LPP-1 origin traces back to the genus origin (De Maayer et al. 2012). Here we term the LPP-1 groups based on previous classification of LPP-1 homologs (De Maayer et al. 2012). LPP-1 plasmids in group I were clustered into three groups that are supported by the *repB* phylogeny (Fig. S2). Plasmids in LPP-1 group I.I are specific to *P. agglomerans*, group I.II plasmids were observed in *P. eucalypti*, and group I.III plasmids are shared to *P. vagans* and *P. anthophila*. Plasmids in group II were observed only among *P. ananatis*, *P. allii*, and *P. stewartii*. Our analysis uncovers two additional LPP-1 groups: group IV in *P. dispersa* and group V that is specific to *P. eucrina*. In addition to gene content differences, the LPP-1 clusters differ in their plasmid size (Fig. S3), where group II in *P. ananatis*, *P. stewartii*, and *P. allii* were smaller compared to group IV in *P. dispersa*. Isolates lacking LPP-1 and/or LPP-2 are observed among *P. eucalypti* and a small cluster of *P. aglomerans* isolates. Isolates where LPP-1 and LPP-2 plasmids are not detected have a significantly higher number of contigs compared to those isolates where the plasmids were detected (Fig. S4). Hence, the absence of evidence for LPP-1 and LPP-2 contigs in some *Pantoea* strains may be

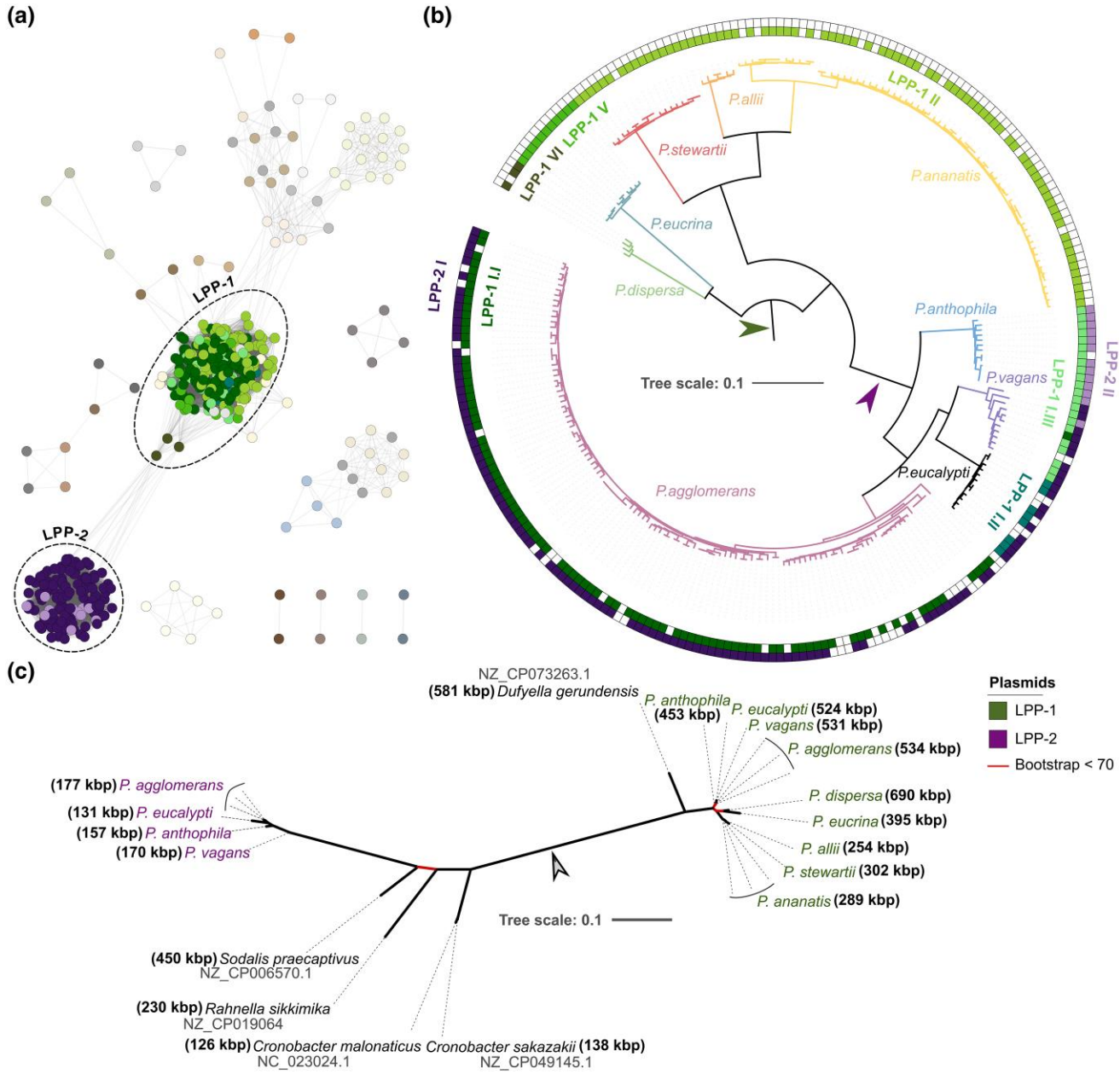
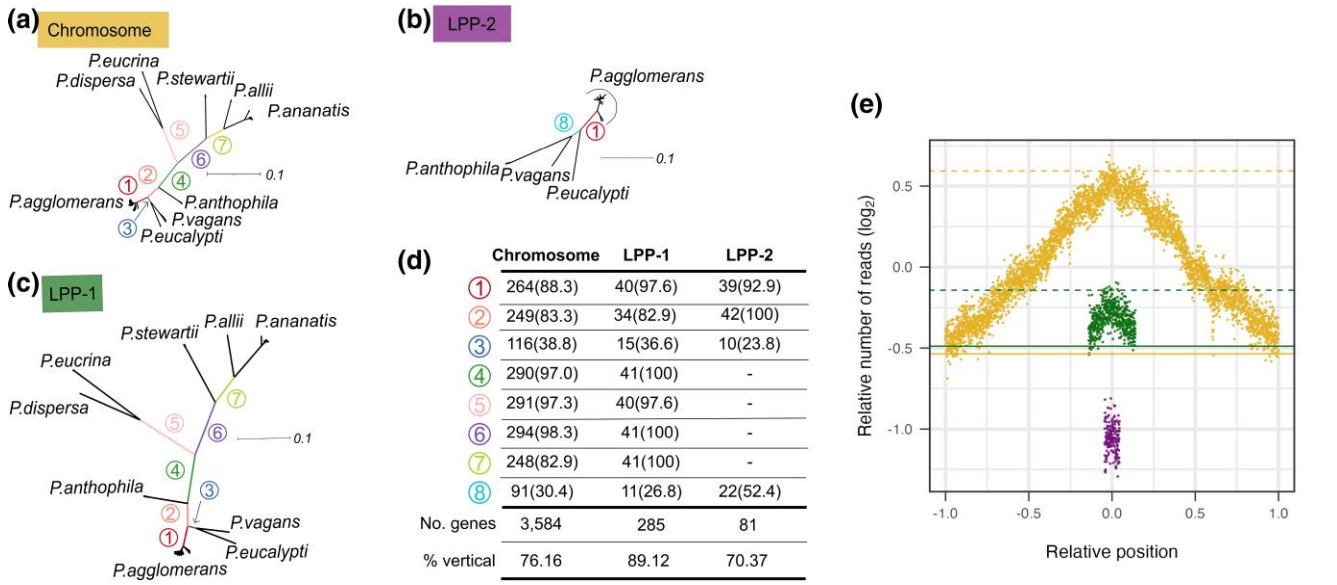


Fig. 1. The LPP-1 and LPP-2 groups correspond to *Pantoea* species taxonomy. a) A network of plasmid gene content similarity. Nodes in the network correspond to plasmids and the edges correspond to shared gene content calculated as the overlap coefficient (see Methods). Node color is according to the LPP-1 and LPP-2 groups shown in b). Edge length is proportional to plasmid similarity, such that similar plasmids appear as aggregates in the network. Edge thickness is proportional to the overlap coefficient value. Only edges having overlap coefficient ≥ 0.4 are shown. b) Phylogenetic tree of 253 *Pantoea* strains from nine species isolated mostly from plant hosts (Table S1). The tree was inferred from 299 chromosomal single-copy gene families. The root was inferred using a phylogenomic rooting approach on this dataset, and additionally, on a set of 205 chromosome-specific CSC gene families identified in a recently published extended dataset of the genus (see Note S1). The tree outer rings show the distribution of LPP-1 (in green) and LPP-2 (in purple) groups. Internal nodes inferred as plasmid acquisition events under the most parsimonious scenario are shown in green and purple, for LPP-1 and LPP-2, respectively. c) Phylogeny of *repB* homologs in LPP-1 and LPP-2 of 13 representative *Pantoea* isolates (Table S1). LPP-1 homologs are shown in green labels; LPP-2 homologs are shown with purple labels. The root of the phylogenetic tree—indicated by an arrow—was inferred using Minimal Ancestor Deviation (MAD) (Tria et al. 2017). The plasmid size is shown for all species. For *Pantoea* species, plasmid size corresponds to the median plasmid size per species. Branches highlighted in red have bootstrap values under 70% (using 1,000 replicates).

explained by a high genome fragmentation that hinders plasmid detection using computational tools. Nevertheless, we cannot rule out genuine plasmid loss in some of these isolates. Notably, the rhizobia symbiosis plasmid that is considered essential may be lost in the soil environment (Soberón-Chávez and Nájera 1989). Indeed, several of the *P. eucalypti* strains were isolated from soil samples, hence the absence of LPP-1 and/or LPP-2 in those strains' genomes may be genuine (Table S1). Homologs of the LPP-2 plasmid were observed only in *Pantoea* isolates from among *P. agglomerans*, *P. vagans*, *P. anthophila*, and *P. eucalypti*. The *repB* phylogeny supports the presence of two LPP-2 groups (Fig. S2). LPP-2-group I is shared among *P. agglomerans*, *P. vagans*, and *P. eucalypti*, while LPP-2 group II is specific to *P. anthophila* isolates with the exception of one *P. vagans* isolate that carries a group II LPP-2 plasmid. The distribution of LPP-2 homologs within the *Pantoea* species suggests that the LPP-2 acquisition corresponds to the divergence of LPP-2



	Chromosome	LPP-1	LPP-2
①	264(88.3)	40(97.6)	39(92.9)
②	249(83.3)	34(82.9)	42(100)
③	116(38.8)	15(36.6)	10(23.8)
④	290(97.0)	41(100)	-
⑤	291(97.3)	40(97.6)	-
⑥	294(98.3)	41(100)	-
⑦	248(82.9)	41(100)	-
⑧	91(30.4)	11(26.8)	22(52.4)
No. genes	3,584	285	81
% vertical	76.16	89.12	70.37

Fig. 2. LPP-1 and LPP-2 are vertically inherited plasmids. Replicon phylogenies were inferred from CSC gene families that are exclusively present in the focal replicon. Internal branches that correspond to major species divergence events are highlighted. a) Chromosome phylogeny reconstructed from 299 CSC gene families present in 253 isolates. b) LPP-2 phylogeny reconstructed from 42 CSC gene families present in 87 isolates. c) LPP-1 phylogeny reconstructed from 41 CSC gene families in 143 isolates. The number of isolates used for the analysis of each replicon was adjusted in order to increase the number of CSC families. d) Number (and percentages) of CSC gene trees where the split was observed. The bottom lines show the total number (and percentage) of gene trees having a topology not significantly different from the species tree topology (ie their evolution is dominated by vertical inheritance). e) MFA for LPP-1 (green), LPP-2 (purple), and chromosome (yellow). The number of reads is presented for 1 kb windows as the $\log_2$ ratio of the number of reads divided by the number of reads of a non-replicating sample. Additionally, the number of reads is normalized by the total number of reads per sample. The *x*-axis indicates the relative position on the chromosome, where 0 corresponds to the *oriC* in all three replicons, and 1 and −1 correspond to the replication termination in the chromosome. Dashed lines indicate the *oriC* while continuous lines indicate termination in LPP-1 and chromosome.

hosts within the genus. Furthermore, the LPP-1 and LPP-2 clusters are largely matching main divergence events in the *Pantoea* species phylogeny, suggesting that both plasmids are vertically inherited.

What is the origin of LPP-1 and LPP-2? The RepB of LPP-1 and LPP-2 are members of the same RepB-type family and exhibit high sequence similarity (eg 66% identical amino acids over 90% of the LPP-1 RepB length with the LPP-2 RepB in *P. agglomerans* R1). The *repB* gene in both LPP-1 and LPP-2 is homologous to the *repB* of the RepFIB replicon in plasmid F (65.09% and 62.58%; WP_000361610.1). Thus, both plasmids are classified into the same plasmid incompatibility group, IncFIB, suggesting a common plasmid origin. To examine the evolutionary relations between LPP-1 and LPP-2, we inferred a joint *repB* phylogeny including additional homologs that were identified outside of the *Pantoea* genus. The sequence similarity search against plasmid sequences yielded only few putative homologs. The homologs found by sequence similarity search have at least 58% identical amino-acids and the alignment covers at least 95% of the query sequence. The LPP-1 *repB* genes had homologs in the megaplasmid of the plant-growth–promoting *Duffyella gerundensis*, a sister species of *Pantoea*, within the same *Erwiniaceae* family (Li et al. 2021; Soutar and Stavrinides 2022). Two homologs of the LPP-1 *repB* were identified in *Cronobacter* plasmids that were previously classified as IncFIB incompatibility group (Grim et al. 2013). Putative homologs of the LPP-2 *repB* were identified in a megaplasmid from *Sodalis praecaptivus*, which was isolated from a hand wound following impalement with a tree branch and is closely related to *sodalis*-allied of insect endosymbionts (Clayton et al. 2012; Oakeson et al. 2014). Another homolog was identified in *Rahnella sikkimica*, which was isolated from soil, yet its gene content hints at a plant-associated lifestyle (Kumar et al. 2022). Notably, the closest relative plasmids identified for both LPP-1 and LPP-2 among non-*Pantoea* species are large, i.e. can be considered megaplasmids (>350 kbp). It is therefore tenable to hypothesize that LPP-1 and LPP-2 diverged from megaplasmid ancestors. The root of the *repB* phylogeny was inferred at the split between the LPP-1 and LPP-2 homologs (Fig. 1c), indicating an independent origin of the two plasmids in *Pantoea*.

LPP-1 and LPP-2 are Vertically Inherited Within Their *Pantoea* Host Lineage

To examine evidence for vertical inheritance in the evolution of LPP-1 and LPP-2, we compared their phylogeny with that of the *Pantoea* chromosome. The phylogenies of the three replicons (ie both plasmids and the chromosome) are largely compatible, except for two splits (marked (3) and (8) in Fig. 2a–c). To quantify the incongruence among the three replicon phylogenies, we examined the individual phylogenies of complete single-copy (CSC) gene families (ie gene families universally present in a single-copy across all studied genomes). The split of *P. eucalypti* and *P. vagans* (marked [3]) was observed in the chromosome and LPP-1 phylogenies but not in the LPP-2 phylogeny. That split is well supported by chromosomal and LPP-1 CSC gene trees and had a lower support in the LPP-2 CSC gene trees (Fig. 2d). The other incompatible split corresponds to the split of *P. anthophila* and *P. vagans* from the remaining species in the LPP-2 phylogeny (marked [8]), which is not present in the chromosome or the LPP-1 phylogenies. This split is supported by about half of the CSC genes in LPP-2, and can be observed in about a third of the chromosomal and LPP-1 CSC gene trees (Fig. 2d). Taken together, the inferred split between *P. eucalypti*, *P. vagans*, and

P. anthophila and the remaining species is well supported by the data, while conflicting splits in gene trees show evidence for an ongoing process of incomplete lineage sorting among these three species.

To examine further evidence for co-divergence of LPP-1, LPP-2, and their hosts, we tested for compatibility of plasmid single-copy gene trees with the main speciation events as observed in the chromosome phylogeny. The test was applied to replicon-specific single-copy gene families (ie complete and partial; partial gene families appear in only some genomes) comprising ≥4 sequences in at least two species. Constraining the tree topologies according to the species topology showed that the topology of most single-copy gene families is not significantly different from the species phylogeny (Fig. 2d). Thus, the LPP-1 and LPP-2 phylogenies indicate the presence of mechanisms for stable plasmid inheritance of both plasmids.

As determined by their *repB* sequence and illustrated by their membership in the IncF subtype incompatibility group, IncFIB, both LPP-1 and LPP-2 are relatives of plasmid F. The co-existence of both plasmids may be explained by the presence of different partitioning systems and by differences in the specificity of their RepB for the binding sites in the plasmid origin of replication (*oriC*) (Villa et al. 2010). Their basic replicon structure is similar to that of iteron-regulated replicons, such as the RepFIB replicon present on the 7.6-kb *Eco*RI f7 fragment of plasmid F or the 4.3-kb *Eco*RI fragment of plasmid P307. This replicon has been found to be maintained at a very low copy number in *E. coli* (Ishino et al. 1987; Saul et al. 1989). In order to test whether LPP-1 and LPP-2 behave as single RepFIB replicons similarly, we determined the plasmid copy number (PCN) of the LPP-1 and LPP-2 by digital droplet PCR (ddPCR) in *P. agglomerans* strain R1 stationary cells. Both plasmids have a PCN of one per chromosome, indicating that plasmid and chromosome replication are coordinated (Fig. S5).

To examine the plasmid replication dynamics relative to the chromosome replication, using marker frequency analysis (MFA; Skovgaard et al. 2011; Val et al. 2016) of *P. agglomerans* strain R1 during early exponential growth. The results show that LPP-1 replication was initiated after 56% of the chromosome has been replicated, and terminates with the chromosome replication. The LPP-1 replication pattern thus follows the "termination synchrony" model as described in *V. cholerae* (Fig. 2e; Rasmussen et al. 2007). The relatively small size of LPP-2 does not enable a reliable evaluation of synchronized replication with the chromosome. Taken together, we conclude that the evolution of both plasmids is governed by vertical inheritance—a key hallmark of plasmid domestication.

The LPP-1 Pangenome is Shared With the Chromosome and Comprises Core Genes

To evaluate the plasmid contribution to the *Pantoea* pangenome, we examined the distribution of gene families in plasmids and chromosome. About 10% of the gene families include homologs in both the chromosome and the LPP-1 (Fig. 3a; Fig. S6a). The core *Pantoea* genome at the genus level includes 191 gene families shared between the chromosome and the LPP-1, where 20 gene families comprise only LPP-1 genes. The number of core gene families (present in ≥95% isolates) shared between the chromosome and LPP-1 is lower at the species level and ranges between 25 gene families in *P. eucrina* and 94 gene families in *P. ananatis*. In contrast, the number of core gene families found exclusively in the LPP-1 is higher at the species level, ranging between 37 gene families in *P. agglomerans* and 446 gene families in *P. dispersa* (Fig. 3a). The distribution of shell gene families (present in 10% to 95% of isolates) is likewise characterized by a high frequency of LPP-1 gene families at the species level, while the cloud gene families (present in ≤10% isolates) are mostly replicon specific (Fig. S6a). Gene families shared exclusively between LPP-2 and the chromosome are rather rare in the *Pantoea* core genome (21 families; Fig. 3a). Nonetheless, gene families including exclusively LPP-2 genes are found in the core genome of *P. vagans* and the shell genome of *P. vagans*, *P. agglomerans*, *P. eucalypti*, and *P. anthophila*. Gene families shared exclusively between the LPP-1 and LPP-2 are found only in the shell and cloud pangenomes (51 gene families). Taken together, the *Pantoea* core pangenome is distributed across LPP-1, LPP-2, and the chromosome, where plasmid-specific gene families are observed depending on the taxonomic level. The presence of LPP-2-specific genes in the LPP-2 host species adds support to our inference of LPP-2 vertical inheritance, while the presence of LPP-1-specific families at the genus level is in accordance with ancient origins of the LPP-1.

The functional composition of gene families shared between LPP-1 and the chromosome is similar to the functional composition of the LPP-1 pangenome at the species level (Fig. 3b; Fig. S6b). Most gene families shared between LPP-1 and the chromosome are annotated to have transcription-related functions, followed in frequency by gene families classified as carbohydrate and amino-acid metabolism (Fig. 3b). Additionally, transcription-related functions are prevalent in multicopy gene families that harbor homologs in the LPP-1 and the chromosome of the same isolate. The proportion of LPP-1 gene families classified into the functional categories is similar among *Pantoea* species for most categories (Fig. 3b; Fig. S6b; Table S3). The distribution of transcription-related gene families stands out as an exception: in *P. ananatis*, *P. allii*, and *P. stewartii* the frequency of chromosomal gene families in the transcription category is markedly higher compared to LPP-1 while in the remaining species, the frequency of LPP-1 gene families in the transcription is slightly higher than the chromosomal ones. The differences in LPP-1 functional composition among *Pantoea* species are in line with the species-specific evolution of the LPP-1 taxonomic units (Fig. 1b). In addition to transcriptional regulators, the core genome exclusively encoded by the LPP-1 includes *thiS/G* and *ribH* genes that are involved in thiamine and riboflavin biosynthesis, respectively. Additionally, the core LPP-1 genome comprises genes involved in amino acid and carbohydrate metabolism, as well as a malate:quinone oxidoreductase that has been shown to link the TCA cycle with respiration in other bacteria (Spahich et al. 2016).

To further examine the evolution of LPP-1 gene families, we compared the genomic location of all gene families that include at least ten LPP-1-encoded genes among *Pantoea* species. The replicon presence/absence pattern (PAP) reveals core *Pantoea* gene families comprising LPP-1 genes as well as chromosomal and LPP-2 homologs. The PAPs reveal two main patterns of gene families comprising LPP-1 and chromosomal homologs, which are reminiscent of the main *Pantoea* species phylogeny (Fig. 3c). The first PAP is characterized by a combination of chromosomal homologs in *P. dispersa* and *P. eucrina* and LPP-1 homologs in the remaining species, or the other way around: LPP-1 homologs in *P. dispersa* and

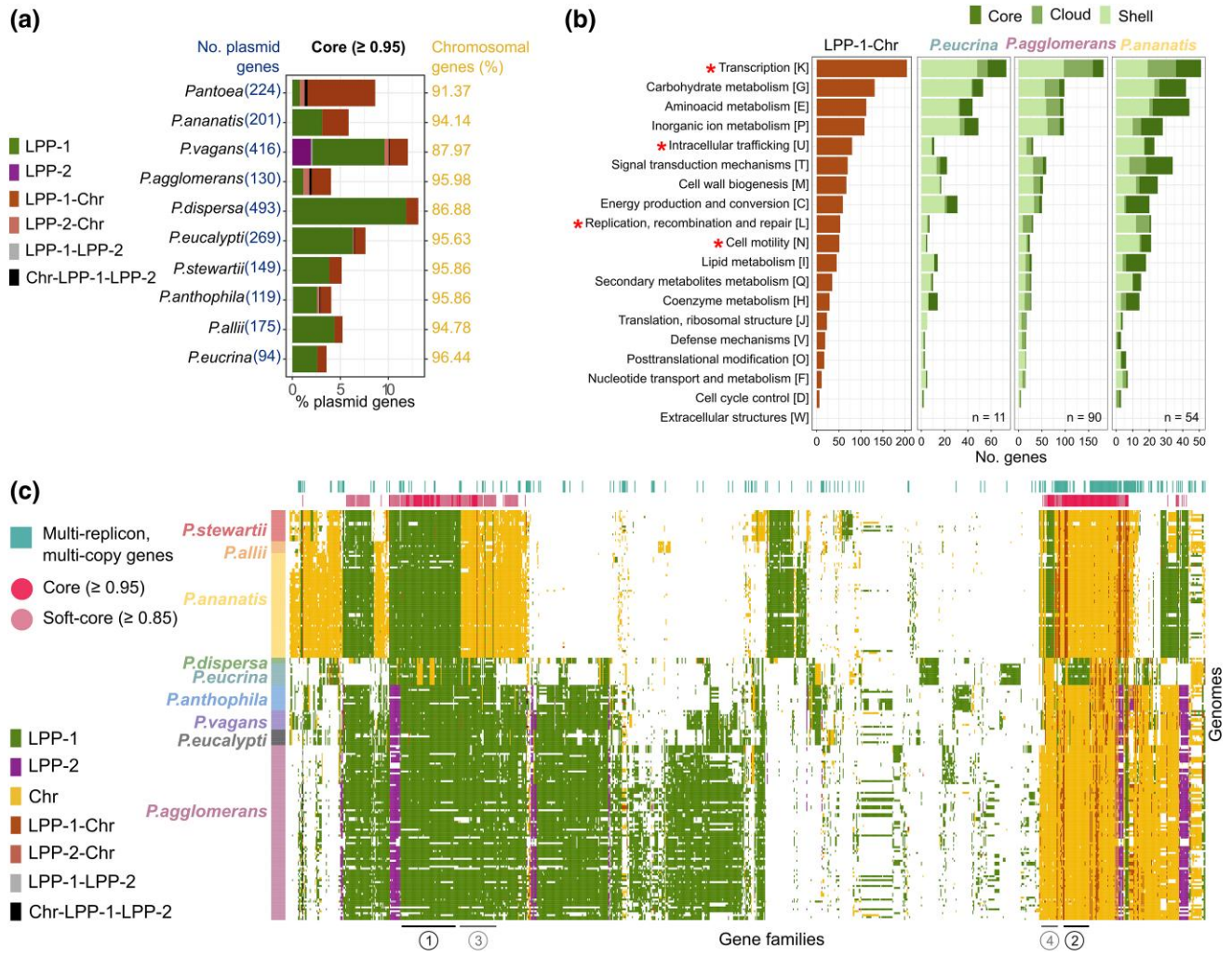


Fig. 3. The LPP-1 pangenome is shared with the chromosome and comprises a high frequency of transcription-related genes. a) Distribution of core gene families in the *Pantoea* pangenome in 211 isolates harboring the LPP-1 plasmid. Three out of the 214 original isolates carrying the LPP-1 were excluded from this analysis due to potential assembly artifacts (Figs. S7 and S8). The stacked bar graph shows the proportion of gene families according to replicons where family-members are present. The number of core gene families that include plasmid homologs is supplied per species on the left (in parentheses). The percent of core gene families exclusively encoded in the chromosome is listed on the right (in yellow). b) Functional annotation of genes shared between LPP-1 and chromosome and the LPP-1 pangenome in representative species: *P. eucrina, P. agglomerans*, and *P. ananatis*. The number of isolates (*n*) in the analysis is shown below. COG categories where the frequency of genes in that category is significantly different among all nine species are marked by a red asterisk. c) Replicon PAP of gene families in the LPP-1 pangenome. Rows correspond to isolates and columns correspond to gene families. The data presented correspond to 1,086 gene families including an LPP-1 homologs in at least 10 isolates. The rows are ordered according to the species phylogeny in (Fig. 1b). Clusters of gene families characterized by species-specific gene presence in plasmids and chromosome are indicated by numbers below the heatmap.

P. eucrina and chromosomal homologs in the remaining species (groups 1 and 2 in Fig. 3c); this PAP is matching the basal *Pantoea* phylogeny split (Fig. 1b). The second pattern corresponds to core genes that are located either on the chromosome in *P. ananatis*, *P. allii*, and *P. stewartii* and on the LPP-1 of the remaining species, or alternatively, genes that are located on the LPP-1 of *P. ananatis*, *P. allii*, and *P. stewartii* and the chromosome of the remaining species (groups 3 and 4 in Fig. 3c). That pattern matches the split between *P. ananatis*, *P. allii*, and *P. stewartii* and the remaining species in the species phylogeny. The correspondence between the replicon PAPs and main splits in the *Pantoea* phylogeny suggests that gene family diversification coincides with two main speciation events in the genus. Notably, most of the gene families comprise a homolog on a single replicon within each isolate, with a total of 465 (14%) gene families including homologs on both LPP-1 and the chromosome within the same isolate (see multi-replicon, multicopy gene families in Fig. 3c). Hence, the presence of homologous genes in multiple replicons within *Pantoea* genomes is rather rare.

Taken together, the *Pantoea* pangenome comprises core LPP-1 genes at the genus level that likely correspond to LPP-1–mediated novel gene acquisition in the common *Pantoea* ancestor. Gene families comprising chromosomal and LPP-1 homologs are the result of either gene translocation to, or from, LPP-1, or alternatively differential loss of homologous genes that were redundant in the chromosome and LPP-1. The phyletic pattern observed in the replicon PAPs indicates that LPP-1 gene acquisition, translocation, and loss coincided with species divergence events in *Pantoea*.

Phylogenies of LPP-1 Genes Provide Support for the Genetic Redundancy Hypothesis

To investigate the relative contributions of the "translocation hypothesis" and the "redundancy hypothesis" to LPP-1

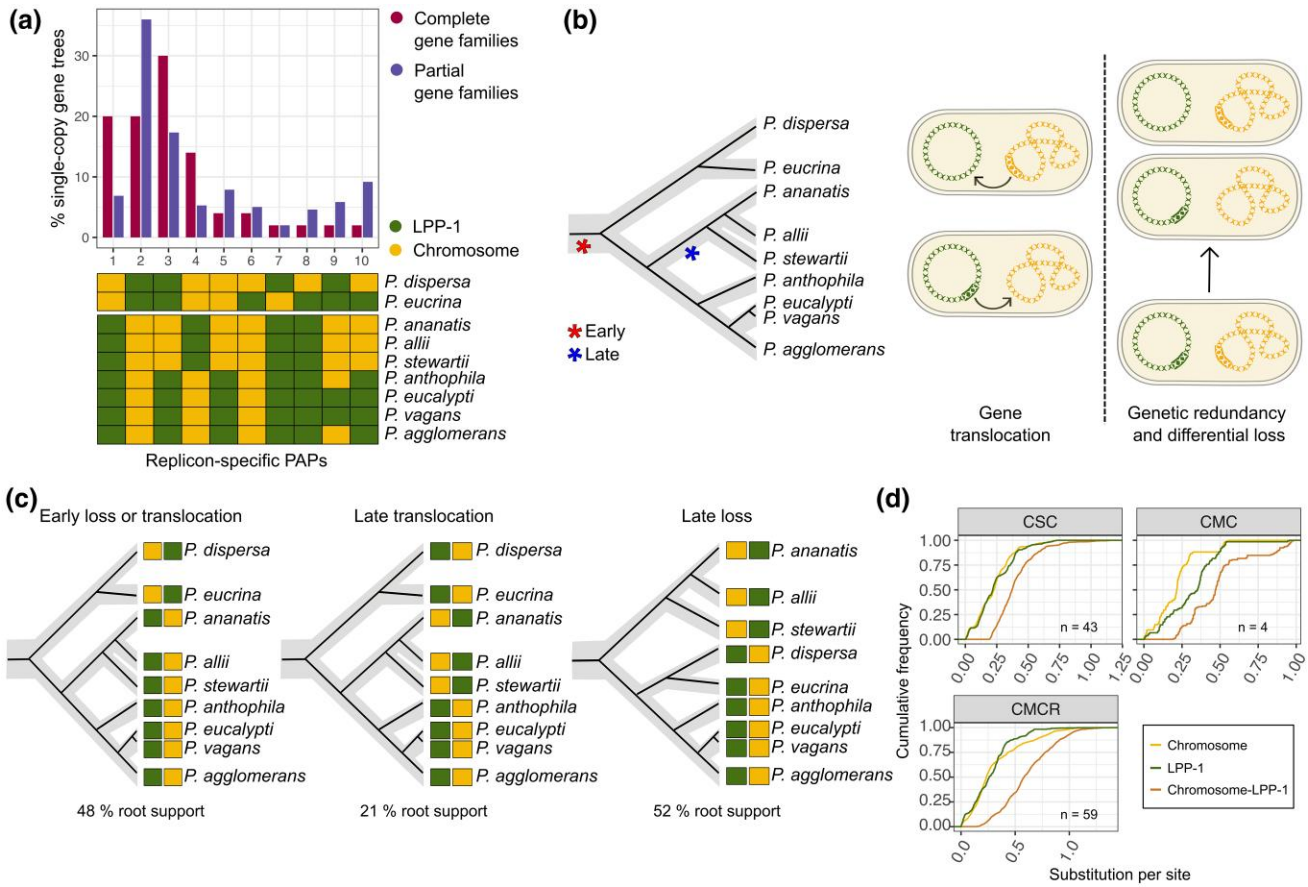


Fig. 4. Phylogenetic analysis and genetic distances provide support for the redundancy hypothesis in LPP-1 evolution. a) Distribution of single-copy gene families in the LPP-1 pangenome that comprise chromosomal homologs. Highly frequent replicon-specific PAPs were identified in 50 CSC gene families that are present in all nine species. A total of 183 partial single-copy gene families found in at least four species, were mapped onto the replicon-specific PAPs. The replicon-specific PAPs are displayed in a heatmap form. b) An illustration of the tested evolutionary hypotheses. Putative branches matching the occurrence of early and late events matching the most common PAPs found among single-copy gene families are marked by an asterisk. The cartoons illustrate hypotheses of gene translocation between the LPP-1 and chromosome (left) and ancestral genetic redundancy followed by differential loss (right). c) Expected gene tree rooted topologies and PAPs for the tested scenarios (illustrated in [b]). The percentage of genes in (a) that support each root position is reported below (see detailed report in Table S4). d) Distribution of pairwise phylogenetic distance calculated as substitution per site across complete gene families shared between chromosome and LPP-1 plasmid. Distance distributions are presented for CSC; complete multicopy gene families, where copies are present in a single replicon (CMC) and complete multicopy gene families, where copies can be present in both replicons (CMCR). The number of gene families analyzed (*n*) is presented.

evolution, we examined evidence for gene translocation and differential loss in the evolution of shared core LPP-1 and chromosome gene families. To constrain the effect of phylogenetic biases in the analysis (eg due to a large number of Operational Taxonomic Units [OTUs]; Roettger et al. 2009), and increase the number of CSCs (Vernikos et al. 2015), we analyzed a reduced dataset of 13 representative complete genomes of the *Pantoea* species (Table S1). The largest category of gene families has a replicon PAP that is matching the basal *Pantoea* phylogeny split (40% CSC genes and 43% partial single-copy genes; Fig. 4a). The second most frequent gene families category has a replicon PAP that is matching the split between *P. ananatis, P. allii*, and *P. stewarti* and the remaining species (44% CSC genes and 22.6% partial single-copy genes; Fig. 4a). The main replicon-specific PAPs of gene families observed in the preliminary dataset are reconstituted in the reduced dataset; hence we conclude that the reduced dataset is representative of the major PAPs we observed.

In accordance with the most frequent patterns of gene family distribution in LPP-1 and the chromosome, we hypothesized two alternative evolutionary scenarios. The *early translocation or loss* scenario is proposed for gene families where the replicon-specificity corresponds to the basal species phylogeny split (PAPs 1, 2 in Fig. 4a and groups 1, 2 in Fig. 3c). According to this *early* scenario, the LPP-1 acquisition was accompanied either by gene translocation between LPP-1 and the chromosome, or, if a chromosomal homolog existed, the LPP-1 acquisition was followed by loss of either the chromosomal or the plasmid homolog (Fig. 4b). The second scenario—*late translocation or loss*—is proposed for gene families where the replicon-specificity matches the split between *P. ananatis, P. allii*, and *P. stewartii* and the remaining species (PAPs 3, 4 in Fig. 4a and groups 3, 4 in Fig. 3c). According to the *late* scenario, the translocation or loss event occurred during the divergence of *P. ananatis, P. allii*, and *P. stewartii* ancestor (Fig. 4b). Note that the most frequent PAP points toward the loss or translocation of plasmid genes rather than the loss or translocation of chromosomal genes in this lineage (see PAP 3 vs. PAP 4 Fig. 4a), which matches the observation of smaller LPP-1 size in *P. ananatis*, *P. allii*, and *P. stewartii* (Fig. S3). The phylogenies of gene families having a PAP matching to the *early translocation or loss* scenario are

expected to have a similar topology to that of the species tree—that is, with a root position matching the basal species phylogeny split. Indeed, we find that the most frequently observed root position in these gene families is the basal *Pantoea* split (48% of gene trees), followed by a root position on the branch leading to *P. ananatis, P. allii*, and *P. stewartii* (22% gene trees) (Fig. 4c; Table S4). Furthermore, the basal *Pantoea* split is significantly supported as the root position among phylogenetic trees reconstructed from the *early* PAPs (False Discovery Rate [FDR] adjusted *P*-value = 0.0221, using two-sided Wilcoxon test). Note that the *early* tree topology (Fig. 4c) does not enable distinguishing between translocation or loss events. In contrast, the root inference for gene families matching the *late* scenario enables a distinction between translocation and loss, since these two events induce different root positions (Fig. 4c). The most frequent root position in the *late* gene families is the branch that splits chromosomal and plasmid homologs (52% of the gene trees), followed by a root position on the branch leading to *P. eucrina* and *P. dispersa* (21% gene trees; Fig. 4c; Table S4). Furthermore, the root position expected for *late loss* is significantly more supported in trees of the *late* PAP category compared to alternative root positions (FDR adjusted *P*-value = 0.0218, using two-sided Wilcoxon rank sum test).

The most frequent rooted tree topologies in both *early* and *late* scenarios comprise a deep split between chromosomal and plasmid homologs, which appears in most gene trees having more than one plasmid homolog (68%; 186 genes). Thus, suggesting an ancient origin of chromosomal and plasmid homologs for most gene families that is more consistent with gene loss. To further distinguish between scenarios of gene loss or translocation, we compared the pairwise phylogenetic distance between homologs present within-replicon and between-replicons in the chromosome and LPP-1 (Fig. 4d). The distance between homologous gene pairs in the same replicon in CSC families was significantly different than the distance between gene pairs in different replicons, with the median within-replicon distances smaller than the median between-replicon distances (see details in Table S5). A significant difference was furthermore observed in all gene family categories, including partial single-copy and multi-copy gene families, as well as those gene families that harbor chromosomal and LPP-1 homologs within the same strain (Fig. 4d; Table S5; Fig. S9). Similar differences were observed in the number of inferred synonymous substitutions (dS) when comparing homologous gene pairs within and between replicons, across all gene family types (including complete and partial gene families; Fig. S10; Table S5). Hence, the observed differences in evolutionary distances are unlikely to stem from differences in selective pressure between the plasmids and chromosome. Comparing the distance distribution to core chromosome-specific gene families further shows that the observed differences in phylogenetic distances between core chromosomal and LPP-1 genes are significantly larger than the expected distances by species divergence alone (P-value < 2.2×10^{-16}, using two-sided Wilcoxon test; Fig. S11). Consequently, we conclude that the divergence of most chromosomal and LPP-1 genes preceded the *Pantoea* species divergence events.

Taken together, we conclude that the LPP-1 diversification occurred during two main evolutionary events matching early and late species divergence episodes. The results of the phylogenetic analysis and comparison of gene distances are in support of early divergence of chromosomal and LPP-1 genes. If translocation of core genes between replicons occurred, it is a very rare event. The phylogenetic signal is more concordant with differential loss of homologs following the LPP-1 acquisition that is consistent with the redundancy hypothesis for the evolution of core genes in domesticated plasmids.

The LPP-2 Evolution Reveals Early Stages of Plasmid Domestication

The plasmid LPP-2 has a limited distribution within the genus and its contribution to *Pantoea* physiology remains understudied. Nonetheless, its presence in species sharing a common ancestor and our inference of LPP-2 vertical inheritance suggests that LPP-2 plasmid is domesticated in the analyzed host species (Figs. 1b and 2). The functional composition of LPP-2 is similar among *Pantoea* species (Fig. S12), which is in support of a recent acquisition (compared to LPP-1). The majority of LPP-2 gene families shared with the chromosome encode for transcription-related protein functions, followed by intracellular trafficking, cell motility, and signal transduction mechanisms (Fig. S12). Examining the replicon PAP of LPP-2 gene families reveals four main PAP patterns. The most frequent PAP corresponds to gene families that are mostly LPP-2-specific (Fig. 5a). Two additional frequent PAPs correspond to gene families comprising LPP-2 and chromosomal homologs; one of these PAPs contains several gene families that have multiple homologs in the same isolate (group 1, Fig. 5a). A total of 113 (25%) gene families in the LPP-2 pangenome have chromosomal homologs in the same isolate (see multi-replicon, multi-copy genes in Fig. 5a). Thus, the frequency of shared gene families with the chromosome is higher in the LPP-2 pangenome compared to the LPP-1 pangenome. The second PAP comprises gene families that have a chromosomal homolog in the genomes of those species that do not harbor LPP-2 (group 2, Fig. 5a). Similarly to LPP-1, the phylogenetic distance between homologous genes in the same replicon (chromosome or LPP-2) is significantly smaller compared to the distance between genes in different replicons (P-value < 2.2×10^{-6}, using two-sided Wilcoxon test; Fig. 5b). This difference indicates that LPP-2 and chromosomal homologs diverged prior to the diversification of LPP-2-hosts within the *Pantoea* genus. Thus, our results suggest that the LPP-2 acquisition introduced genetic redundancy that was accompanied by differential loss of homologous genes in the chromosome and LPP-2, as expected according to the redundancy hypothesis.

The replicon PAP reveals a fourth PAP pattern of 51 gene families shared between LPP-2 and the LPP-1 of LPP-2-less species (group 3, Fig. 5a). These gene families include several typical plasmid functions (eg active partition and toxin-antitoxin systems). Plasmid LPP-2 furthermore harbors genes that mediate lipid A modifications that have been shown to supply resistance to polymyxin (eg *arnBCADTEF*; Raetz et al. 2007). Moreover, a homolog of organophosphorus hydrolase (*opdE*) is widely present in the LPP-2; a homolog of this enzyme from the plant-associated *Leuconostoc mesenteroides* was recently shown to degrade diverse insecticides (Siddavattam et al. 2003).

The similarity in the pattern of LPP-2 gene sharing with the chromosome to that of LPP-1 and the evidence for LPP-2 vertical inheritance (Fig. 2), indicate that LPP-2 is a domesticated plasmid. Compared to LPP-1, the LPP-2 acquisition was late in the *Pantoea* evolution; the LPP-2 contribution to the *Pantoea* species core genome is rather small and LPP-2 genes

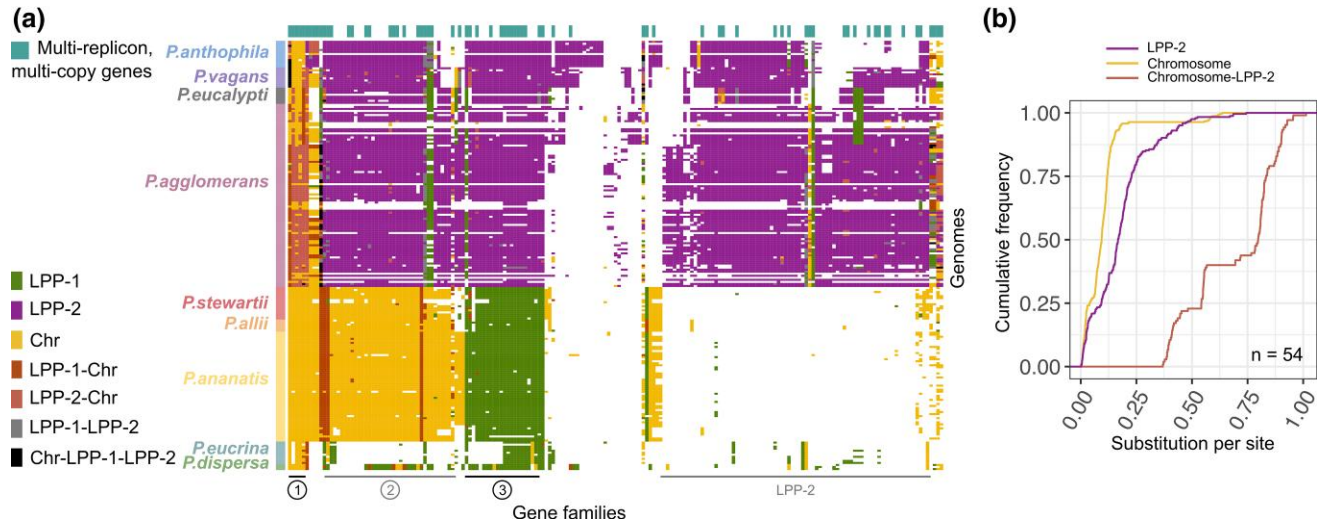


Fig. 5. The LPP-2 pangenome shows a signature of early stages of plasmid domestication. a) PAP of LPP-2 pangenome according to replicons. The analysis was restricted to 211 isolates included in Fig. 3c and 189 protein families which are present in at least 10 LPP-2 plasmids. Each row corresponds to a genome and each column corresponds to one gene family. The rows are clustered according to the species phylogeny (Fig. 1b). Groups of gene families characterized by species-specific replicon PAP are indicated by numbers below the heatmap. b) Distribution of pairwise phylogenetic distance calculated as substitution per site across gene families shared between LPP-2 and the chromosome. Only species carrying the LPP-2 plasmid were considered in the analysis.

having homologs in the chromosome are more frequent compared to LPP-1. Taken together, the LPP-2 pangenome properties support the inference that LPP-2 is at an early stage of plasmid domestication.

Discussion

The evolution of adaptive traits via horizontal gene transfer can facilitate the acquisition of novel ecological niches in prokaryotes (Hacker and Kaper 2000; Ochman et al. 2000). Horizontal gene transfer is a major driver of prokaryote speciation; however, most models of species divergence focused on the role of homologous recombination (Gogarten et al. 2002; Vos 2011; Wiedenbeck and Cohan 2011; Thane Papke et al. 2013), while plasmid acquisition was considered as a driver of species divergence in only few taxa, e.g. rhizobia (Johnston et al. 1978) and Yersinia (Achtman et al. 1999; Drew et al. 2021). The *Pantoea* species we examined here are frequently associated with plants. Nonetheless, species in the genus are known to be found in diverse habitats (Dutkiewicz et al. 2016). A recent phylogenetic reconstruction of the genus evolutionary history shows species not associated with plants in a basal position, thus pointing toward *Pantoea* adaptation to the plant habitat as a later adaptation. Our results on the coincidence of LPP-1 and LPP-2 origins with the divergence of *Pantoea* inhabiting the plant environment suggest a contribution of LPP-1 and LPP-2 to the adaptation of a plant-associated lifestyle in *Pantoea*.

The presence of plasmid genes in the species core genome is one hallmark of bacterial taxa that harbor chromids (Harrison et al. 2010). Core gene families comprising plasmid genes are of two types: gene families that are exclusively present in the plasmid and gene families comprising homologs in both the plasmid and chromosome. The latter type of mosaic plasmid–chromosome gene families makes up the majority of the core plasmid genes in *Pantoea* (Fig. 3). The emergence of plasmid-specific core genes in the species pangenome is well explained by species divergence following plasmid acquisition. But what is the origin of mosaic plasmid–chromosome core gene families? The translocation hypothesis posits the occurrence of gene transfer or translocation from the chromosome to the plasmid (Moreno 1998). However, a recent study suggests that gene transfer between plasmids and chromosomes is rather rare and mostly depends on gene association with mobile genetic elements (eg transposons) (Kadibalban et al. 2024). Hence, while translocation of genes from the chromosome to the chromid has been reported in rhizobia and *Burkholderia* (Guo et al. 2010; diCenzo et al. 2013), such instances may be anecdotal or transient. While our results do not rule out the occurrence of gene translocation between the plasmids and chromosome in *Pantoea*, the phylogenetic analysis support the notion that translocations have been rare. Our inference for the evolution of mosaic plasmid–chromosome core gene families is rather in agreement with the redundancy hypothesis that implicates genetic redundancy and differential loss (diCenzo and Finan 2017; Hall et al. 2022). The redundancy hypothesis is further supported by the observation that ca. 14% of the chromosomal genes in *S. meliloti* are functionally redundant with the resident plasmids (diCenzo et al. 2018). Notably, in most of the mosaic core gene families, we identified a robust split between the plasmid and chromosomal homologs that is highly supported by phylogenetic inference and comparison of genetic distances (Fig. 4d). Our results thus indicate that chromosomal and plasmid homologs in *Pantoea* diverged prior to main speciation events in the genus. The plasmid homologs in those core gene families shared between plasmid and chromosome are thus better described as xenologs (Fitch 1970), rather than paralogs.

The evolution of genetic redundancy has been extensively studied in the context of whole genome duplication (WGD) in eukaryotes. Examples are studies of WGD in the evolution of yeast (Wolfe and Shields 1997) and *Arabidopsis* (Kowalski et al. 1994), which showed that most duplicated genes are differentially lost over time and return to a single-copy state (Li et al. 2016). An important difference between WGD and megaplasmid acquisition is the level of genetic redundancy, which is a determinant of the fate of duplicated genes (Kuzmin et al.

2022). WGD-derived paralogs are preliminarily identical while chromosomal genes and the plasmid xenologs are most likely preliminarily diverged, both in their sequence (Figs. 4d and 5b) and the mode of transcriptional regulation. Our inference points toward the differential loss of homologous genes following the LPP-1 (or LPP-2) acquisition, similarly to post-WGD events. The *late* scenario we tested here (Fig. 4) coincides with the emergence of commonly recognized plant pathogens in the species *P. ananatis, P. allii*, and *P. stewarti*. It is tenable to hypothesize that the LPP-1 size reduction in those species is associated with an adaptation of a pathogenic lifestyle in those species. Our results further show a high frequency of redundancy in gene families involved in transcription-related functions. These findings are in agreement with previous reports on the enrichment of transcriptional regulators in chromids (diCenzo and Finan 2017). The retention of functionally redundant homologous genes potentiates the evolution of transcriptional plasticity that can contribute to rapid physiological response to diverse environmental conditions (as shown in yeast; Mattenberger et al. 2017). Phenotypic plasticity is an important property in bacteria leading a bi-phasic lifestyle (reviewed in Obeng et al. 2021). Indeed, *Pantoea* species are frequently associated with plant (or insect) hosts; however, the mode of *Pantoea* transmission along the host life cycle remains to be clarified and reports of isolates from diverse soil habitats indicate that *Pantoea* is well adapted to host-associated as well as free-living lifestyles (Walterson and Stavrinides 2015). The LPP-1 and LPP-2 domestication most likely contributed to the versatility of *Pantoea* niche occupation.

Both LPP-1 and LPP-2 are related to plasmid lineages that are characterized by medium to large size and a gene content enriched in metabolic functions, similarly to other plasmids that are prevalent in plant microbiomes (Finks and Martiny 2023). LPP-1 and LPP-2 do not harbor homologs of *mob* relaxases; hence, they are unlikely to be mobile. Nonetheless, the absence of plasmid mobility mechanisms does not preclude an evolutionary origin of these plasmids from a mobile ancestor, as plasmid mobility may be lost during plasmid domestication (Coluzzi et al. 2022; Hanke et al. 2024). Plasmid domestication following the loss of plasmid mobility is bound to generate genetic polymorphism within the population, comprising plasmid-hosts and non-hosts sub-populations—that is the initial step of sympatric speciation. The evolution of stable plasmid inheritance will preclude the emergence of non-hosts from the host population, while the loss of plasmid mobility prevents gene flow from hosts to the non-hosts subpopulation. The emergence of species isolation (termed also "reproductive barriers") may be further promoted by selective sweep for the plasmid presence (ie for a trait supplied by the plasmid) and the consequent elimination of non-hosts from the population. Alternatively, migration of the hosts subpopulation to a new niche made accessible by the plasmid presence may likewise lead to the emergence of new species. In the absence of genetic mechanisms for the translocation of beneficial plasmid genes to the host chromosome (eg transposable elements), domesticated plasmids are expected to become an integral component of their host genome.

Methods

Genomes Dataset

Genome assemblies of draft and complete genomes of *Pantoea* isolates were initially downloaded from RefSeq (version of 03/2021; O'Leary et al. 2016). Three additional *P. dispersa* genomes were downloaded from NCBI and added to the analysis *a posteriori*, due to the basal position of the species in the species tree (Fig. 1) and putative importance to the analysis and conclusions. The genomes of insect symbionts were excluded as their genomes underwent extreme genome reduction (Otero-Bravo et al. 2018). Such fast-evolving genomes may hinder the reconstruction of the species phylogeny. Therefore, we decided to restrict our study to plant-associated species (Bergsten 2005). Additionally, we analyzed only classified species for which there is at least one complete genome. The final dataset comprises 33 complete and 220 draft *Pantoea* genomes. The complete list of genomes can be found in Table S1.

Gene Family Reconstruction

The reconstruction of gene families was performed as previously described in (Wang and Dagan 2024). The protein sequences of each possible pair of isolates were compared with BLAST (Altschul et al. 1990) using blastp (version 2.11.0; Camacho et al. 2009) and applying a threshold of e-value $\leq 10^{-6}$. Reciprocal best blast hits were extracted and Needleman-Wunsch global alignments were computed with parasail using gap open and extension penalties of 11 and 1, respectively and the blosum62 matrix for amino-acid replacement (Daily 2016). All pairs of sequences with ≥30% identical amino acids were clustered into gene families with MCL using an inflation index of 2 (Enright et al. 2002). The gene family reconstruction pipeline was used two times to reconstruct gene families among all *Pantoea* genomes, and among the *P. dispersa* genomes that were added *a posteriori*. The gene families of the two datasets were merged as follows: sequence similarity search among genomes of the two groups was performed and hits with a sequence similarity of ≥30% identical amino acids were retained. The average sequence similarity between each possible gene pair in a gene family was calculated. Gene family pairs with an average sequence similarity of ≥30% identical amino acids were retained, and the best match were identified as the same gene family (with no distinction between orthologs, paralogs, or xenologs).

Replicon Assignment and Plasmid Classification

To identify and classify plasmid sequences, we compared the gene content across all contigs in the dataset. The overlap coefficient (Vijaymeena and Kavitha 2016) in protein families encoded by each pair of contigs estimated as:

$$Overlap\ (Contig1,\ Contig2) = \frac{|Contig1\ \cap\ Contig2|}{min(|Contig1|, |Contig2|)}.$$

Contig pairs having ≥0.8 overlap coefficient and a length ratio ≤1/2 were retained for clustering. These thresholds were selected based on the minimum overlap found among 11 LPP-1 plasmid and 36 chromosomal replicons (Fig. S13). Contig pairs were clustered using MCL with an inflation index of 2 (Enright et al. 2002). A total of 41 clusters including at least one plasmid identified as such in a completely sequenced genomes included plasmid-like 462 contigs. Most of these contigs (366 contigs, 80%) harbored one out of the two most common replication initiation (Rep) gene families in the dataset, indicating that the genus hosts mainly two major plasmid lineages. The largest plasmid lineage comprises sequences that were identified to show high similarity to the LPP-1 (De Maayer et al. 2012; Fig. S14), followed by a

plasmid lineage comprising sequences with high similarity to the LPP-2 plasmid (Soluch et al. 2021; Fig. S15). Rep gene family phylogenies were reconstructed to further validate the presence of such plasmid lineages (Fig. S2). In total, we identified 214 LPP-1-related contigs in six clusters of homologous plasmids and 114 LPP-2-related contigs in four clusters of homologous plasmids. To control for the quality of our plasmid groups, we additionally estimated the Average Nucleotide Identity (ANI; Jain et al. 2018) among LPP-1 and LPP-2 contigs and inspected the consistency in gene content within plasmid group. Figure S13 shows the ANI shared among LPP-1 plasmids. Notably, all LPP-1 plasmids were inferred as a single contig. The clustering using ANI as a distance metric yields similar results to that using a gene content distance (Fig. 1) and the plasmid size within group is overall consistent, further confirming the quality of our plasmid groups. The inferred LPP-1 plasmids in this study have similar size and show high sequence similarity to previously annotated LPP-1 plasmids. Figure S14 shows the ANI among LPP-2 plasmids. The inspection of sequence similarity among plasmid contigs revealed the presence of two plasmid clusters (LPP-2 groups Ia and Ib) that show low sequence similarity to one another while they both show high similarity with another plasmid group (LPP-2 group I). Notably, LPP-2 groups Ia and Ib co-occur in five isolates, which made us suspect of a possible over-segmentation. Thus, we decided to examine the gene content of the plasmids (Fig. S16). We observed that LPP-2 Ia and LPP-2 Ib complement the presence of different regions from plasmids belonging to group LPP-2 I. We suspect that contigs in plasmid groups LPP-2 Ia and Ib likely belong to the same original plasmid and that their fragmentation is likely an assembly artifact. Therefore, we treated these two contigs as a single plasmid sequence in the following analyses. Clusters of homologous LPP-1 were termed based on previous classification of LPP-1 groups (De Maayer et al. 2012). Note that the LPP-1 group III is not described in our dataset. Preliminary analyses including a larger number of isolates showed that the two strains reported to harbor this plasmid group (*Pantoea sp.* At-9b and *Pantoea cypripedii* LMG265^T) are rather far relatives from the here studied taxa, and therefore we decided to exclude them from the current study. The remaining plasmid groups are either inconsistently present in the lineages or have a small number of contigs and therefore were not considered further in the analysis (Fig. S17). A plasmid network was reconstructed with Cytoscape v3.10.1 (Shannon et al. 2003) using a threshold of overlap coefficient ≥ 0.4.

Chromosomal contigs were further validated as such using PlasClass (Pellow et al. 2020), which calculates the probability of a contig to be a plasmid. We used contigs of known plasmids and chromosomes in the dataset as a training set and classified contigs as chromosomal when the probability of being a plasmid was ≤ 0.045. This threshold was chosen based on the probabilities inferred for chromosomes from completely sequenced genomes in the dataset. As a final quality check, we examined the distribution of chromosome size per isolate after assigning chromosomal sequences and compared it to the expected chromosomal genome size in each species (Fig. S18).

Inference of Species, Plasmid, and Gene Phylogenies

The *Pantoea* species tree was inferred from 299 single-copy universal gene families that are exclusively found on the chromosome. Alignments for each gene family were reconstructed with MAFFT with default parameters (Katoh et al. 2002). Maximum likelihood trees were inferred with IQ-TREE version 2.0.3, using the partitioned analysis (Nguyen et al. 2015). We used a general time reversible substitution model (GTR), using empirical base frequencies (+F). Gene trees were also reconstructed for replicon exclusive and replicon-shared gene families, using the same substitution model parameters. The root of the species tree was inferred by applying a phylogenomic rooting approach (Tria et al. 2023) using our dataset and, additionally, using the dataset provided by Crosby et al. (2023) (see also Note S1). The root inference of individual gene families was done with Minimal Ancestor Deviation (MAD) (Tria et al. 2017). Isolates branching within the nine clades corresponding to nine main *Pantoea* species were classified as the corresponding species. Plasmid trees were inferred using a subset of isolates to include the maximum number of CSC gene families. To identify putatively related plasmids outside of the *Pantoea* genus, we searched for similar sequences to the RepB proteins in the NCBI nonredundant database (version of 05/2024) using the online version of blastp (Altschul et al. 1990). The search was done using *Pantoea agglomerans* strain R1 as a query (WP_010246820.1; see Table S1). Only hits identified in annotated plasmids were considered. The inference of multiple sequence alignment and phylogenetic trees of plasmid genes was performed using the tools and parameters described above for the species tree.

Verticality Tests

Replicon-specific single-copy gene families that comprise at least four sequences from at least two species were tested for deviation from the species phylogeny. A constrained topology was generated according to the species tree, where the sequences are split into the nine studied species. Multiple genes within the same species were added as a polytomic group. A constrained species trees was adapted to each gene family by pruning the species tree to contain only the sequences found in a focal gene family using ETE3 Toolkit (Huerta-Cepas et al. 2016). For each gene family, two phylogenies are inferred: an unconstrained maximum-likelihood phylogeny and a constrained-topology phylogeny using IQ-TREE version 2.0.3 (Kishino and Hasegawa 1989; Nguyen et al. 2015). The likelihood of the resulting topologies was compared using the Kishino-Hasegawa test using IQ-TREE. If the likelihood of the two phylogenies were not significantly different, the gene family is concluded to have evolved vertically.

P. agglomerans Strain R1 Growth Conditions and DNA Isolation

The *P. agglomerans* strain R1 is a member of the native, seed-borne microbiota of *Triticum aestivum* that was isolated from surface-sterilized seeds of the winter wheat cultivar Runal harvested in 2017 at the experimental farm Hohenschulen of Kiel University (Soluch et al. 2021). The strain was routinely grown in Reasoner's 2A medium at 30 °C. Total DNA was isolated from stationary cultures of *P. agglomerans* using the Wizard Genomic DNA Purification Kit (Promega). Concentration and quality of the extracted DNA was assessed the Qubit device (Invitrogen by Life Technologies).

PCN of LPP-1 and LPP-2

ddPCR was used to quantify the replicon copy number/PCN of LPP-1 and LPP-2 in samples of genomic DNA (total

DNA) prepared from stationary cultures of *P. agglomerans* R1 following the protocol of (Hechard and Wang 2023). A combination of four primer/probe pairs, targeting the *oriC* and the region of replication termination (*terC*) of the chromosome (as determined by the MFA analysis) as well as two pairs targeting the *oriC* on LPP-1 and LPP-2, respectively, were designed (Table S6). The fluorophores of the probes were chosen in manner to distinguish between both chromosomal targets as well as the two plasmids allowing a quadruplex reaction setup for the simultaneous detection of two chromosomal targets and each plasmid target (Table S6). The primer pairs were tested for efficiency and specificity by qPCR separately (see Table S6 for details). Standard curves for all primer pairs were linear, with average slopes of −3.97 and −3.64, respectively. Based on the standard curve slopes, amplification efficiencies ranged between 1.78 and 1.88 for all four amplicons. For ddPCR, the concentration of primers and probes was adjusted for a clear separation of negative and positive droplets of the four different amplicons. Good discrimination of all droplet types was observed at a primer concentration of 400 nM and probe concentration of 100 nM (Fig. S5a). In stationary cells used for normalization in the MFA, the ratio of *oriC* to *terC* is almost one, indicating that the *oriC* region and the *terC* region occur in equal numbers. The ratio of the LPP-1 marker relative to the chromosomal *terC* marker, shows that the LPP-1 has a similar abundance than the chromosome, ie its copy number is identical to that of the chromosome. The copy number of LPP-2 appears lower, possibly due to the efficiency of the primer/probe combination used, which may be reflected in the low cluster separation between positive and negative droplets (Fig. S5a, right panel, *x*-axis).

Droplet Digital PCR

A quadruplex reaction setup for the simultaneous detection of two chromosomal and two plasmid-located targets was used for ddPCR. Single reaction volumes for droplet generation were 25 µl and contained 12.5 µl 2× ddPCR Supermix (Bio-Rad), 0.1 µl of each primer (final concentration of 400 nM), 0.25 µl of each probe (final concentration of 100 nM), and 2 µl of template solution (final DNA amount per reaction: 30 pg). The master mix containing all ingredients except the template was prepared and added to the samples of genomic DNA. The samples were thoroughly mixed, briefly centrifuged, and transferred to DG8 cartridges (Bio-Rad) for droplet generation with the QX200 system (Bio-Rad) according to the manufacturer. The generated droplet emulsions were transferred into a semi-skirted 96-well PCR plate (Bio-Rad) and heat-sealed with pierceable foil in the PX1 PCR plate sealer (Bio-Rad). The PCR reaction was performed in a C1000 thermo-cycler (Bio-Rad) with the following parameters: Initial denaturation step (95 °C for 10 min), 40 cycles of 94 °C for 30 s and 60 °C for 60 s, followed by a final denaturation step of 98 °C for 10 min, with a ramp rate of 2 °C/s. Droplets were analyzed with the QX600 droplet reader with simultaneous detection of Cy5, FAM, HEX, and ROX. Data analysis was performed with the QX Manager software, version 2.1.0.25 (Bio-Rad), and with R, version 4.3.0.

Marker Frequency Analysis

Whole genome DNA preparations (gDNA) for sequencing were prepared from either stationary-phase overnight cultures or from cells harvested during early logarithmic growth phase at OD450 ∼ 0.16 using the Wizard® Genomic DNA Purification Kit (Promega). For MFA, gDNA was isolated from independently grown cultures (30 ml) of exponentially growing and stationary cells, inoculated with a single colony. Sequencing using paired-end libraries with an insert size of 350-bp was performed on the Illumina NovaSeq platform with 150-bp read length on each end by NovoGene (Munich, Germany). Samples were sequenced to an average depth of ca. x437. Reads were aligned to the scaffold of the chromosome of strain R1, for which the contigs (NZ_WSSS01000001 to NZ_WSSS01000008) were sorted with MAUVE (Darling et al. 2010) using the genomes the of *P. agglomerans* strains DBM3797 (GenBank: CP086133.1) and L15 (GenBank: CP034148.1) as reference. The contigs of LPP-1 and LPP-2 represent the complete plasmids—confirmed by Sanger sequencing of PCR products bridging the contig junctions. The MFA was performed using the MFA-seq protocol (Batrakou et al. 2020), which uses the localMapper bash script for aligning reads and calculating the sequencing depth in genomic windows, and the R-package Repliscope version 1.1.1 (Müller et al. 2014) for normalization.

Functional Annotation

Functional annotation and COG classification of the shared genes between chromosomes and LPP-1 plasmids found among the set of 13 representative genomes (Table S1) were performed with eggNOG-mapper version 2 (Huerta-Cepas et al. 2019). Briefly, we mapped each gene sequence to the eggNOG database. When annotating gene families, the most common category or categories in each gene family was identified as the COG category for the whole gene family. The majority rule was not applied when annotating individual genes (ie in the species pangenome analyses).

Statistical Comparison of Rooted Tree Topologies

To test for statistical support of the "translocation" and "redundancy" hypotheses, we rooted gene trees in Fig. 4 using MAD (Tria et al. 2017). The percentage of gene families within each scenario ("early" and "late") that support each candidate root was calculated (Table S4, Fig. S1). Gene families included in "early" scenarios feature PAPs 1 and 2, while genes families included in "late" scenarios feature PAPs 3 and 4 as shown in Fig. 4a. The tested candidate roots comprise the set of inferred roots found among the complete set of genes (following the phylogenomic rooting approach; Tria et al. 2023). These roots were also inferred among gene families in the complete dataset and are depicted in Fig. S1. The inferred root in a gene family corresponds to the candidate root with the minimal ancestral deviation. In gene families where the ratio between the best root and the second-best root (Ambiguity Index) is equal or larger than 0.9, the gene family was concluded to support each root with half their weight.

Phylogenetic Pairwise Distance Analysis

The distribution of phylogenetic distances was examined between sequence pairs belonging to partial and CSC gene families shared between the plasmid (LPP-1 or LPP-2) and the chromosome in the set of 13 representative genomes (Table S1). Only families having three or more homologs were included in the analysis. The phylogenetic distance was calculated as the sum of branch length between homologous

gene pairs (ie OTUs) in family gene trees inferred using Maximum Likelihood and the GTR + F evolutionary model using IQ-TREE version 2.0.3 (Nguyen et al. 2015). Genetic distances were calculated also for chromosomal CSC gene families for comparison (CSCs; Fig. S11). Chromosomal genes correspond to 97% of the CSC families (median: 97%, IQR: 1.2%). LPP-1 genes correspond to 3% of the CSC families (median: 2.6%, IQR: 0.4%). LPP-2 genes comprise less than 1% of the CSC LPP-2 families (median: 0.9%, IQR: 0.13%). Similarly, we inferred the number of synonymous substitutions per site (dS) in gene families shared between the LPP-1 and the chromosome. For that purpose, we converted amino-acid alignments into codon alignments using PAL2NAL version v14 (Suyama et al. 2006). Then, gene trees were inferred from the codon alignments with IQ-TREE v2.0.3 (Nguyen et al. 2015), using Maximum Likelihood and GY codon evolutionary model (Goldman and Yang 1994). Finally, dS distances were inferred by Maximum Likelihood with codeml version 4.10.7 (Yang 2007).

Supplementary material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Giddy Landan for insightful comments on the phylogenetic analyses. We thank Kai Ripcke and Shreya Vichare for fruitful discussions and Yiqing Wang, Fabian Nies, and Dustin Hanke for critical comments on the manuscript. This work was supported by the DFG funded CRC1182 Origin and Function of Metaorganisms and ERC grant pMolEvol (grant number 101043835).

Author Contributions

D.R.P. and T.D. designed the study; D.R.P. collected the data and performed all comparative genomics and phylogenetic analyses; N.F.H., F.M., L.H., and P.K. performed the marker frequency analysis. D.R.P. and T.D. interpreted the results with contribution from N.F.H., F.M., L.H., and P.K. D.R.P. and T.D. wrote the manuscript with contributions from all authors.

Conflict of Interest

The authors declare no competing interest.

Data Availability

Sequence reads are available at the European Nucleotide Archive (ENA; accession numbers: ERS27186183 and ERS27186184 for samples in stationary and exponential growth phase, respectively; biosample ids: SAMEA120485497 and SAMEA120485498, for samples in stationary and exponential growth phase, respectively). All other material is supplied in the supplementary information.

References

Achtman M *et al.* *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci U S A*. 1999:96:14043–14048. https://doi.org/10.1073/pnas.96.24.14043.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990:215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Ares-Arroyo M, Coluzzi C, Rocha EPC. Origins of transfer establish networks of functional dependencies for plasmid transfer by conjugation. *Nucleic Acids Res*. 2023:51:3001–3016. https://doi.org/10.1093/nar/gkac1079.

Bánfalvi Z *et al.* Location of nodulation and nitrogen fixation genes on a high molecular weight plasmid of *R. meliloti*. *Mol Gen Genet*. 1981:184:318–325. https://doi.org/10.1007/BF00272925.

Barash I, Manulis-Sasson S. Recent evolution of bacterial pathogens: The Gall-Forming *Pantoea agglomerans* case. *Annu Rev Phytopathol*. 2009:47:133–152. https://doi.org/10.1146/annurev-phyto-080508-081803.

Batrakou DG, Müller CA, Wilson RHC, Nieduszynski CA. DNA copy-number measurement of genome replication dynamics by high-throughput sequencing: the sort-seq, sync-seq and MFA-seq family. *Nat Protoc*. 2020:15:1255–1284. https://doi.org/10.1038/s41596-019-0287-7.

Bergsten J. A review of long-branch attraction. *Cladistics*. 2005:21:163–193. https://doi.org/10.1111/j.1096-0031.2005.00059.x.

Brady CL *et al.* *Pantoea allii* sp. nov., isolated from onion plants and seed. *Int J Syst Evol Microbiol*. 2011:61:932–937. https://doi.org/10.1099/ijs.0.022921-0.

Camacho C *et al.* BLAST+: architecture and applications. *BMC Bioinformatics*. 2009:10:421. https://doi.org/10.1186/1471-2105-10-421.

Castagno LN, Estrella MJ, Sannazzaro AI, Grassano AE, Ruiz OA. Phosphate-solubilization mechanism and in vitro plant growth promotion activity mediated by *Pantoea eucalypti* isolated from Lotus tenuis rhizosphere in the Salado River Basin (Argentina). *J Appl Microbiol*. 2011:110:1151–1165. https://doi.org/10.1111/j.1365-2672.2011.04968.x.

Clayton AL *et al.* A novel human-infection-derived bacterium provides insights into the evolutionary origins of mutualistic insect–bacterial symbioses. *PLoS Genet*. 2012:8:e1002990. https://doi.org/10.1371/journal.pgen.1002990.

Coluzzi C, Garcillán-Barcia MP, de la Cruz F, Rocha EPC. Evolution of plasmid mobility: origin and fate of conjugative and nonconjugative plasmids. *Mol Biol Evol*. 2022:39:msac115. https://doi.org/10.1093/molbev/msac115.

Coutinho T, Venter S. *Pantoea ananatis*: an unconventional plant pathogen. *Mol Plant Pathol*. 2009:10:325–335. https://doi.org/10.1111/j.1364-3703.2009.00542.x.

Crosby KC *et al.* Genomic delineation and description of species and within-species lineages in the genus Pantoea. *Front Microbiol*. 2023:14:1254999. https://doi.org/10.3389/fmicb.2023.1254999.

Daily J. Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. *BMC Bioinformatics*. 2016:17:81. https://doi.org/10.1186/s12859-016-0930-z.

Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*. 2010:5:e11117. https://doi.org/10.1371/journal.pone.0011147.

De Maayer P *et al.* The large universal Pantoea plasmid LPP-1-1 plays a major role in biological and ecological diversification. *BMC Genomics*. 2012:13:625. https://doi.org/10.1186/1471-2164-13-625.

diCenzo G, Milunovic B, Cheng J, Finan TM. The tRNAarg gene and engA are essential genes on the 1.7-Mb pSymB megaplasmid of *Sinorhizobium meliloti* and were translocated together from the chromosome in an ancestral strain. *J Bacteriol*. 2013:195:202–212. https://doi.org/10.1128/JB.01758-12.

diCenzo GC *et al.* Robustness encoded across essential and accessory replicons of the ecologically versatile bacterium Sinorhizobium meliloti. *PLoS Genet*. 2018:14:e1007357. https://doi.org/10.1371/journal.pgen.1007357.

diCenzo GC, Finan TM. The divided bacterial genome: structure, function, and evolution. *Microbiol Mol Biol Rev*. 2017:81:e00019-17. https://doi.org/10.1128/MMBR.00019-17.

Drew GC, Stevens EJ, King KC. Microbial evolution and transitions along the parasite–mutualist continuum. *Nat Rev Microbiol*. 2021:19:623–638. https://doi.org/10.1038/s41579-021-00550-7.

Dutkiewicz J, Mackiewicz B, Lemieszek MK, Golec M, Milanowski J. *Pantoea agglomerans*: a mysterious bacterium of evil and good. Part IV. Beneficial effects. *Ann Agric Environ Med*. 2016:23:206–222. https://doi.org/10.5604/12321966.1203879.

Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*. 2002:30:1575–1584. https://doi.org/10.1093/nar/30.7.1575.

Finks SS, Martiny JBH. Plasmid-encoded traits vary across environments. *mBio*. 2023:14:e0319122. https://doi.org/10.1128/mbio.03191-22.

Fitch WM. Distinguishing homologous from analogous proteins. *Syst Zool*. 1970:19:99. https://doi.org/10.2307/2412448.

Gogarten JP, Doolittle WF, Lawrence JG. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol*. 2002:19:2226–2238. https://doi.org/10.1093/oxfordjournals.molbev.a004046.

Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*. 1994:11:725–736. https://doi.org/10.1093/oxfordjournals.molbev.a040153.

Grim CJ *et al*. Pan-genome analysis of the emerging foodborne pathogen *Cronobacter* spp. suggests a species-level bidirectional divergence driven by niche adaptation. *BMC Genomics*. 2013:14:366. https://doi.org/10.1186/1471-2164-14-366.

Guo F-B, Ning L-W, Huang J, Lin H, Zhang H-X. Chromosome translocation and its consequence in the genome of Burkholderia cenocepacia AU-1054. *Biochem Biophys Res Commun*. 2010:403:375–379. https://doi.org/10.1016/j.bbrc.2010.11.039.

Hacker J, Kaper JB. Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol*. 2000:54:641–679. https://doi.org/10.1146/annurev.micro.54.1.641.

Hall JPJ, Botelho J, Cazares A, Baltrus DA. What makes a megaplasmid? *Philos Trans R Soc B Biol Sci*. 2022:377:20200472. https://doi.org/10.1098/rstb.2020.0472.

Hanke DM, Wang Y, Dagan T. Pseudogenes in plasmid genomes reveal past transitions in plasmid mobility. *Nucleic Acids Res*. 2024:52:7049–7062. https://doi.org/10.1093/nar/gkae430.

Harrison PW, Lower RPJ, Kim NKD, Young JPW. Introducing the bacterial 'chromid': not a chromosome, not a plasmid. *Trends Microbiol*. 2010:18:141–148. https://doi.org/10.1016/j.tim.2009.12.010.

Hechard T, Wang H. Determination of growth rate and virulence plasmid copy number during Yersinia pseudotuberculosis infection using droplet digital PCR. In: Nordenfelt P, Collin M, editors. *Bacterial pathogenesis: methods and protocols*. Springer US; 2023. p. 101–115.

Huerta-Cepas J *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*. 2019:47:D309–D314. https://doi.org/10.1093/nar/gky1085.

Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*. 2016:33:1635. https://doi.org/10.1093/molbev/msw046.

Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J Bacteriol*. 1987:169:5429–5433. https://doi.org/10.1128/jb.169.12.5429-5433.1987.

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*. 2018:9:5114. https://doi.org/10.1038/s41467-018-07641-9.

Johnston AWB *et al*. High frequency transfer of nodulating ability between strains and species of *Rhizobium*. *Nature*. 1978:276:634–636. https://doi.org/10.1038/276634a0.

Kadibalban AS, Landan G, Dagan T. The extent and characteristics of DNA transfer between plasmids and chromosomes. *Curr Biol*. 2024:34:3189–3200.e5. https://doi.org/10.1016/j.cub.2024.06.030.

Kamber T *et al*. 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*. 2012:78:4412–4419. https://doi.org/10.1128/AEM.07351-11.

Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002:30:3059–3066. https://doi.org/10.1093/nar/gkf436.

Kishino H, Hasegawa M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol*. 1989:29:170–179. https://doi.org/10.1007/BF02100115.

Klein JM, Loper JE, Stockwell VO. Influence of endogenous plasmids on phenotypes of *Pantoea vagans* strain C9-1 associated with epiphytic fitness. *J Plant Pathol*. 2017:99:81–89.

Kowalski SP, Lan TH, Feldmann KA, Paterson AH. Comparative mapping of *Arabidopsis thaliana* and *Brassica oleracea* chromosomes reveals islands of conserved organization. *Genetics*. 1994:138:499–510. https://doi.org/10.1093/genetics/138.2.499.

Kumar A *et al*. *Rahnella sikkimica* sp. nov., a novel cold-tolerant bacterium isolated from the glacier of Sikkim Himalaya with plant growth-promoting properties. *Extremophiles*. 2022:26:35. https://doi.org/10.1007/s00792-022-01283-y.

Kuzmin E, Taylor JS, Boone C. Retention of duplicated genes in evolution. *Trends Genet*. 2022:38:59–72. https://doi.org/10.1016/j.tig.2021.06.016.

Lekired A, Cherif-Silini H, Silini A, Ben Yahia H, Ouzari H-I. Comparative genomics reveals the acquisition of mobile genetic elements by the plant growth-promoting Pantoea eucrina OB49 in polluted environments. *Genomics*. 2023:115:110579. https://doi.org/10.1016/j.ygeno.2023.110579.

Li T, Mann R, Kaur J, Spangenberg G, Sawbridge T. Transcriptome analyses of barley roots Inoculated with novel *Paenibacillus* sp. and *Erwinia gerundensis* strains rerveal beneficial early-stage plant–bacteria interactions. *Plants*. 2021:10:1802. https://doi.org/10.3390/plants10091802.

Li Z *et al*. Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell*. 2016:28:326–344. https://doi.org/10.1105/tpc.15.00877.

Lim JY, La HJ, Sheng H, Forney LJ, Hovde CJ. Influence of plasmid pO157 on *Escherichia coli* O157:H7 Sakai biofilm formation. *Appl Environ Microbiol*. 2010:76:963–966. https://doi.org/10.1128/AEM.01068-09.

Mattenberger F, Sabater-Muñoz B, Toft C, Fares MA. The phenotypic plasticity of duplicated genes in Saccharomyces cerevisiae and the origin of adaptations. *G3 (Bethesda)*. 2017:7:63–75. https://doi.org/10.1534/g3.116.035329.

McKay LL, Baldwin KA, Efstathiou JD. Transductional evidence for plasmid linkage of lactose metabolism in streptococcus lactis C2. *Appl Environ Microbiol*. 1976:32:45–52. https://doi.org/10.1128/aem.32.1.45-52.1976.

Meibom KL *et al*. The Vibrio cholerae chitin utilization program. *Proceedings of the National Academy of Sciences*. 2004:101:2524–2529. https://doi.org/10.1073/pnas.0308707101.

Moreno E. Genome evolution within the alpha Proteobacteria: why do some bacteria not possess plasmids and others exhibit more than one different chromosome? *FEMS Microbiol Rev*. 1998:22:255–275. https://doi.org/10.1111/j.1574-6976.1998.tb00370.x.

Müller CA *et al*. The dynamics of genome replication using deep sequencing. *Nucleic Acids Res*. 2014:42:e3. https://doi.org/10.1093/nar/gkt878.

Naito T, Kusano K, Kobayashi I. Selfish behavior of restriction-modification systems. *Science*. 1995:267:897–899. https://doi.org/10.1126/science.7846533.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015:32:268–274. https://doi.org/10.1093/molbev/msu300.

Nuti MP, Lepidi AA, Prakash RK, Schilperoort RA, Cannon FC. Evidence for nitrogen fixation (nif) genes on indigenous

Rhizobium plasmids. *Nature*. 1979:282:533–535. https://doi.org/10.1038/282533a0.
Oakeson KF *et al.* Genome degeneration and adaptation in a nascent stage of symbiosis. *Genome Biol Evol.* 2014:6:76–93. https://doi.org/10.1093/gbe/evt210.
Obeng N, Bansept F, Sieber M, Traulsen A, Schulenburg H. Evolution of microbiota–host associations: the microbe's perspective. *Trends Microbiol.* 2021:29:779–787. https://doi.org/10.1016/j.tim.2021.02.005.
Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. *Nature*. 2000:405:299–304. https://doi.org/10.1038/35012500.
O'Leary NA *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016:44:D733–D745. https://doi.org/10.1093/nar/gkv1189.
Otero-Bravo A, Goffredi S, Sabree ZL. Cladogenesis and genomic streamlining in extracellular endosymbionts of tropical stink bugs. *Genome Biol Evol.* 2018:10:680–693. https://doi.org/10.1093/gbe/evy033.
Pellow D, Mizrahi I, Shamir R. PlasClass improves plasmid sequence classification. *PLoS Comput Biol.* 2020:16:e1007781. https://doi.org/10.1371/journal.pcbi.1007781.
Raetz CRH, Reynolds CM, Trent MS, Bishop RE. Lipid a modification systems in gram-negative bacteria. *Annu Rev Biochem.* 2007:76:295–329. https://doi.org/10.1146/annurev.biochem.76.010307.145803.
Rasmussen T, Jensen RB, Skovgaard O. The two chromosomes of *Vibrio cholerae* are initiated at different time points in the cell cycle. *EMBO J.* 2007:26:3124–3131. https://doi.org/10.1038/sj.emboj.7601747.
Roettger M, Martin W, Dagan T. A machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies. *Mol Biol Evol.* 2009:26:1931–1939. https://doi.org/10.1093/molbev/msp105.
Ruppel S, Merbach W. Effect of ammonium and nitrate on 15N2-fixation of *Azospirillum* spp. and *Pantoea agglomerans* in association with wheat plants. *Microbiol Res.* 1997:152:377–383. https://doi.org/10.1016/S0944-5013(97)80055-9.
Sanz-Puente I *et al.* Seed-mediated vertical transmission of Pantoea core endophytes. *ISME J.* 2025:19:wraf192. https://doi.org/10.1093/ismejo/wraf192.
Saul D *et al.* Nucleotide sequence and replication characteristics of RepFIB, a basic replicon of IncF plasmids. *J Bacteriol.* 1989:171:2697–2707. https://doi.org/10.1128/jb.171.5.2697-2707.1989.
Sergeeva E, Hirkala DLM, Nelson LM. Production of indole-3-acetic acid, aromatic amino acid aminotransferase activities and plant growth promotion by *Pantoea agglomerans* rhizosphere isolates. *Plant Soil*. 2007:297:1–13. https://doi.org/10.1007/s11104-007-9314-5.
Shannon P *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003:13:2498–2504. https://doi.org/10.1101/gr.1239303.
Shin GY *et al.* Plasmids encode and can mobilize onion pathogenicity in *Pantoea agglomerans*. *ISME J*. 2025:19:wraf019. https://doi.org/10.1093/ismejo/wraf019.
Siddavattam D, Khajamohiddin S, Manavathi B, Pakala SB, Merrick M. Transposon-like organization of the plasmid-borne organophosphate degradation (opd) gene cluster found in *Flavobacterium* sp. *Appl Environ Microbiol.* 2003:69:2533–2539. https://doi.org/10.1128/AEM.69.5.2533-2539.2003.
Singh P *et al.* Diazotrophic bacteria *Pantoea dispersa* and Enterobacter asburiae promote sugarcane growth by inducing nitrogen uptake and defense-related gene expression. *Front Microbiol.* 2021:11:600417. https://doi.org/10.3389/fmicb.2020.600417.
Skovgaard O, Bak M, Løbner-Olesen A, Tommerup N. Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. *Genome Res.* 2011:21:1388–1393. https://doi.org/10.1101/gr.117416.110.
Smits THM *et al.* Genomic and phenotypic characterization of a non-pigmented variant of *Pantoea vagans* biocontrol strain C9-1 lacking the 530-kb megaplasmid pPag3. *FEMS Microbiol Lett.* 2010:308:48–54. https://doi.org/10.1111/j.1574-6968.2010.01994.x.
Soberón-Chávez G, Nájera R. Isolation from soil of *Rhizobium leguminosarum* lacking symbiotic information. *Can J Microbiol.* 1989:35:464–468. https://doi.org/10.1139/m89-071.
Soluch R *et al.* Colonization dynamics of *Pantoea agglomerans* in the wheat root habitat. *Environ Microbiol.* 2021:23:2260–2273. https://doi.org/10.1111/1462-2920.15430.
Soutar CD, Stavrinides J. Phylogenomic analysis of the *Erwiniaceae* supports reclassification of *Kalamiella piersonii* to *Pantoea piersonii* comb. nov. and *Erwinia gerundensis* to the new genus *Duffyella* gen. nov. as *Duffyella gerundensis* comb. nov. *Mol Genet Genomics*. 2022:297:213–225. https://doi.org/10.1007/s00438-021-01829-3.
Spahich NA, Vitko NP, Thurlow LR, Temple B, Richardson AR. Staphylococcus aureus lactate- and malate-quinone oxidoreductases contribute to nitric oxide resistance and virulence. *Mol Microbiol.* 2016:100:759–773. https://doi.org/10.1111/mmi.13347.
Stewart FC. A bacterial disease of sweet corn. *N Y Agric Exp Sta Bull.* 1897:130:422–439.
Stice SP *et al.* Thiosulfinate tolerance is a virulence strategy of an atypical bacterial pathogen of onion. *Curr Biol.* 2020:30:3130–3140.e6. https://doi.org/10.1016/j.cub.2020.05.092.
Stockwell VO, Johnson KB, Sugar D, Loper JE. Mechanistically compatible mixtures of bacterial antagonists improve biological control of fire blight of pear. *Phytopathology*®. 2011:101:113–123. https://doi.org/10.1094/PHYTO-03-10-0098.
Sulja A *et al.* Comparative genomics to examine the endophytic potential of *Pantoea agglomerans* DAPP-PG 734. *BMC Genomics* 2022:23:742. https://doi.org/10.1186/s12864-022-08966-y.
Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 2006:34:W609–W612. https://doi.org/10.1093/nar/gkl315.
Thane Papke R, Naor A, Gophna U. Speciation in the shadow of recombination and lateral gene transfer. In: Gophna U, editor. *Lateral gene transfer in evolution*. Springer; 2013. p. 275–289.
Tria FDK, Landan G, Dagan T. Phylogenetic rooting using minimal ancestor deviation. *Nat Ecol Evol.* 2017:1:193. https://doi.org/10.1038/s41559-017-0193.
Tria FDK, Landan G, Picazo DR, Dagan T. Phylogenomic testing of root hypotheses. *Genome Biol Evol.* 2023:15:evad096. https://doi.org/10.1093/gbe/evad096.
Val M-E *et al.* A checkpoint control orchestrates the replication of the two chromosomes of *Vibrio cholerae*. *Sci Adv*. 2016:2:e1501914. https://doi.org/10.1126/sciadv.1501914.
Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. *Curr Opin Microbiol.* 2015:23:148–154. https://doi.org/10.1016/j.mib.2014.11.016.
Vijaymeena MK, Kavitha K. A survey on similarity measures in text mining. *Mach Learn Appl Int J*. 2016:3:19–28. https://doi.org/10.5121/mlaij.2016.3103.
Villa L, García-Fernández A, Fortini D, Carattoli A. Replicon sequence typing of IncF plasmids carrying virulence and resistance determinants. *J Antimicrob Chemother*. 2010:65:2518–2529. https://doi.org/10.1093/jac/dkq347.
Vos M. A species concept for bacteria based on adaptive divergence. *Trends Microbiol.* 2011:19:1–7. https://doi.org/10.1016/j.tim.2010.10.003.
Walterson AM, Stavrinides J. Pantoea: insights into a highly versatile and diverse genus within the Enterobacteriaceae. *FEMS Microbiol Rev*. 2015:39:968–984. https://doi.org/10.1093/femsre/fuv027.

Wang Y, Dagan T. The evolution of antibiotic resistance islands occurs within the framework of plasmid lineages. *Nat Commun*. 2024:15:4555. https://doi.org/10.1038/s41467-024-48352-8.

Wein T, Dagan T. Plasmid evolution. *Curr Biol*. 2020:30:R1158–R1163. https://doi.org/10.1016/j.cub.2020.07.003.

Wiedenbeck J, Cohan FM. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev*. 2011:35:957–976. https://doi.org/10.1111/j.1574-6976.2011.00292.x.

Winkelmann G, Rupp R, Jung G. Herbicolins-new peptide antibiotics from *Erwinia herbicola*. *J Antibiot (Tokyo)*. 1980:33:353–358. https://doi.org/10.7164/antibiotics.33.353.

Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*. 1997:387:708–713. https://doi.org/10.1038/42711.

Xu S *et al*. Fusarium fruiting body microbiome member *Pantoea agglomerans* inhibits fungal pathogenesis by targeting lipid rafts. *Nat Microbiol*. 2022:7:831–843. https://doi.org/10.1038/s41564-022-01131-x.

Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007:24:1586–1591. https://doi.org/10.1093/molbev/msm088.

Zhao X *et al*. Exploring the pathogenic function of *Pantoea ananatis* endogenous plasmid by an efficient and simple plasmid elimination strategy. *Microbiol Res*. 2021:246:126710. https://doi.org/10.1016/j.micres.2021.126710.