nature genetics

ARTICLES

https://doi.org/10.1038/s41588-017-0012-9

Genomic features of bacterial adaptation to plants

Asaf Levy[1], Isai Salas Gonzalez[2,3], Maximilian Mittelviefhaus[4], Scott Clingenpeel[1], Sur Herrera Paredes[2,3,15], Jiamin Miao[5,16], Kunru Wang[5], Giulia Devescovi[6], Kyra Stillman[1], Freddy Monteiro[2,3], Bryan Rangel Alvarez[1], Derek S. Lundberg[2,3], Tse-Yuan Lu[7], Sarah Lebeis[8], Zhao Jin[9], Meredith McDonald[2,3], Andrew P. Klein[2,3], Meghan E. Feltcher[2,3,17], Tijana Glavina Rio[1], Sarah R. Grant[2], Sharon L. Doty[10], Ruth E. Ley[11], Bingyu Zhao[5], Vittorio Venturi[6], Dale A. Pelletier[7], Julia A. Vorholt[4], Susannah G. Tringe[1,12*], Tanja Woyke[1,12*] and Jeffery L. Dangl[2,3,13,14*]

Plants intimately associate with diverse bacteria. Plant-associated bacteria have ostensibly evolved genes that enable them to adapt to plant environments. However, the identities of such genes are mostly unknown, and their functions are poorly characterized. We sequenced 484 genomes of bacterial isolates from roots of Brassicaceae, poplar, and maize. We then compared 3,837 bacterial genomes to identify thousands of plant-associated gene clusters. Genomes of plant-associated bacteria encode more carbohydrate metabolism functions and fewer mobile elements than related non-plant-associated genomes do. We experimentally validated candidates from two sets of plant-associated genes: one involved in plant colonization, and the other serving in microbe–microbe competition between plant-associated bacteria. We also identified 64 plant-associated protein domains that potentially mimic plant domains; some are shared with plant-associated fungi and oomycetes. This work expands the genome-based understanding of plant–microbe interactions and provides potential leads for efficient and sustainable agriculture through microbiome engineering.

The microbiota of plants and animals have coevolved with their hosts for millions of years[1–3]. Through photosynthesis, plants serve as a rich source of carbon for diverse bacterial communities. These include mutualists and commensals, as well as pathogens. Phytopathogens and growth-promoting bacteria have considerable effects on plant growth, health, and productivity[4–7]. Except for intensively studied relationships such as root nodulation in legumes[8], T-DNA transfer by *Agrobacterium*[9], and type III secretion–mediated pathogenesis[10], the molecular mechanisms that govern plant–microbe interactions are not well understood. It is therefore important to identify and characterize the bacterial genes and functions that help microbes thrive in the plant environment. Such knowledge should improve the ability to combat plant diseases and harness beneficial bacterial functions for agriculture, with direct effects on global food security, bioenergy, and carbon sequestration.

Cultivation-independent methods based on profiling of marker genes or shotgun metagenome sequencing have considerably improved the overall understanding of microbial ecology in the plant environment[11–15]. In parallel, reduced sequencing costs have enabled the genome sequencing of plant-associated bacterial isolates at a large scale[16]. Importantly, isolates enable functional validation of in silico predictions. Isolate genomes also provide genomic and evolutionary context for individual genes, as well as the potential to access genomes of rare organisms that might be missed by metagenomics because of limited sequencing depth. Although metagenome sequencing has the advantage of capturing the DNA of uncultivated organisms, multiple 16S rRNA gene surveys have reproducibly shown that the most common plant-associated bacteria are derived mainly from four phyla[13,17] (Proteobacteria, Actinobacteria, Bacteroidetes, and Firmicutes) that are amenable to cultivation. Thus, bacterial cultivation is not a major limitation in sampling of the abundant members of the plant microbiome[16].

Our objective was to characterize the genes that contribute to bacterial adaptation to plants (plant-associated genes) and those genes that specifically aid in bacterial root colonization (root-associated genes). We sequenced the genomes of 484 new bacterial isolates and single bacterial cells from the roots of Brassicaceae, maize, and poplar trees. We combined the newly sequenced genomes with existing genomes to create a dataset of 3,837 high-quality, non-redundant genomes. We then developed a computational approach to identify plant-associated genes and root-associated genes based on comparison of phylogenetically related genomes with knowledge of the origin of isolation. We experimentally validated two sets of plant-associated genes, including a previously unrecognized gene family that functions in plant-associated microbe–microbe competition. In addition, we characterized many plant-associated genes that are shared between bacteria of different phyla, and even between bacteria and plant-associated eukaryotes. This study represents

[1]DOE Joint Genome Institute, Walnut Creek, CA, USA. [2]Department of Biology, University of North Carolina, Chapel Hill, NC, USA. [3]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [4]Institute of Microbiology, ETH Zurich, Zurich, Switzerland. [5]Department of Horticulture, Virginia Tech, Blacksburg, VA, USA. [6]International Centre for Genetic Engineering and Biotechnology, Trieste, Italy. [7]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. [8]Department of Microbiology, University of Tennessee, Knoxville, TN, USA. [9]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA. [10]School of Environmental and Forest Sciences, University of Washington, Seattle, WA, USA. [11]Max Planck Institute for Developmental Biology, Tübingen, Germany. [12]School of Natural Sciences, University of California, Merced, Merced, CA, USA. [13]The Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, NC, USA. [14]Department of Microbiology and Immunology, University of North Carolina, Chapel Hill, NC, USA. Present address: [15]Department of Biology, Stanford University, Stanford, CA, USA; [16]The Grassland College, Gansu Agricultural University, Lanzhou, Gansu, China; [17]BD Technologies and Innovation, Research Triangle Park, NC, USA. Asaf Levy and Isai Salas Gonzalez contributed equally to this work. *e-mail: dangl@email.unc.edu; twoyke@lbl.gov; sgtringe@lbl.gov

a comprehensive and unbiased effort to identify and characterize candidate genes required at the bacteria–plant interface.

Results

Expanding the plant-associated bacterial reference catalog. To obtain a comprehensive reference set of plant-associated bacterial genomes, we isolated and sequenced 191, 135, and 51 novel bacterial strains from the roots of Brassicaceae (91% from *Arabidopsis thaliana*), poplar trees (*Populus trichocarpa* and *Populus deltoides*), and maize, respectively (Methods, Table 1, Supplementary Tables 1–3). The bacteria were specifically isolated from the interior (endophytic compartment) or surface (rhizoplane) of plant roots, or from soil attached to the root (rhizosphere). In addition, we isolated and sequenced 107 single bacterial cells from surface-sterilized roots of *A. thaliana*. All genomes were assembled, annotated, and deposited in public databases and in a dedicated website ("URLs," Supplementary Table 3, Methods).

A broad, high-quality bacterial genome collection. In addition to the newly sequenced genomes noted above, we collected 5,587 bacterial genomes belonging to the four most abundant phyla of plant-associated bacteria[13] from public databases (Methods). We manually classified each genome as plant-associated, non-plant-associated (NPA), or soil-derived on the basis of its unambiguous isolation niche (Methods, Supplementary Tables 1 and 2). The plant-associated genomes included organisms isolated from plants or rhizospheres. A subset of the plant-associated bacteria was also annotated as 'root-associated' when isolated from the rhizoplane or the root endophytic compartment. Genomes from bacteria isolated from soil were considered as a separate group, as it is unknown whether these strains can actively associate with plants. Finally, the remaining genomes were labeled as NPA genomes; these were isolated from diverse sources, including humans, non-human animals, air, sediments, and aquatic environments.

We carried out stringent quality control to remove low-quality or redundant genomes (Methods). This led to a final dataset of 3,837 high-quality and nonredundant genomes, including 1,160 plant-associated genomes, 523 of which were also root-associated. We grouped these 3,837 genomes into nine monophyletic taxa to allow comparative genomics analysis among phylogenetically related genomes (Fig. 1a, Supplementary Tables 1 and 2, Methods, "URLs").

To determine whether our genome collection from cultured isolates was representative of plant-associated bacterial communities, we analyzed cultivation-independent 16S rDNA surveys and metagenomes from the plant environments of *Arabidopsis*[11,12], barley[18], wheat, and cucumber[14] (Methods). The nine taxa analyzed here account for 33–76% (median, 41%; Supplementary Table 4) of the total bacterial communities found in plant-associated environments and therefore represent a substantial portion of the plant microbiota, consistent with previous reports[13,16,19].

Increased carbohydrate metabolism and fewer mobile elements in plant-associated genomes. We compared the genomes of bacteria isolated from plant environments with those from bacteria of shared ancestry that were isolated from non-plant environments. We assumed that the two groups should differ in the set of accessory genes that evolved as part of their adaptation to a specific niche. Comparison of the size of plant-associated, soil, and NPA genomes showed that plant-associated and/or soil genomes were significantly larger than NPA genomes ($P < 0.05$, PhyloGLM and *t*-tests; Supplementary Fig. 1a, Supplementary Table 5). We observed this trend in six to seven of the nine analyzed taxa (depending on the test), representing all four phyla. Pangenome analyses of a few genera with plant-associated and NPA isolation sites showed that pangenome sizes were similar between plant-associated and NPA genomes (Supplementary Fig. 2).

Next, we examined whether certain gene categories are enriched or depleted in plant-associated genomes versus in their NPA counterparts, using 26 broad functional gene categories (Supplementary Table 6). We used the PhyloGLM test (Fig. 1b) and *t*-test (Supplementary Fig. 3) to detect enrichment. Two gene categories demonstrated similar phylogeny-independent trends suggestive of an environment-dependent selection process. The "Carbohydrate metabolism and transport" gene category was expanded in the plant-associated organisms of six taxa (Fig. 1b). This was the most expanded category in Alphaproteobacteria, Bacteroidetes, Xanthomonadaceae, and *Pseudomonas* (Supplementary Fig. 3). In contrast, mobile genetic elements (phages and transposons) were underrepresented in four plant-associated taxa (Fig. 1b and Supplementary Fig. 3). Plant-associated genomes showed increased genome sizes despite a reduction in the number of mobile elements that often serve as vehicles for horizontal gene transfer and genome expansion. A comparison of root-associated bacteria to soil bacteria showed less drastic changes than those seen between plant-associated and NPA groups, as expected for organisms that live in more similar habitats (Fig. 1b and Supplementary Fig. 3).

Identification and validation of plant- and root-associated genes. We sought to identify specific genes enriched in plant- and root-associated genomes compared with NPA and soil-derived

Table 1 | Novel and previously sequenced genomes used in this analysis

Taxon	Taxonomic rank	Novel sequenced PA genomes	Scanned genomes	Genomes used in analysis	PA	NPA	Soil	RA
Alphaproteobacteria[1]	Class	126	784	610	368	199	43	169
Burkholderiales[1]	Order	85	612	433	160	209	64	86
Acinetobacter[1]	Genus	4	926	454	7	442	5	3
Pseudomonas[1]	Genus	75	506	349	169	137	43	61
Xanthomonadaceae[1]	Family	15	264	147	110	26	11	26
Bacillales[2]	Order	54	664	454	97	185	172	54
Actinobacteria 1[3]	NA	69	504	394	164	142	88	89
Actinobacteria 2[3]	NA	19	845	587	29	526	32	18
Bacteroidetes[4]	Phylum	37	481	409	56	293	60	17
Total		484	5,586	3,837	1,160	2,159	518	523

[1]Proteobacteria. [2]Firmicutes. [3]Actinobacteria phylum. PA, plant-associated bacteria; NPA, non-plant-associated bacteria; soil, soil-associated bacteria; RA, root-associated bacteria; NA, not available (an artificial taxon).

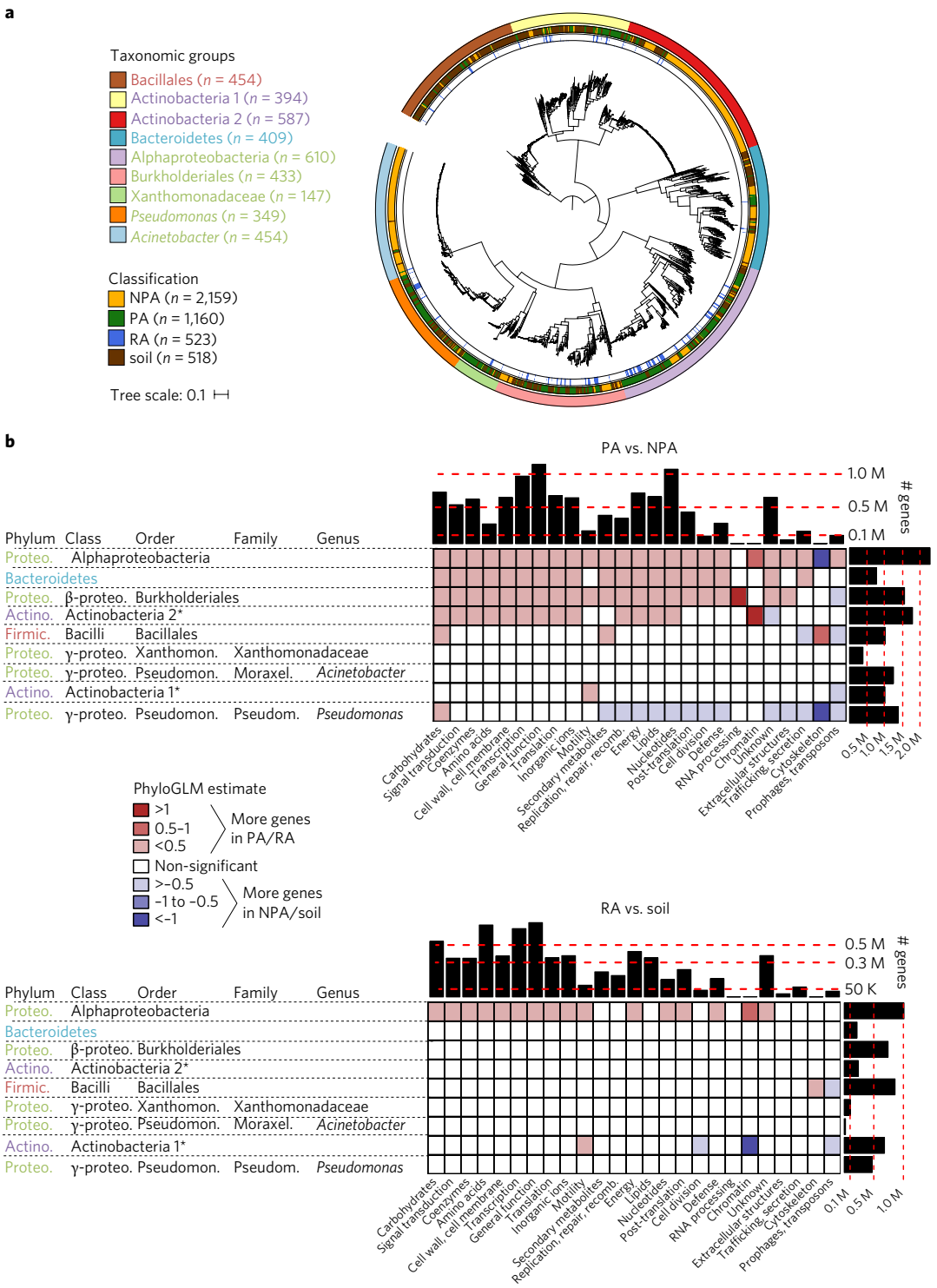

Fig. 1 | The genome dataset used in analysis, and differences in gene category abundances. a, The maximum-likelihood phylogenetic tree of 3,837 high-quality and nonredundant bacterial genomes, based on the concatenated alignment of 31 single-copy genes. The outer ring shows the taxonomic group, the central ring shows the isolation source, and the inner ring shows the root-associated (RA) genomes within plant-associated (PA) genomes. Taxon names are color-coded according to phylum: green, Proteobacteria; red, Firmicutes; blue, Bacteroidetes; purple, Actinobacteria. See "URLs" for the iTOL interactive phylogenetic tree. **b**, Differences in gene categories between plant-associated and NPA genomes (top) and between root-associated and soil-associated genomes (bottom) of the same taxon. Both heat maps indicate the level of enrichment or depletion based on a PhyloGLM test. Significant cells (color-coded according to the key) represent P values of < 0.05 (FDR-corrected). Pink-red cells indicate significantly more genes in plant-associated and root-associated genomes in the top and bottom heat maps, respectively. Histograms at the top and right of each heat map represent the total number of genes compared in each column and row, respectively. Asterisks indicate non-formal class names. "Carbohydrates" denotes the carbohydrate metabolism and transport gene category. Full COG category names for the x-axis labels are presented in Supplementary Table 6. Note that cells representing high absolute estimate values (dark colors) are based on categories of few genes and are therefore more likely to be less accurate. Phylum names are color-coded as in **a**. Xanthomon., Xanthomonadales; Pseudomon., Pseudomonadales; Pseudom., Pseudomonadaceae; Moraxel., Moraxellaceae.

genomes, respectively (Supplementary Fig. 4, Methods). First, we clustered the proteins and/or protein domains of each taxon on the basis of homology, using the annotation resources COG[20], KEGG Orthology[21], and TIGRFAM[22], which typically comprise 35–75% of all genes in bacterial genomes[23]. To capture genes that do not have existing functional annotations, we also used OrthoFinder[24] (after benchmarking; Supplementary Fig. 5) to cluster all protein sequences within each taxon into homology-based orthogroups. Finally, we clustered protein domains with Pfam[25] (Methods, "URLs"). We used these five protein/domain-clustering approaches in parallel comparative genomics pipelines. Each protein/domain sequence was additionally labeled as originating from either a plant-associated genome or an NPA genome.

Next, we determined whether protein/domain clusters were significantly associated with a plant-associated lifestyle by using five independent statistical approaches: hypergbin, hypergcn (two versions of the hypergeometric test), phyloglmbin, phyloglmcn (two phylogenetic tests based on PhyloGLM[26]), and Scoary[27] (a stringent combined test) (Methods). These analyses were based on either gene presence/absence or gene copy number. We defined a gene as significantly plant-associated if at least one test showed that it belonged to a significant plant-associated gene cluster, and if it originated from a plant-associated genome. We defined significant NPA, root-associated, and soil genes in the same way. Significant gene clusters identified by the different methods had varying degrees of overlap (Supplementary Figs. 6 and 7). In general, we noted a high degree of overlap between plant-associated and root-associated genes and overlap between NPA and soil-associated genes (Supplementary Fig. 8). Overall, plant-associated genes were depleted from NPA genomes from heterogeneous isolation sources (Supplementary Figs. 9 and 10). Principal coordinates analysis with matrices that contained only the plant-associated and NPA genes derived from each method as features increased the separation of plant-associated from NPA genomes along the first two axes (Supplementary Fig. 11). We provide full lists of statistically significant plant-associated, root-associated, soil-associated, and NPA proteins and domains according to the five clustering techniques and five statistical approaches for each taxon in Supplementary Tables 7–15 (also see "URLs").

To validate our predictions, we assessed the abundance patterns of plant-associated and root-associated genes in natural environments. We retrieved 38 publicly available plant-associated, NPA, root-associated, and soil-associated shotgun metagenomes, including some from plant-associated environments that were not used for isolation of the bacteria analyzed here[14,28,29] (Supplementary Table 16a). We mapped reads from these culture-independent metagenomes to plant-associated genes found with all statistical approaches (Methods, Supplementary Figs. 12–16). Plant-associated genes in up to seven taxa were more abundant ($P < 0.05$, t-test) in plant-associated metagenomes than in NPA metagenomes (Fig. 2a, Supplementary Table 16b). Root-associated, soil-associated, and NPA genes, in contrast, were not necessarily more abundant in their expected environments (Supplementary Table 16b).

In addition, we selected eight genes that were predicted to be plant-associated by multiple approaches (Supplementary Table 17a) for experimental validation via an *in planta* bacterial fitness assay (Methods). We inoculated the roots of surface-sterilized rice seedlings ($n = 9$–30 seedlings per experiment) with wild-type *Paraburkholderia kururiensis* M130 (a rice endophyte[30]) or a knockout mutant strain for each of the eight genes. We grew the plants for 11 d and then collected and quantified the bacteria that were tightly attached to the roots (Methods, Supplementary Table 17b). Mutations in two genes led to fourfold to sixfold reductions in colonization (false discovery rate (FDR)-corrected Wilcoxon rank sum test, $q < 0.1$) relative to that by wild-type bacteria (Fig. 2b), without an observed effect on growth rate (Supplementary Fig. 17). These two genes encode an outer-membrane efflux transporter from the nodT family and a Tir chaperone protein (CesT), respectively. It is plausible that the other six genes assayed function in facets of plant association not captured in this experimental context.

Functions for which coexpression of and cooperation between different proteins are needed are often encoded by gene operons in bacteria. We therefore tested whether our methods could correctly predict known plant-associated operons. We grouped plant-associated and root-associated genes into putative plant-associated and root-associated operons on the basis of their genomic proximity and orientation (Supplementary Fig. 4, Methods, "URLs"). This analysis yielded some well-known plant-associated functions, such as those of the *nodABCSUIJZ* and *nifHDKENXQ* operons (Fig. 2c,d). Nod and Nif proteins are integral for biological nitrogen cycling and mediate root nodulation[31] and nitrogen fixation[32], respectively. We also identified the biosynthetic gene cluster for the precursor of the plant hormone gibberellin[33,34] (Fig. 2e). Other known plant-associated operons identified are related to chemotaxis[35], secretion systems such as T3SS[36] and T6SS[37], and flagellum biosynthesis[38–40] (Fig. 2f–i).

Thus, we identified thousands of plant-associated and root-associated gene clusters by using five different statistical approaches (Supplementary Table 18) and validated them by means of computational and experimental approaches, broadening our understanding of the genetic basis of plant–microbe interactions and providing a valuable resource to drive further experimentation.

Protein domains reproducibly enriched in diverse plant-associated genomes. Plant-associated and root-associated proteins and protein domains conserved across evolutionarily diverse taxa are potentially pivotal to the interaction between bacteria and plants. We identified 767 Pfam domains as significant plant-associated domains in at least three taxa, on the basis of multiple tests (Supplementary Table 19a). Below we elaborate on a few domains that were plant-associated or root-associated in all four phyla. Two of these domains, a DNA-binding domain (pfam00356) and a ligand-binding (pfam13377) domain, are characteristic of the LacI transcription factor (TF) family. These TFs regulate gene expression in response to different sugars[41], and their copy numbers were expanded in the genomes of plant-associated and root-associated bacteria in eight of the nine taxa analyzed (Fig. 3a). Examination of the genomic neighbors of *lacI*-family genes identified strong enrichment for genes involved in carbohydrate metabolism and transport in all of these taxa, consistent with their expected regulation by a LacI-family protein[41] (Supplementary Fig. 18). We analyzed the promoter regions of these putative regulatory targets of LacI-family TFs, and identified three AANCGNTT palindromic octamers that were statistically enriched in all but one taxon, and which may serve as the TF-binding site (Supplementary Table 20). These data suggest that accumulation of a large repertoire of LacI-family-controlled regulons is a common strategy across bacterial lineages during adaptation to the plant environment.

Another domain, the metabolic domain aldo-keto reductase (pfam00248), was enriched in the genomes of plant-associated and root-associated bacteria from eight taxa belonging to all four phyla investigated (Fig. 3b). This domain is involved in the metabolic conversion of a broad range of substrates, including sugars and toxic carbonyl compounds[42]. Thus, bacteria that inhabit plant environments may consume similar substrates. Additional plant-associated and root-associated proteins and domains that were enriched in at least six taxa are described in Supplementary Fig. 19.

We also identified domains that were reproducibly enriched in NPA and/or soil-associated genomes, including many domains of mobile genetic elements (Supplementary Fig. 20).

Putative plant protein mimicry by plant- and root-associated proteins. Convergent evolution and horizontal transfer of protein domains from eukaryotes to bacteria have been suggested for some

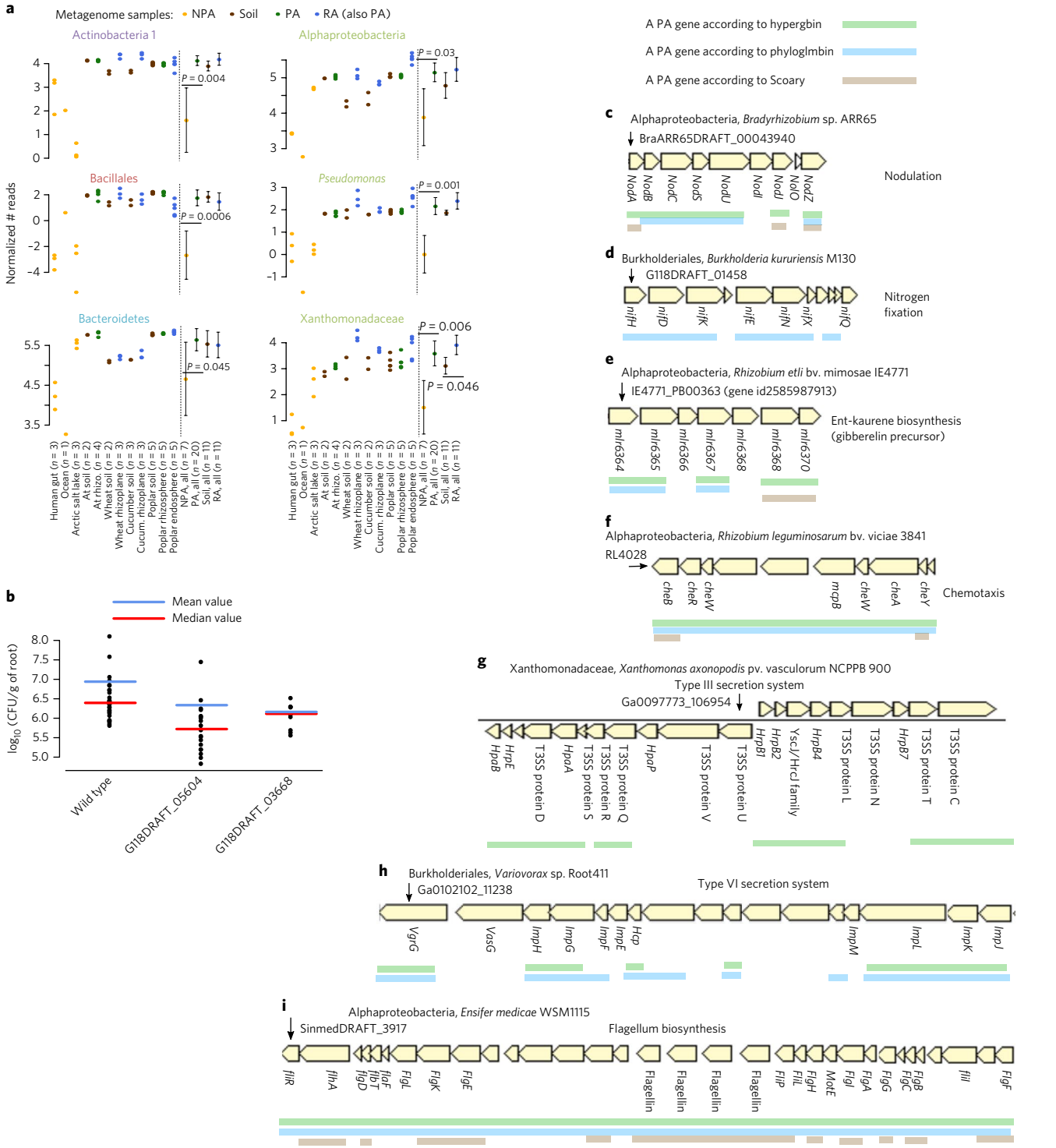

Fig. 2 | Validation of predicted plant-associated genes by multiple approaches. a, Plant-associated (PA) genes, which were predicted from isolate genomes, were more abundant in PA metagenomes than in NPA metagenomes. Reads from 38 shotgun metagenome samples were mapped to significant PA, NPA, RA, and soil-associated genes predicted by Scoary. *P* values are indicated for the significant differences between PA and NPA genes or RA and soil-associated genes in each taxon (two-sided *t*-test). Full results and an explanation for normalization are presented in Supplementary Fig. 14. **b**, Results of a rice root colonization experiment using wild-type *Paraburkholderia kururiensis* M130 or knockout mutants for two predicted plant-associated genes. Two mutants showed reduced colonization compared with the wild type: G118DRAFT_05604 (q-value = 0.00013, Wilcoxon rank sum test), which encodes an outer membrane efflux transporter from the nodT family, and G118DRAFT_03668 (q-value = 0.0952, Wilcoxon rank sum test), a Tir chaperone protein (CesT). Each point represents the average count of a minimum of three to six plates derived from the same plantlet, expressed as colony-forming units (CFU) per gram of root. **c**–**i**, Examples of known functional plant-associated operons captured by different statistical approaches. The plant-associated genes are highlighted by shaded bars, colored according to the key. **c**, *Nod* genes. **d**, *NIF* genes. **e**, Ent-kaurene (gibberelin precursor). **f**, Chemotaxis proteins in bacteria from different taxa. **g**, Type III secretion system. **h**, Type VI secretion system, including the *imp* genes (impaired in nodulation). **i**, Flagellum biosynthesis in Alphaproteobacteria. Labels show the gene symbol or the protein name for which such information was available.

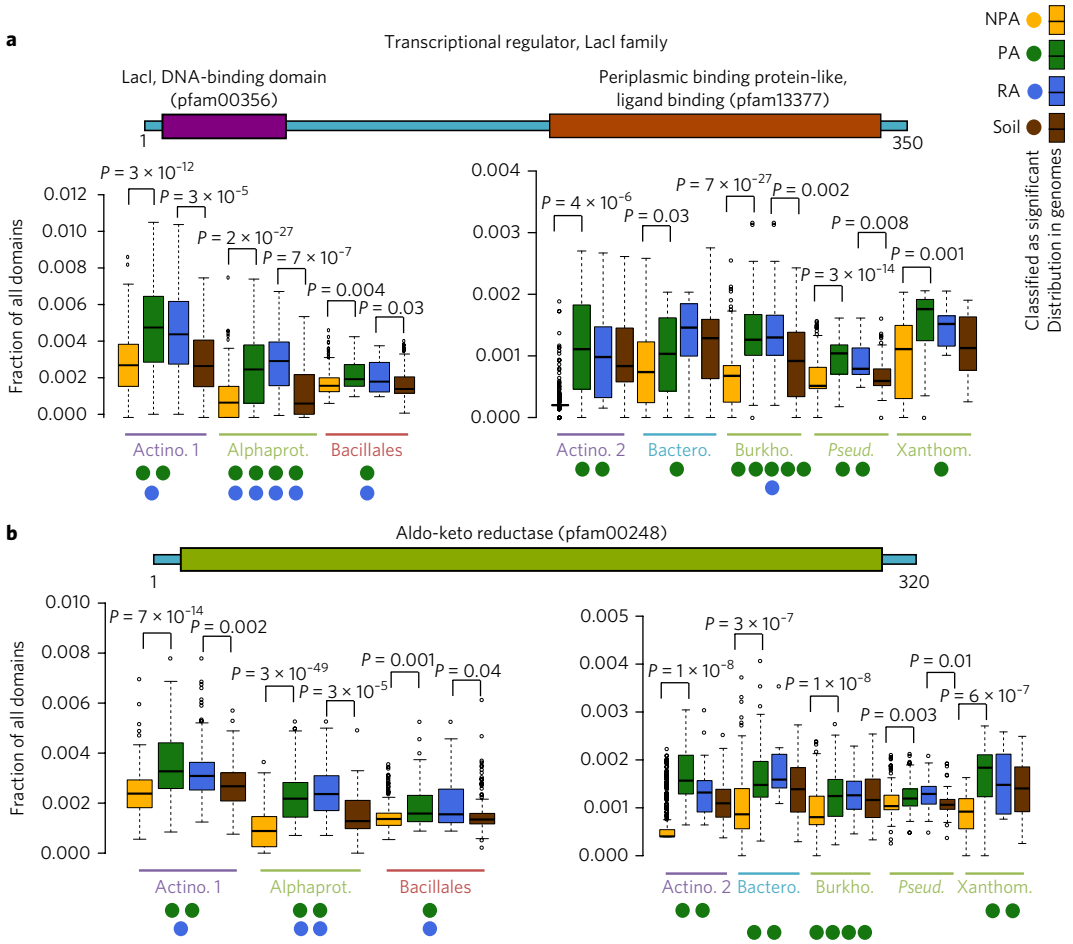

Fig. 3 | Proteins and protein domains that were reproducibly enriched as plant-associated or root-associated in multiple taxa. We compared the occurrence of protein domains (from Pfam) between plant-associated (PA) and NPA bacteria and between root-associated (RA) and soil-associated bacteria. Color-coding is as in Fig. 1a. **a**, Transcription factors with LacI (Pfam00356) and periplasmic-binding protein domains (Pfam13377). These proteins are often annotated as COG1609. **b**, Aldo-keto reductase domain (Pfam00248). Proteins with this domain are often annotated as COG0667. We used a two-sided *t*-test to test for the presence of the genes in **a** and **b** in genomes that shared the same label and to verify the enrichment reported by the various tests. FDR-corrected *P* values are shown for significant results (q-value < 0.05). Colored circles indicate the number of different statistical tests (≤ 5) supporting plant, non-plant, root, or soil association of a gene or domain, with each circle representing one test. Gene illustrations above each graph represent random protein models. Note that **a** and **b** each contain two graphs because of the different scales. Actino., Actinobacteria; Alphaprot., Alphaproteobacteria; Burkho., Burkholderiales; Bactero., Bacteroidetes; *Pseud.*, Pseudomonas; Xanthom., Xanthomonadaceae. Box-and-whisker plots show the median (center lines), 25th and 75th percentiles (box edges), extreme data points within 1.5 times the interquartile range from the box edge (whiskers), and outliers (isolated data points). Full results are in Supplementary Table 19.

microbial effector proteins that are secreted into eukaryotic host cells to suppress defense and facilitate microbial proliferation[43–45]. We searched for new candidate effectors or other functional plant-protein mimics. We retrieved a set of significant plant-associated and root-associated Pfam domains that were reproducibly predicted by multiple approaches or in multiple taxa, and we cross-referenced these with protein domains that were also more abundant in plant genomes than in bacterial genomes (Methods). This analysis yielded 64 plant-resembling plant-associated and root-associated domains (PREPARADOs) encoded by 11,916 genes (Supplementary Fig. 21, Supplementary Table 21). The number of PREPARADOs was four-fold higher than the number of domains that overlapped reproducible NPA/soil-associated domains and plant domains ($n = 15$). The PREPARADOs were relatively abundant in genomes of plant-associated Bacteroidetes and Xanthomonadaceae (>0.5% of all domains on average; Supplementary Fig. 22). Some PREPARADOs were previously described as domains within effector proteins, such as Ankyrin repeats[46], regulator of chromosome condensation repeat (RCC1)[47], leucine-rich repeat (LRR)[48], and pectate lyase[49]. PREPARADOs from plant genomes were enriched 3–14-fold ($P < 10^{-5}$, Fisher's exact test) as domains predicted to be 'integrated effector decoys' when fused to plant intracellular innate immune receptors of the NLR class[50–53] (compared with two random domain sets; Methods, Supplementary Figs. 21 and 23, Supplementary Table 21). We found that 2,201 bacterial proteins that encode 17 of the 64 PREPARADOs shared ≥40% identity across the entire protein sequence with eukaryotic proteins from plants, plant-associated fungi, or plant-associated oomycetes, and therefore are likely to maintain a similar function (Supplementary Fig. 24, Supplementary Tables 21 and 22). The varied phylogenetic distribution among this protein class could have resulted from convergent evolution or from cross-kingdom horizontal gene transfer between phylogenetically distant organisms subjected to the shared selective forces of the plant environment.

Seven PREPARADO-containing protein families were characterized by N-terminal eukaryotic or bacterial signal peptides followed

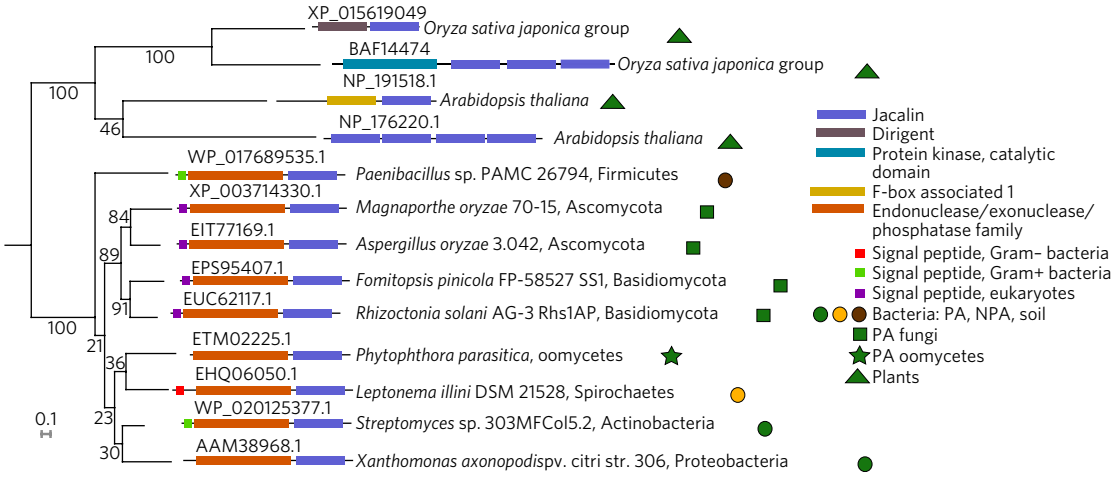

Fig. 4 | A protein family shared by plant-associated bacteria, fungi, and oomycetes that resemble plant proteins. A maximum-likelihood phylogenetic tree of representative proteins with Jacalin-like domains across plants and plant-associated (PA) organisms. Endonuclease/exonuclease/phosphatase-Jacalin proteins are present across PA eukaryotes (fungi and oomycetes) and PA bacteria. In most cases these proteins contain a signal peptide in the N terminus. The Jacalin-like domain is found in many plant proteins, often in multiple copies. The protein accession is shown above each protein illustration.

by a PREPARADO dedicated to carbohydrate binding or metabolism (Supplementary Table 21). One of these domains, Jacalin, is a mannose-binding lectin domain that is found in 48 genes in the *A. thaliana* genome, compared with three genes in the human genome[25]. Mannose is found on the cell wall of different bacterial and fungal pathogens and could serve as a microbial-associated molecular pattern that is recognized by the plant immune system[54–61]. We identified a family of ~430-amino-acid-long microbial proteins with a signal peptide followed by a functionally ill-defined endonuclease/exonuclease/phosphatase family domain (pfam03372), and ending with a Jacalin domain (pfam01419). This domain architecture is absent in plants but is found in diverse microorganisms, many of which are phytopathogens, including Gram-negative and Gram-positive bacteria, fungi from the Ascomycota and Basidiomycota phyla, and oomycetes (Fig. 4). We speculate that these microbial lectins may be secreted to outcompete plant immune receptors for mannose-binding on the microbial cell wall, effectively serving as camouflage.

We thus discovered a large set of protein domains that are shared between plants and the microbes that colonize them. In many cases the entire protein is conserved across evolutionarily distant plant-associated microorganisms.

Co-occurrence of plant-associated gene clusters. We identified numerous cases of plant-associated gene clusters (orthogroups) that demonstrate high co-occurrence between genomes ("URLs"). When the plant-associated genes were derived by phylogeny-aware tests (i.e., PhyloGLM and Scoary), they were candidates for inter-taxon horizontal gene transfer events. For example, we identified a cluster predicted by Scoary of up to 11 co-occurring genes (mean pairwise Spearman correlation: 0.81) in a flagellum-like locus from sporadically distributed plant-associated or soil-associated genomes across 12 different genera in Burkholderiales (Fig. 5). Two of the annotated flagellar-like proteins, FlgB (COG1815) and FliN (pfam01052), are also encoded by plant-associated genes in Actinobacteria 1 and Alphaproteobacteria taxa. Six of the remaining genes encode hypothetical proteins, all but one of which are specific to Betaproteobacteria, suggestive of a flagellar structure variant that evolved in this class in the plant environment. Flagellum-mediated motility or flagellum-derived secretion systems (for example, T3SS) are important for plant colonization and virulence[39,40,62,63] and can be horizontally transferred[64].

Novel putative plant- and root-associated gene operons. In addition to successfully capturing several known plant-associated operons (Fig. 2c–i), we also identified putative plant-associated bacterial operons ("URLs"). Two previously uncharacterized plant-associated gene families were conspicuous. These genes are organized in multiple loci in plant-associated genomes, each with up to five tandem gene copies. They encode short, highly divergent, high-copy-number proteins that are predicted to be secreted, as explained below. These two plant-associated protein families never co-occurred in the same genome, and their genomic presence was perfectly correlated with lifestyles of pathogenic or nonpathogenic bacteria of the genus *Acidovorax* (order Burkholderiales) (Fig. 6a). We named the gene families present in non-pathogens and pathogens *Jekyll* and *Hyde*, respectively, after the characters in Robert Louis Stevenson's classic novel.

The typical *Jekyll* gene is 97 amino acids long, contains an N-terminal signal peptide, lacks a transmembrane domain, and, in 98.5% of cases, appears in non-pathogenic plant-associated or soil-associated *Acidovorax* isolates (Fig. 6a, Supplementary Fig. 25d, Supplementary Table 23a). A single genome may encode up to 13 *Jekyll* gene copies (Fig. 6a) distributed in up to nine loci (Supplementary Table 23a). We recently isolated four *Acidovorax* strains from the leaves of naturally grown *Arabidopsis*[16]. Even these nearly identical isolates carried hypervariable *Jekyll* loci that were substantially more divergent than neighboring genes and included copy-number variations and various mutations (Fig. 6b, Supplementary Fig. 25, Supplementary Table 24).

The *Hyde* putative operons, in contrast, are composed of two distinct gene families unrelated to *Jekyll*. A typical Hyde1 protein has 135 amino acids and an N-terminal transmembrane helix. Hyde1 proteins are also highly variable, as demonstrated by copy-number variation, sequence divergence, and intralocus transposon insertions (Fig. 6a,c, Supplementary Fig. 26a–c, Supplementary Table 23b). *Hyde1* was found in 99% of cases in phytopathogenic *Acidovorax*. These genomes carried up to 15 *Hyde1* gene copies distributed in up to ten loci (Fig. 6a, Supplementary Table 23b). In 70% of cases *Hyde1* was located directly downstream from a more conserved ~300-amino-acid-long plant-associated protein-coding gene that we named *Hyde2* (Fig. 6c,d, Supplementary Table 23d). We identified loci with *Hyde2* followed by *Hyde1*-like genes in different members of the Proteobacteria phylum. These contained a highly variable Hyde1-like protein family that maintained only the

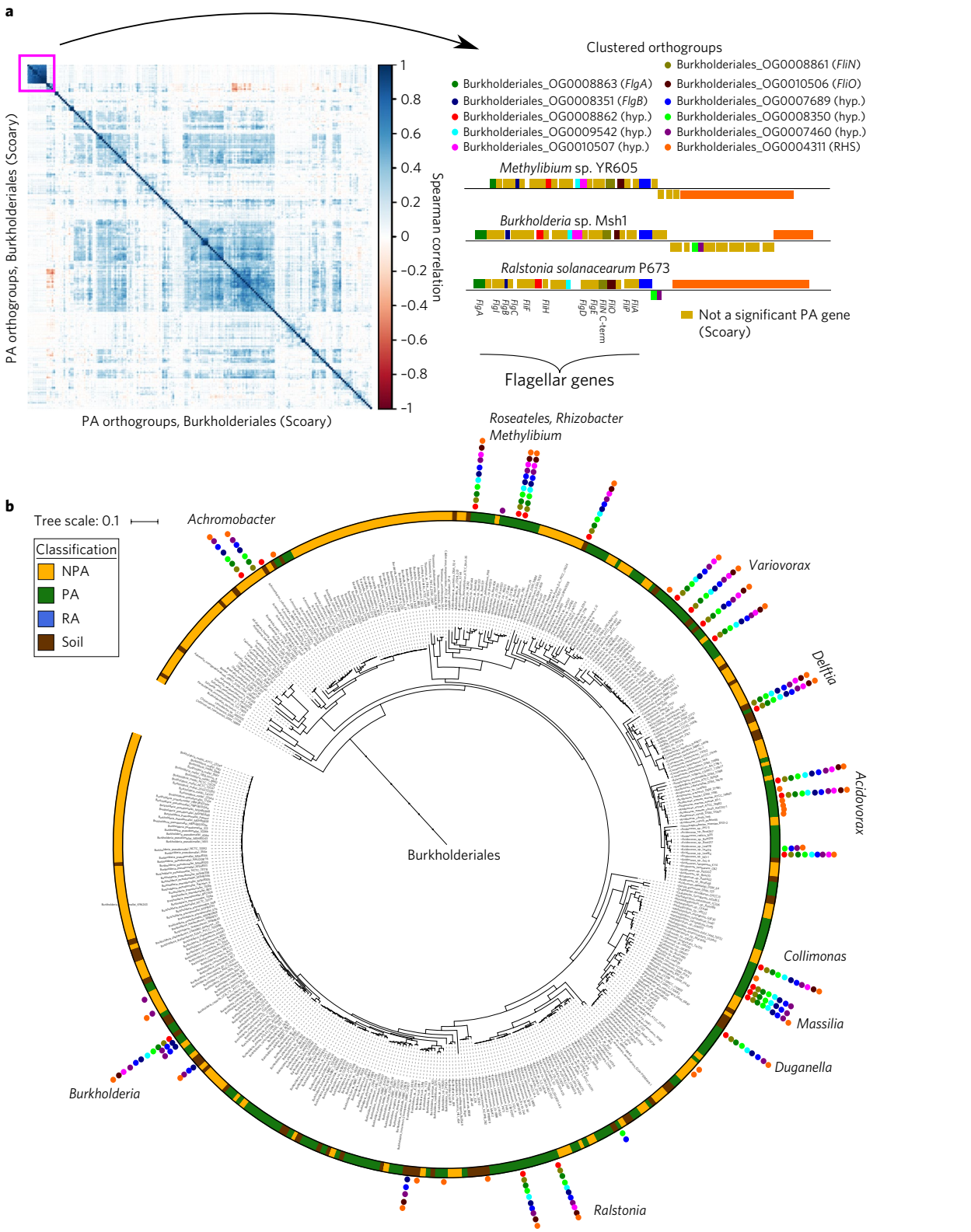

Fig. 5 | Co-occurring plant-associated and soil-associated flagellum-like gene clusters are sporadically distributed across Burkholderiales. a, Left, a hierarchically clustered correlation matrix of all 202 significant plant-associated (PA) orthogroups (gene clusters) from Burkholderiales, predicted by Scoary. Right, the orthogroups present within and adjacent to the flagellar-like locus of different genomes. Gene names based on a BLAST search are shown in parentheses. Hyp., hypothetical protein; RHS, RHS repeat protein. Genes illustrated above and below the black horizontal line for each species are located on the positive and negative strand, respectively. **b**, The Burkholderiales phylogenetic tree based on the concatenated alignment of 31 single-copy genes. Colored circles represent the 11 orthogroups presented in **a**, with the same color-coding as in a. Genus names are shown next to pillars of stacked circles. RA, root-associated.

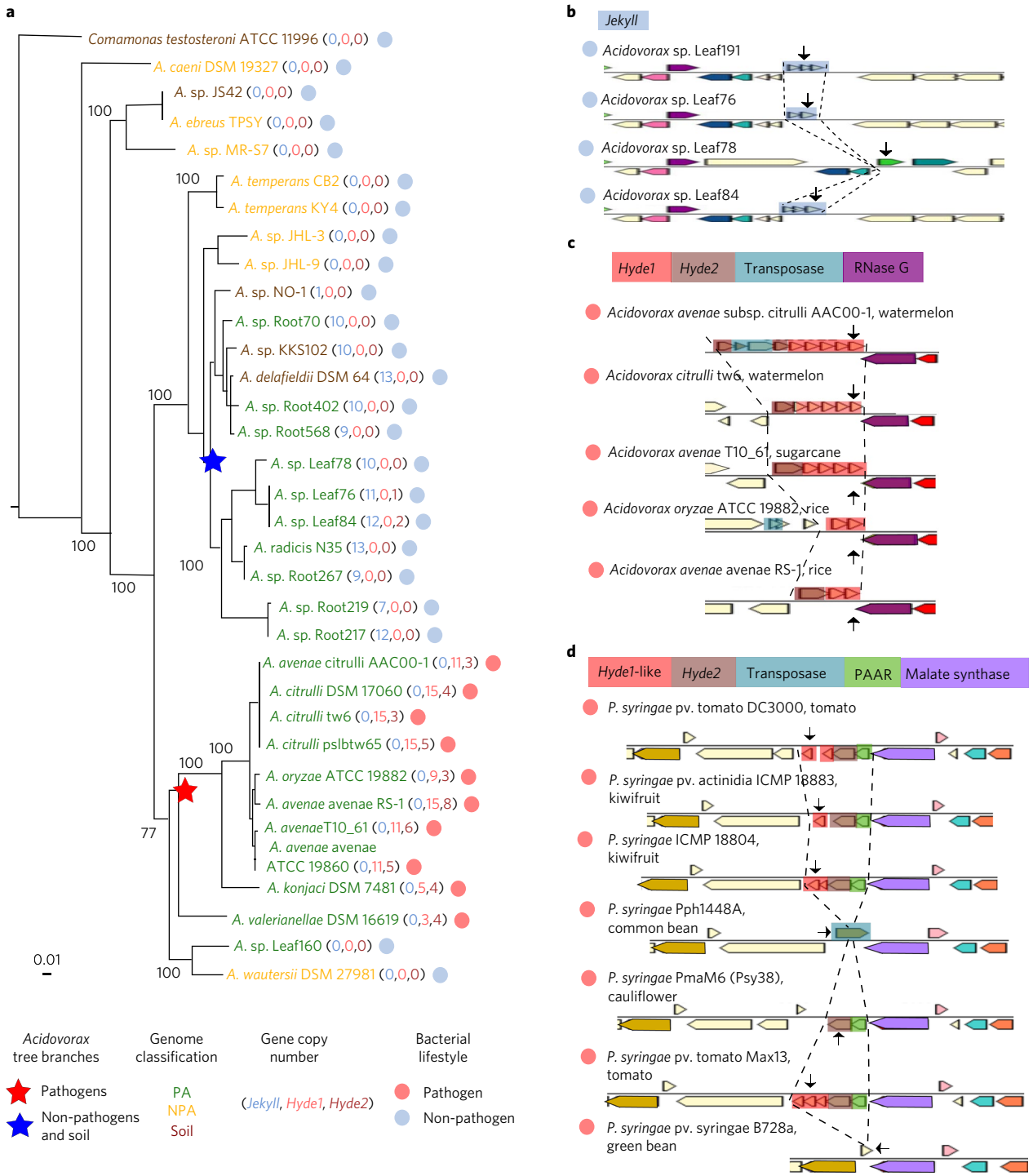

Fig. 6 | Rapidly diversifying, high-copy-number *Jekyll* and *Hyde* plant-associated genes. a, A maximum likelihood phylogenetic tree of *Acidovorax* isolates based on concatenation of 35 single-copy genes. The pathogenic and non-pathogenic branches of the tree are perfectly correlated with the presence of *Hyde1* and *Jekyll* genes, respectively. **b**, An example of a variable *Jekyll* locus in highly related *Acidovorax* species isolated from leaves of wild *Arabidopsis* from Brugg, Switzerland. Arrows indicate the following locus tags (from top to bottom): Ga0102403_10161, Ga0102306_101276, Ga0102307_107159, and Ga0102310_10161. **c**, An example of a variable *Hyde* locus from pathogenic *Acidovorax* infecting different plants (the host plant is shown after the species name). The transposase in the first operon fragmented a *Hyde2* gene. Arrows indicate the following locus tags (from top to bottom): Aave_3195, Ga0078621_123525, Ga0098809_1087148, T336DRAFT_00345, and AASARDRAFT_03920. **d**, An example of a variable *Hyde* locus from pathogenic *Pseudomonas syringae* infecting different plants. Arrows indicate the following locus tags (from top to bottom): PSPTOimg_00004880 (a.k.a. PSPTO_0475), A243_06583, NZ4DRAFT_02530, Pphimg_00049570, PmaM6_0066.00000100, PsyrptM_010100007142, and Psyr_4701. Genes color-coded with the same colors in **b**–**d** are homologous, with the exception of genes colored in ivory (unannotated genes) and *Hyde1* and *Hyde1*-like genes, which are analogous in terms of their similar size, high diversification rate, position downstream of *Hyde2*, and tendency to have a transmembrane domain. PAAR, proline-alanine-alanine-arginine repeat superfamily.

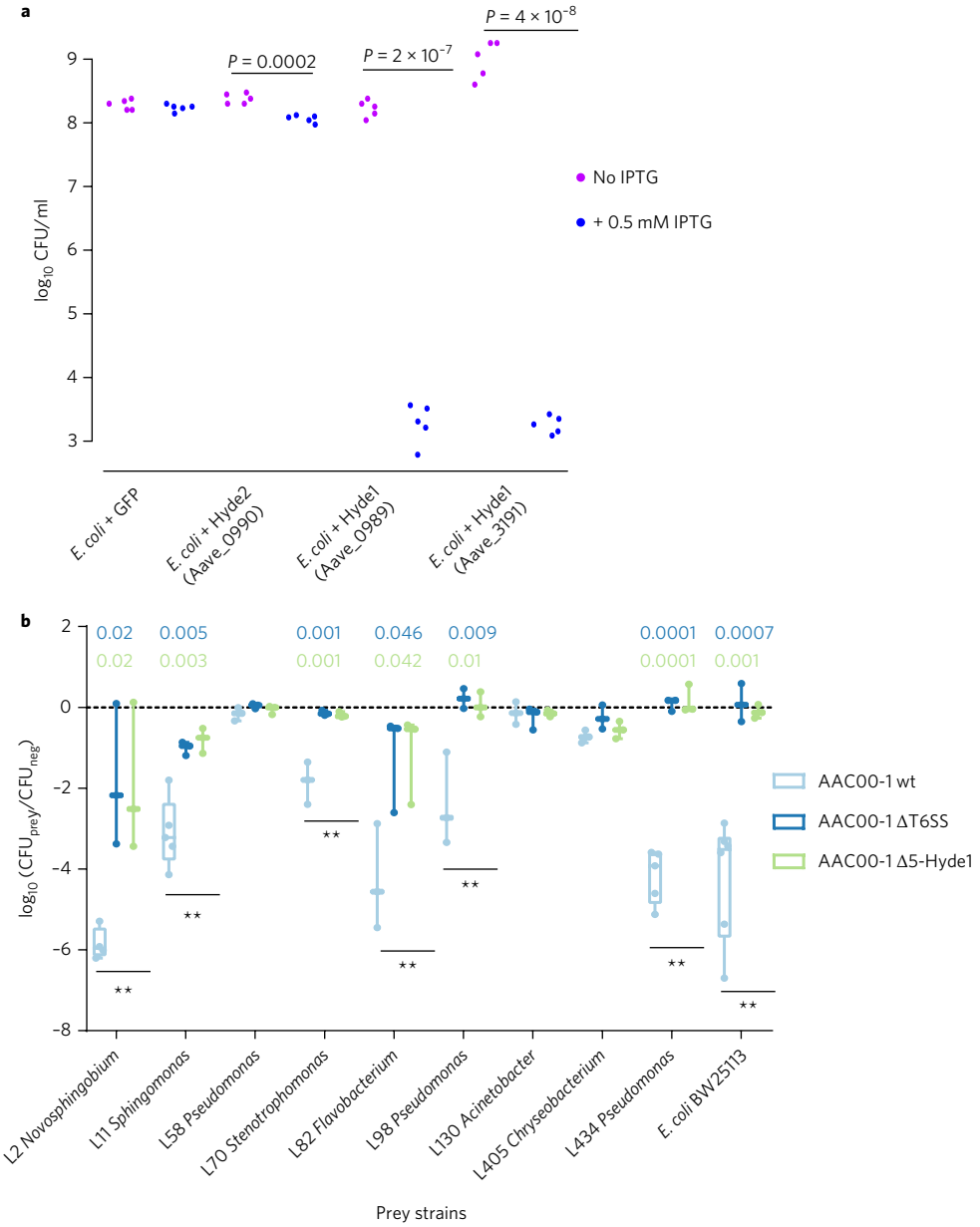

Fig. 7 | Hyde1 proteins of *Acidovorax citrulli* AAC00-1 are toxic to *E. coli* and various plant-associated bacterial strains. a, Toxicity assay of Hyde proteins expressed in *E. coli*. GFP, *Hyde2*-Aave_0990, and two *Hyde1* genes from two loci, Aave_0989 and Aave_3191, were cloned into pET28b and transformed into *E. coli* C41 cells. Aave_0989 and Aave_3191 proteins were 53% identical. Bacterial cultures from five independent colonies were spotted on an LB plate. Gene expression of the cloned genes was induced with 0.5 mM IPTG. *P* values are shown for significant results (two-sided *t*-test). **b**, Quantification of recovered prey cells after coincubation with *Acidovorax* aggressor strains. Antibiotic-resistant prey strains *E. coli* BW25113 and nine different *Arabidopsis* leaf isolates were mixed at equal ratios with different aggressor strains or with NB medium (negative control). Five *Hyde1* loci (including 9 out of 11 *Hyde1* genes) are deleted in Δ5-Hyde1. ΔT6SS contains a *vasD* (Aave_1470) deletion. After coincubation for 19 h on NB agar plates, mixed populations were resuspended in NB medium and spotted on selective antibiotic-containing NB agar. The box plots represent results from at least three independent experiments, with individual values superimposed as dots. The center line represents the median, the box limits represent the 25th and 75th percentiles, and the edges represent the minimal and maximal values. *P* values are shown at the top; double asterisks denote a significant difference (one-way ANOVA followed by Tukey's honest significant difference test) between results for wild type versus ΔT6SS and for wild type versus Δ5-Hyde1. Full strain names and statistical information are presented in Supplementary Table 25. For a time course experiment with exemplary strains, see Supplementary Fig. 29.

short length and a transmembrane helix (Supplementary Fig. 26d). *Hyde*-carrying organisms included other phytopathogens, such as *Pseudomonas syringae*, in which the *Hyde1-like-Hyde2* locus was again highly variable between closely related strains (Fig. 6d, Supplementary Table 23c). However, the striking *Hyde* genomic expansion was specific to the phytopathogenic *Acidovorax* lineage (Supplementary Table 23e). Notably, we observed that *Hyde* genes often are directly preceded by genes that encode core structural T6SS proteins, such as PAAR, VgrG, and Hcp[65], or are fused to PAAR (Fig. 6d, Supplementary Fig. 27a,b, Supplementary Table 23e). We therefore suggest that Hyde1 and/or Hyde2 might constitute a new T6SS effector family.

The high sequence diversity of *Jekyll* and *Hyde1* genes suggests that the two plant-associated protein families encoded by these

genes could be involved in molecular arms races with other organisms in the plant environment. As many type VI effectors are used in interbacterial warfare, we tested *Acidovorax* Hyde1 proteins for antibacterial properties. Expression of two variants of the gene in *Escherichia coli* led to a 10^5–10^6-fold reduction in cell numbers (Fig. 7a, Supplementary Table 25). We constructed a mutant strain of the phytopathogen *Acidovorax citrulli* AAC00-1with deletion of five *Hyde1* loci (Δ5-Hyde1), encompassing 9 of 11 *Hyde1* genes (Supplementary Fig. 28, Supplementary Table 25). Wild-type, Δ5-Hyde1, and T6SS-mutant (ΔT6SS) *Acidovorax* strains were coincubated with an *E. coli* strain that is susceptible to T6SS killing[66] and nine phylogenetically diverse *Arabidopsis* leaf bacterial isolates[16]. Survival of wild-type *E. coli* and six of the leaf isolates after coincubation with wild-type *Acidovorax* was reduced 10^2–10^6-fold compared with that after coincubation with Δ5-Hyde1 or ΔT6SS *Acidovorax* (Fig. 7b, Supplementary Fig. 29, Supplementary Table 25). Combined with the genomic association of *Hyde* loci with T6SS, these results suggest that the T6SS antibacterial phenotype of *Acidovorax* is mediated by Hyde proteins and that these toxins could be used in competition against other plant-associated organisms. Consistent with a function in microbe–microbe interactions, we did not detect compromised virulence of the Δ5-Hyde1 strain on host plants (watermelon; data not shown). However, clearance of competitors via T6SS can promote the persistence of *Acidovorax citrulli* on its host[67].

Discussion

There is increasing awareness that plant-associated microbial communities have important roles in host growth and health. An understanding of plant–microbe relationships at the genomic level could enable scientists to use microbes to enhance agricultural productivity. Most studies have focused on specific plant microbiomes, with more emphasis on microbial diversity than on gene function[12,14,16,18,68–74]. Here we sequenced nearly 500 root-associated bacterial genomes isolated from different plant hosts. These new genomes were combined in a collection of 3,837 high-quality bacterial genomes for comparative analysis. We developed a systematic approach to identify plant-associated and root-associated genes and putative operons. Our method is accurate as reflected by its ability to capture numerous operons previously shown to have a plant-associated function, the enrichment of plant-associated genes in plant-associated metagenomes, the validation of Hyde1 proteins as likely type VI effectors in *Acidovorax* directed against other plant-associated bacteria, and the validation of two new genes in *P. kuruiensis* that affect rice root colonization. We note that bacterial genes that are enriched in genomes from the plant environment are also likely to be involved in adaptation to the many other organisms that share the same niche, as we demonstrated for *Hyde1*.

We used five different statistical approaches to identify genes that were significantly associated with the plant/root environment, each with its advantages and disadvantages. The phylogeny-correcting approaches (phyloglmbin, phyloglmcn, and Scoary) allow accurate identification of genes that are polyphyletic and correlate with an environment independently of ancestral state. On the basis of our metagenome validation, the hypergeometric test predicts more genes that are abundant in plant-associated communities than PhyloGLM does. It also identifies monophyletic plant-associated genes, but it yields more false positives than the phylogenetic tests, because in every plant-associated lineage many lineage-specific genes will be considered plant-associated. Scoary is the most stringent method of all and yielded the fewest predictions (Supplementary Table 18). Future experimental validation should prioritize genes predicted in multiple taxa and/or by multiple approaches (Supplementary Figs. 5 and 6, Supplementary Tables 20 and 26).

We discovered 64 PREPARADOs. Proteins containing 19 of these domains are predicted to be secreted by the Sec or T3SS protein secretion systems (Supplementary Table 21). Notably, plant proteins carrying 35 of these domains belonged to the NLR class of intracellular innate immune receptors (Supplementary Fig. 23, Supplementary Table 21). Thus, these PREPARADO protein domains may serve as molecular mimics. Some may interfere with plant immune functions through disruption of key plant protein interactions[75,76]. Likewise, the Jacalin-containing proteins we detected in plant-associated bacteria, fungi, and oomycetes may represent a strategy of avoiding immunity triggered by microbial-associated molecular patterns, by binding to extracellular microbial mannose molecules and thereby serving as a molecular invisibility cloak[77,78].

Finally, we demonstrated that numerous plant-associated functions are consistent across phylogenetically diverse bacterial taxa, and that some functions are even shared with plant-associated eukaryotes. Some of these traits may facilitate plant colonization by microbes and therefore might prove useful in genome engineering of agricultural inoculants to eventually yield a more efficient and sustainable agriculture.

URLs. iTOL Interactive tree (Fig. 1a), https://itol.embl.de/tree/1522232018227362150877262O; datasets at the Dangl lab's dedicated website, http://labs.bio.unc.edu/Dangl/Resources/gfobap_website/index.html (Dataset 1, FNA—nucleotide FASTA files of the 3,837 genomes; Dataset 2, FAA—FASTA files of all proteins used in the analysis; Dataset 3, COG/KEGG Orthology/Pfam/TIGRFAM IMG annotations of all genes used in analysis; Dataset 4, metadata of all genomes; Dataset 5, phylogenetic trees of each of the nine taxa; Dataset 6, pangenome matrices; Dataset 7, pangeneome data frames; Dataset 8, OrthoFinder orthogroup FASTA files; Dataset 9, Mafft MSA of all orthogroups; Dataset 10, hidden Markov models of all orthogroups; Dataset 11, plant-associated/NPA and root-associated/soil-associated enrichment tables; Dataset 12, correlation matrices; Dataset 13, predicted operons); DSMZ, https://www.dsmz.de/; ATCC, https://www.atcc.org/; NCBI Biosample, https://www.ncbi.nlm.nih.gov/biosample/; IMG, https://img.jgi.doe.gov/cgi-bin/mer/main.cgi; GOLD, https://gold.jgi.doe.gov/; Phytozome, https://phytozome.jgi.doe.gov/pz/portal.html; BrassicaDB, http://brassicadb.org/brad/; sm R package, http://www.stats.gla.ac.uk/~adrian/sm; vegan R package, https://cran.r-project.org/web/packages/vegan/index.html; ape R package, https://cran.r-project.org/web/packages/ape/ape.pdf; fpc R package, https://cran.r-project.org/web/packages/fpc/index.html; phylolmR package, https://cran.r-project.org/web/packages/phylolm/index.html; scripts used to compute the orthogroups, https://github.com/isaisg/gfobap/tree/master/orthofinder_diamond; scripts used to run the gene enrichment tests, https://github.com/isaisg/gfobap/tree/master/enrichment_tests; scripts used to perform the PCoA, https://github.com/isaisg/gfobap/tree/master/pcoa_visualization_ogs_enriched.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-017-0012-9.

Received: 20 January 2017; Accepted: 10 November 2017;
Published online: 18 December 2017

References

1. Ley, R. E. et al. Evolution of mammals and their gut microbes. *Science* **320**, 1647–1651 (2008).
2. Baumann, P. Biology bacteriocyte-associated endosymbionts of plant sap-sucking insects. *Annu. Rev. Microbiol.* **59**, 155–189 (2005).
3. Sprent, J. I. 60Ma of legume nodulation. What's new? What's changing? *J. Exp. Bot.* **59**, 1081–1084 (2008).
4. Pfeilmeier, S., Caly, D. L. & Malone, J. G. Bacterial pathogenesis of plants: future challenges from a microbial perspective: Challenges in Bacterial Molecular Plant Pathology. *Mol. Plant Pathol.* **17**, 1298–1313 (2016).

5. Chowdhury, S. P., Hartmann, A., Gao, X. & Borriss, R. Biocontrol mechanism by root-associated *Bacillus amyloliquefaciens* FZB42—a review. *Front. Microbiol.* **6**, 780 (2015).
6. Fibach-Paldi, S., Burdman, S. & Okon, Y. Key physiological properties contributing to rhizosphere adaptation and plant growth promotion abilities of *Azospirillum brasilense*. *FEMS Microbiol. Lett.* **326**, 99–108 (2012).
7. Santhanam, R. et al. Native root-associated bacteria rescue a plant from a sudden-wilt disease that emerged during continuous cropping. *Proc. Natl. Acad. Sci. USA* **112**, E5013–E5020 (2015).
8. Peters, N. K., Frost, J. W. & Long, S. R. A plant flavone, luteolin, induces expression of *Rhizobium meliloti* nodulation genes. *Science* **233**, 977–980 (1986).
9. Hiei, Y., Ohta, S., Komari, T. & Kumashiro, T. Efficient transformation of rice (*Oryza sativa L.*) mediated by Agrobacterium and sequence analysis of the boundaries of the T-DNA. *Plant J.* **6**, 271–282 (1994).
10. Hueck, C. J. Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiol. Mol. Biol. Rev.* **62**, 379–433 (1998).
11. Bulgarelli, D. et al. Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. *Nature* **488**, 91–95 (2012).
12. Lundberg, D. S. et al. Defining the core *Arabidopsis thaliana* root microbiome. *Nature* **488**, 86–90 (2012).
13. Bulgarelli, D., Schlaeppi, K., Spaepen, S., Ver Loren van Themaat, E. & Schulze-Lefert, P. Structure and functions of the bacterial microbiota of plants. *Annu. Rev. Plant Biol.* **64**, 807–838 (2013).
14. Ofek-Lalzar, M. et al. Niche and host-associated functional signatures of the root surface microbiome. *Nat. Commun.* **5**, 4950 (2014).
15. Gottel, N. R. et al. Distinct microbial communities within the endosphere and rhizosphere of *Populus deltoides* roots across contrasting soil types. *Appl. Environ. Microbiol.* **77**, 5934–5944 (2011).
16. Bai, Y. et al. Functional overlap of the *Arabidopsis* leaf and root microbiota. *Nature* **528**, 364–369 (2015).
17. Hardoim, P. R. et al. The hidden world within plants: ecological and evolutionary considerations for defining functioning of microbial endophytes. *Microbiol. Mol. Biol. Rev.* **79**, 293–320 (2015).
18. Bulgarelli, D. et al. Structure and function of the bacterial root microbiota in wild and domesticated barley. *Cell Host Microbe* **17**, 392–403 (2015).
19. Hacquard, S. et al. Microbiota and host nutrition across plant and animal kingdoms. *Cell Host Microbe* **17**, 603–616 (2015).
20. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
21. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
22. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).
23. Huntemann, M. et al. The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4). *Stand. Genomic Sci.* **10**, 86 (2015).
24. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
25. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).
26. Ives, A. R. & Garland, T. Jr. Phylogenetic logistic regression for binary dependent variables. *Syst. Biol.* **59**, 9–26 (2010).
27. Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* **17**, 238 (2016).
28. Hultman, J. et al. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature* **521**, 208–212 (2015).
29. Louca, S. et al. Integrating biogeochemistry with multiomic sequence information in a model oxygen minimum zone. *Proc. Natl. Acad. Sci. USA* **113**, E5925–E5933 (2016).
30. Coutinho, B. G., Licastro, D., Mendonça-Previato, L., Cámara, M. & Venturi, V. Plant-influenced gene expression in the rice endophyte *Burkholderia kururiensis* M130. *Mol. Plant Microbe Interact.* **28**, 10–21 (2015).
31. Long, S. R. Rhizobium-legume nodulation: life together in the underground. *Cell* **56**, 203–214 (1989).
32. Ruvkun, G. B., Sundaresan, V. & Ausubel, F. M. Directed transposon Tn5 mutagenesis and complementation analysis of *Rhizobium meliloti* symbiotic nitrogen fixation genes. *Cell* **29**, 551–559 (1982).
33. Hershey, D. M., Lu, X., Zi, J. & Peters, R. J. Functional conservation of the capacity for ent-kaurene biosynthesis and an associated operon in certain rhizobia. *J. Bacteriol.* **196**, 100–106 (2014).
34. Nett, R. S. et al. Elucidation of gibberellin biosynthesis in bacteria reveals convergent evolution. *Nat. Chem. Biol.* **13**, 69–74 (2017).
35. Scharf, B. E., Hynes, M. F. & Alexandre, G. M. Chemotaxis signaling systems in model beneficial plant-bacteria associations. *Plant Mol. Biol.* **90**, 549–559 (2016).
36. Büttner, D. & He, S. Y. Type III protein secretion in plant pathogenic bacteria. *Plant Physiol.* **150**, 1656–1664 (2009).
37. Gao, R. et al. Genome-wide RNA sequencing analysis of quorum sensing-controlled regulons in the plant-associated *Burkholderia glumae* PG1 strain. *Appl. Environ. Microbiol.* **81**, 7993–8007 (2015).
38. Weller-Stuart, T., Toth, I., De Maayer, P. & Coutinho, T. Swimming and twitching motility are essential for attachment and virulence of *Pantoea ananatis* in onion seedlings. *Mol. Plant Pathol.* **18**, 734–745 (2017).
39. De Weger, L. A. et al. Flagella of a plant-growth-stimulating *Pseudomonas fluorescens* strain are required for colonization of potato roots. *J. Bacteriol.* **169**, 2769–2773 (1987).
40. de Weert, S. et al. Flagella-driven chemotaxis towards exudate components is an important trait for tomato root colonization by *Pseudomonas fluorescens*. *Mol. Plant Microbe Interact.* **15**, 1173–1180 (2002).
41. Ravcheev, D. A. et al. Comparative genomics and evolution of regulons of the LacI-family transcription factors. *Front. Microbiol.* **5**, 294 (2014).
42. Yamauchi, Y., Hasegawa, A., Taninaka, A., Mizutani, M. & Sugimoto, Y. NADPH-dependent reductases involved in the detoxification of reactive carbonyls in plants. *J. Biol. Chem.* **286**, 6999–7009 (2011).
43. Burstein, D. et al. Genome-scale identification of *Legionella pneumophila* effectors using a machine learning approach. *PLoS Pathog.* **5**, e1000508 (2009).
44. Dean, P. Functional domains and motifs of bacterial type III effector proteins and their roles in infection. *FEMS Microbiol. Rev.* **35**, 1100–1125 (2011).
45. Stebbins, C. E. & Galán, J. E. Structural mimicry in bacterial virulence. *Nature* **412**, 701–705 (2001).
46. Price, C. T. et al. Molecular mimicry by an F-box effector of *Legionella pneumophila* hijacks a conserved polyubiquitination machinery within macrophages and protozoa. *PLoS Pathog.* **5**, e1000704 (2009).
47. Rothmeier, E. et al. Activation of Ran GTPase by a *Legionella* effector promotes microtubule polymerization, pathogen vacuole motility and infection. *PLoS Pathog.* **9**, e1003598 (2013).
48. Xu, R.-Q. et al. AvrAC(Xcc8004), a type III effector with a leucine-rich repeat domain from *Xanthomonas campestris* pathovar campestris confers avirulence in vascular tissues of *Arabidopsis thaliana* ecotype Col-0. *J. Bacteriol.* **190**, 343–355 (2008).
49. Shevchik, V. E., Robert-Baudouy, J. & Hugouvieux-Cotte-Pattat, N. Pectate lyase PelI of *Erwinia chrysanthemi* 3937 belongs to a new family. *J. Bacteriol.* **179**, 7321–7330 (1997).
50. Cesari, S., Bernoux, M., Moncuquet, P., Kroj, T. & Dodds, P. N. A novel conserved mechanism for plant NLR protein pairs: the "integrated decoy" hypothesis. *Front. Plant Sci.* **5**, 606 (2014).
51. Sarris, P. F. et al. A plant immune receptor detects pathogen effectors that target WRKY transcription factors. *Cell* **161**, 1089–1100 (2015).
52. Sarris, P. F., Cevik, V., Dagdas, G., Jones, J. D. & Krasileva, K. V. Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. *BMC Biol.* **14**, 8 (2016).
53. Le Roux, C. et al. A receptor pair with an integrated decoy converts pathogen disabling of transcription factors to immunity. *Cell* **161**, 1074–1088 (2015).
54. Brown, G. D. & Netea, M. G. (eds.). *Immunology of Fungal Infections*. (Springer, Dordrecht, The Netherlands, 2007).
55. Gadjeva, M., Takahashi, K. & Thiel, S. Mannan-binding lectin—a soluble pattern recognition molecule. *Mol. Immunol.* **41**, 113–121 (2004).
56. Ma, Q.-H., Tian, B. & Li, Y.-L. Overexpression of a wheat jasmonate-regulated lectin increases pathogen resistance. *Biochimie* **92**, 187–193 (2010).
57. Xiang, Y. et al. A jacalin-related lectin-like gene in wheat is a component of the plant defence system. *J. Exp. Bot.* **62**, 5471–5483 (2011).
58. Yamaji, Y. et al. Lectin-mediated resistance impairs plant virus infection at the cellular level. *Plant Cell* **24**, 778–793 (2012).
59. Weidenbach, D. et al. Polarized defense against fungal pathogens is mediated by the Jacalin-related lectin domain of modular *Poaceae*-specific proteins. *Mol. Plant* **9**, 514–527 (2016).
60. Sahly, H. et al. Surfactant protein D binds selectively to *Klebsiella pneumoniae* lipopolysaccharides containing mannose-rich O-antigens. *J. Immunol.* **169**, 3267–3274 (2002).
61. Osborn, M. J., Rosen, S. M., Rothfield, L., Zeleznick, L. D. & Horecker, B. L. Lipopolysaccharide of the gram-negative cell wall. *Science* **145**, 783–789 (1964).
62. Tans-Kersten, J., Huang, H. & Allen, C. *Ralstonia solanacearum* needs motility for invasive virulence on tomato. *J. Bacteriol.* **183**, 3597–3605 (2001).
63. Cole, B. J. et al. Genome-wide identification of bacterial plant colonization genes. *PLoS Biol.* **15**, e2002860 (2017).
64. Poggio, S. et al. A complete set of flagellar genes acquired by horizontal transfer coexists with the endogenous flagellar system in *Rhodobacter sphaeroides*. *J. Bacteriol.* **189**, 3208–3216 (2007).

65. Ho, B. T., Dong, T. G. & Mekalanos, J. J. A view to a kill: the bacterial type VI secretion system. *Cell Host Microbe* **15**, 9–21 (2014).
66. MacIntyre, D. L., Miyata, S. T., Kitaoka, M. & Pukatzki, S. The *Vibrio cholerae* type VI secretion system displays antimicrobial properties. *Proc. Natl. Acad. Sci. USA* **107**, 19520–19524 (2010).
67. Tian, Y. et al. The type VI protein secretion system contributes to biofilm formation and seed-to-seedling transmission of *Acidovorax citrulli* on melon. *Mol. Plant Pathol.* **16**, 38–47 (2015).
68. Peiffer, J. A. et al. Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proc. Natl. Acad. Sci. USA* **110**, 6548–6553 (2013).
69. Agler, M. T. et al. Microbial hub taxa link host and abiotic factors to plant microbiome variation. *PLoS Biol.* **14**, e1002352 (2016).
70. Bokulich, N. A., Thorngate, J. H., Richardson, P. M. & Mills, D. A. Microbial biogeography of wine grapes is conditioned by cultivar, vintage, and climate. *Proc. Natl. Acad. Sci. USA* **111**, E139–E148 (2014).
71. Coleman-Derr, D. et al. Plant compartment and biogeography affect microbiome composition in cultivated and native *Agave* species. *New Phytol.* **209**, 798–811 (2016).
72. Shade, A., McManus, P. S. & Handelsman, J. Unexpected diversity during community succession in the apple flower microbiome. *MBio* **4**, e00602–e00612 (2013).
73. Turner, T. R. et al. Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere microbiome of plants. *ISME J.* **7**, 2248–2258 (2013).
74. Edwards, J. et al. Structure, variation, and assembly of the root-associated microbiomes of rice. *Proc. Natl. Acad. Sci. USA* **112**, E911–E920 (2015).
75. Kroj, T., Chanclud, E., Michel-Romiti, C., Grand, X. & Morel, J.-B. Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread. *New Phytol.* **210**, 618–626 (2016).
76. Mukhtar, M. S. et al. Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science* **333**, 596–601 (2011).
77. Vimr, E. & Lichtensteiger, C. To sialylate, or not to sialylate: that is the question. *Trends Microbiol.* **10**, 254–257 (2002).
78. de Jonge, R. et al. Conserved fungal LysM efector Ecp6 prevents chitin-triggered immunity in plants. *Science* **329**, 953–955 (2010).

Acknowledgements

The work conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the US Department of Energy under contract no. DE-AC02-05CH11231. J.L.D. and S.G.T. were supported by NSF INSPIRE grant IOS-1343020, and J.L.D. was also supported by DOE–USDA Feedstock Award DE-SC001043 and by the Office of Science (BER), US Department of Energy, grant no. DE-SC0014395. S.H.P. was supported by NIH Training Grant T32 GM067553-06 and was a Howard Hughes Medical Institute (HHMI) International Student Research Fellow. D.S.L. was supported by NIH Training Grant T32 GM07092-34. J.L.D. is an Investigator of the HHMI, supported by the HHMI and the Gordon and Betty Moore Foundation (GBMF3030). M.E.F. was supported by NIH Dr. Ruth L. Kirschstein NRSA Fellowship F32-GM112345. D.A.P. and T.-Y.L. were supported by the Genomic Science Program, US Department of Energy, Office of Science, Biological and Environmental Research as part of the Oak Ridge National Laboratory Plant Microbe Interfaces Scientific Focus Area (http://pmi.ornl.gov) and Plant Feedstock Genomics Award DE-SC001043. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the US Department of Energy under contract DE-AC05-00OR22725. J.A.V. was supported by a SystemsX.ch grant (Micro2X) and a European Research Council (ERC) advanced grant (PhyMo). We thank I. Bertani, C. Bez, R. Bowers, D. Burstein, A. Chun Chen, D. Chiniquy, B. Cole, O. Cohen, A. Copeland, J. Eisen, E. Eloe-Fadrosh, M. Hadjithomas, O. Finkel, H. Schnitzel Meule Fux, N. Ivanova, J. Knelman, R. Malmstrom, R. Perez-Torres, D. Salomon, R. Sorek, T. Mucyn, R. Seshadri, T.K. Reddy, L. Ryan, and H. Sberro Livnat for general help, text editing, and ideas for this work. We thank R. Walcott (University of Georgia, Athens, GA, USA) for providing the *Acidovorax citrulli* VasD mutant strain.

Author contributions

A.L. performed most data analysis and wrote the paper. I.S.G. performed phylogenetic inference, performed phylogenetically aware analyses, analyzed the data, provided the supporting website, and contributed to manuscript writing. M. Mittelviefhaus and J.A.V. designed and performed experiments related to *Hyde1* gene function and contributed to manuscript writing. S.C. isolated single bacterial cells and prepared metadata for data analysis. F.M. analyzed data. S.H.P. analyzed data and contributed to manuscript writing. J.M. produced a mutant strain for Hyde1. K.W. tested Hyde1 toxicity in *E. coli*. G.D. and V.V. produced deletion mutants and designed and performed rice root colonization experiments. K.S. helped in data analysis. B.R.A. prepared metadata for data analysis. D.S.L., T.-Y.L., S.L., Z.J., M. McDonald, A.P.K., M.E.F., and S.L.D. isolated bacteria from different plants or managed this process. T.G.d.R. managed the sequencing project. S.R.G., D.A.P., and R.E.L. managed bacterial isolation efforts and contributed to manuscript writing. B.Z. managed *Hyde1* deletion and toxicity testing. S.G.T. contributed to manuscript writing. T.W. managed single-cell isolation efforts and contributed to manuscript writing. J.L.D. directed the overall project and contributed to manuscript writing.

Competing interests

J.L.D. is a cofounder of and shareholder in, and S.H.P. collaborates with, AgBiome LLC, a corporation that aims to use plant-associated microbes to improve plant productivity.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-017-0012-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.G.T. or T.W. or J.L.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Additional method descriptions appear in Supplementary Note 1.

Bacterial isolation and genome sequencing. The detailed isolation procedure is described in Supplementary Note 1. Bacterial strains from Brassicaceae and poplar were isolated via previously described protocols[79,80]. Poplar strains were cultured from root tissues collected from *Populus deltoides* and *Populus trichocarpa* trees in Tennessee, North Carolina, and Oregon. Root samples were processed as described previously[15,80]. Briefly, we isolated rhizosphere strains by plating serial dilutions of root wash, whereas for endosphere strains, we pulverized surface-sterilized roots with a sterile mortar and pestle in 10 mL of $MgSO_4$ (10 mM) solution before plating serial dilutions. Strains were isolated on R2A agar media, and the resulting colonies were picked and re-streaked a minimum of three times to ensure isolation. Isolated strains were identified by 16S rDNA PCR followed by Sanger sequencing.

For maize isolates, we selected soils associated with Il14h and Mo17 maize genotypes grown in Lansing, NY, and Urbana, IL. The rhizosphere soil samples of each maize genotype were grown at each location and were collected at week 12 as previously described[68]. From each rhizosphere soil sample, soil was washed and samples were plated onto *Pseudomonas* isolation agar (BD Diagnostic Systems). The plates were incubated at 30 °C until colonies formed, and DNA was extracted from cells.

For isolation of single cells, *A. thaliana* accessions Col-0 and Cvi-0 were grown to maturity. Roots were washed in distilled water multiple times. Root surfaces were sterilized with bleach. Surface-sterilized roots were then ground with a sterile mortar and pestle. Individual cells were isolated by flow cytometry followed by DNA amplification with MDA, and 16S rDNA screening as described previously[81].

DNA from isolates and single cells was sequenced on next-generation sequencing platforms, mostly using Illumina HiSeq technology (Supplementary Table 3). Sequenced genomic DNA was assembled via different assembly methods (Supplementary Table 3). Genomes were annotated using the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4)[23] and deposited at the IMG database ("URLs"), ENA, or Genbank for public use.

Data compilation of 3,837 isolate genomes and their isolation-site metadata. We retrieved 5,586 bacterial genomes from the IMG system ("URLs," Supplementary Table 1). Isolation sites were identified through a manual curation process that included scanning of IMG metadata, DSMZ, ATCC, NCBI Biosample ("URLs"), and the scientific literature. On the basis of its isolation site, each genome was labeled as plant-associated, NPA, or soil-associated. Plant-associated organisms were also labeled as root-associated when isolated from the endophytic compartments or from the rhizoplane. We applied stringent quality control measures to ensure a high-quality and minimally biased set of genomes:

a. Known isolation site: genomes with missing isolation-site information were filtered out.
b. High genome quality and completeness: all isolate genomes passed this filter if N_{50} (the shortest sequence length at 50% of the genome) was more than 50,000 bp. Single amplified genomes passed the quality filter if they had at least 90% of 35 universal single-copy clusters of orthologous groups (COGs)[82]. In addition, we used CheckM[83] to assess isolate genome completeness and contamination. Only genomes that were at least 95% complete and no more than 5% contaminated were used.
c. High-quality gene annotation: genomes that passed this filter had at least 90% genome sequence coding for genes, with an exception—in the *Bartonella* genus most genomes have coding base percentages below 90%.
d. Nonredundancy: we computed whole-genome average nucleotide identity and alignment fraction values for each pair of genomes[84]. When the alignment fraction exceeded 90% and the whole-genome average nucleotide identity was greater than 99.995% we considered the genome pair redundant. In such cases one genome was randomly selected, and the other genome was marked as redundant and was filtered out.
e. Consistency in the phylogenetic tree: we filtered out 14 bacterial genomes that showed discrepancy between their given taxonomy and their actual phylogenetic placement in the bacterial tree.

Construction of the bacterial genome tree. To generate a phylogenetic tree of the 3,837 high-quality and nonredundant bacterial genomes, we retrieved 31 universal single-copy genes from each genome with AMPHORA2[85]. For each individual marker gene, we used Muscle with default parameters to construct an alignment. We masked the 31 alignments by using Zorro[86] and filtered the low-quality columns of the alignment. Finally, we concatenated the 31 alignments into an overall merged alignment, from which we built an approximately maximum-likelihood phylogenetic tree with the WAG model implemented in FastTree 2.1[87]. Trees of each taxon are provided in Dataset S5 at http://labs.bio.unc.edu/Dangl/Resources/gfobap_website/faa_trees_metadata.html.

Clustering of 3,837 genomes into nine taxa. We divided the dataset into different taxa (taxonomic groups) to allow downstream identification of genes enriched in the plant-associated or root-associated genomes of each taxon compared with the NPA or soil-associated genomes from the same taxon, respectively. To determine the number of taxonomic groups to analyze, we converted the phylogenetic tree into a distance matrix, using the cophenetic function implemented in the R package ape ("URLs"). We then clustered the 3,837 genomes into nine groups using *k*-medoids clustering as implemented in the PAM (partitioning around medoids) algorithm from the R package fpc ("URLs"). The *k*-medoids algorithm clusters a dataset of *n* objects into *k* a priori–defined clusters. To identify the optimal *k* value for the dataset, we compared the silhouette coefficients for values of *k* ranging from 1 to 30. We selected a value of $k=9$ because it yielded the maximal average silhouette coefficient (0.66). In addition, at $k=9$ the taxa were monophyletic, contained hundreds of genomes, and were relatively balanced between plant-associated and NPA genomes in most taxa (Table 1). The resulting genome clusters generally overlapped with annotated taxonomic units. One exception was in the Actinobacteria phylum. Here our clustering divided the genomes into two taxa that we named, for simplicity, Actinobacteria 1 and Actinobacteria 2. However, our rigorous phylogenetic analysis supports previous suggestions for revisions in the taxonomy of phylum Actinobacteria[88].

In addition, the tree showed very divergent bacterial taxa in the Bacteroidetes phylum that could not be separated into monophyletic groups. Specifically, the Sphingobacteriales order (from class Sphingobacteria) and the Cytophagaceae (from class Cytophagia) are paraphyletic. Therefore, we decided to unify all Bacteroidetes into one phylum-level taxon. Analysis of the prevalence of the nine taxa in 16S rDNA and metagenome appears in the Supplementary Information.

Pangenome analysis. For each comparison in Supplementary Fig. 2, a random set of ten genomes from each environment (plant-associated and NPA from specific environments) was selected, and the mean and s.d. of the phylogenetic distance in the set were calculated. This step was repeated 50 times to produce two random sets of genomes (plant-associated and NPA) that were comparable and had minimum differences between their mean and s.d. of phylogenetic distances. Genes for pangenome analysis were taken from the orthogroups (see below). Core genome, accessory genome, and unique genes were defined as genes that appeared in all ten genomes, in two to nine genomes, and in only one genome, respectively. For core and accessory genomes, the median copy number in each relevant orthogroup was used.

Genome size comparison and gene category enrichment analysis. Genome sizes were retrieved from the IMG database ("URLs") and compared by *t*-test and PhyloGLM[26]. Kernel density plots from the R sm package ("URLs") were used to prepare Supplementary Fig. 1. Protein-coding genes were retrieved and mapped to COG IDs with the program RPS-BLAST at an e-value cutoff of 1e–2 and an alignment length of at least 70% of the consensus sequence length. Each COG ID was mapped to at least one COG category (Supplementary Table 6). For each genome, we counted the number of genes from a given category. A *t*-test and PhyloGLM test were used to compare the number of genes in the genomes that shared the same taxon and category but different labels (e.g., plant-associated versus NPA).

Benchmarking gene clustering with UCLUST and OrthoFinder. We computed clusters of coding sequences across each of the nine taxa defined above with two algorithms: UCLUST[89] (v 7.0) and OrthoFinder[24] (v 1.1.4). UCLUST was run with 50% identity and 50% coverage in the target to call the clusters. Command used: usearch7.0.1090_i86linux64 -cluster_fast <input_file> -id 0.5 -maxaccepts 0 -maxrejects 0 -target_cov 0.5 –uc <output_file>. To improve pairwise alignment performance, we used the accelerated protein alignment algorithm implemented in DIAMOND[90] (v 0.8.36.98) with the --very-sensitive option in the DIAMOND BLASTP algorithm. After computing the alignments, we ran OrthoFinder with default parameters. See "URLs" for the scripts used to compute the orthogroups.

Supplementary Fig. 5 shows benchmarking of OrthoFinder against UCLUST. To estimate the quality of the clusters output by UCLUST and OrthoFinder, we mapped the proteins from our datasets to the curated set of taxon markers from Phyla-AMPHORA[91]. Next, we compared the distribution of each of the taxon-specific markers identified by Phyla-AMPHORA across the clusters output by UCLUST and OrthoFinder. To compare the two approaches, we estimated two metrics: the purity and the fragmentation index, explained in Supplementary Fig. 5 and in the Supplementary Information.

Identification of plant-associated, NPA, root-associated, and soil genes/domains. The following description applies to plant-associated, NPA, root-associated, and soil genes. For conciseness, only plant-associated genes are described here. Plant-associated genes were identified via a two-step process that included protein/domain clustering on the basis of amino acid sequence similarity and subsequent identification of the protein/domain clusters significantly enriched in proteins/domains from plant-associated bacteria (Supplementary Fig. 4). Clustering of genes and protein domains involved five independent methods: OrthoFinder[24], COG[20], KEGG Orthology (KO)[21], TIGRFAM[22], and Pfam[25]. OrthoFinder was selected (after the aforementioned benchmarking) as a clustering approach that included all proteins, including those that lack any functional annotation. We first compiled, for each taxon separately, a list of all proteins in the genomes. For COG, KO, TIGRFAM, and Pfam, we used the existing annotations

of IMG genes that are based on BLAST alignments to the different protein/domain models[23]. This process yielded gene/domain clusters. Next, we determined which clusters were significantly enriched with genes derived from plant-associated genomes. These clusters were termed plant-associated clusters. In the statistical analysis, we used only clusters of more than five members. We corrected *P* values with Benjamini–Hochberg FDR and used $q < 0.05$ as the significance threshold, unless stated otherwise. The proteins in each cluster were categorized as either plant-associated or NPA, on the basis of the label of the encoding genome. Namely, a plant-associated gene is a gene derived from a plant-associated gene cluster and a plant-associated genome.

The three main approaches were the hypergeometric test (Hyperg), PhyloGLM, and Scoary. Hyperg looks for overall enrichment of gene copies across a group of genomes but ignores the phylogenetic structure of the dataset. PhyloGLM[26] takes into account phylogenetic information to eliminate apparent enrichments that can be explained by shared ancestry. The Hyperg and PhyloGLM tests were used in two versions, based on either gene presence/absence data (hypergbin, phyloglmbin) or gene copy-number data (hypergcn, phyloglmcn). We also used a stringent version of Scoary[27], a gene presence/absence approach that combines Fisher's exact test, a phylogenetic test, and a label-permutation test. The first hypergeometric test, hypergcn, used the gene copy-number data, with the cluster being the sample, the total number of plant-associated and NPA genes being the population, and the number of plant-associated genes within the cluster being considered as 'successes'. The second version, hybergbin, used gene presence/absence data. *P* values were corrected by Benjamini–Hochberg FDR[92] for clusters of COG/KO/TIGRFAM/Pfam. For the abundant OrthoFinder clusters, we used Bonferroni correction with a threshold of $P < 0.1$, as downstream validation with metagenomes showed fewer false positives with the more significant clusters. The third and fourth statistical approaches used PhyloGLM[26], implemented in the phylolm (v 2.5) R package ("URLs"). PhyloGLM combines a Markov process of lifestyle (e.g., plant-associated versus NPA) evolution with a regularized logistic regression. This approach takes advantage of the known phylogeny to specify the residual correlation structure between genomes that share common ancestry, and so it does not need to make the incorrect assumption that observations are independent. Intuitively PhyloGLM favors genes found in multiple lineages of the same taxon. For each taxon we used the subtree from Fig. 1a to estimate the correlation matrix between observations and used the copy number (in phyloglmcn) or presence/absence pattern (in phyloglmbin) of each gene as the only independent variable. Positive and negative estimates in phyloglmbin/phyloglmcn indicated plant-associated/root-associated and NPA/soil-associated proteins/domains, respectively.

Finally, the fifth statistical approach was Scoary[27], which uses a gene presence/absence dataset. Scoary combines Fisher's exact test, a phylogeny-aware test, and an empirical label-switching permutation analysis. A gene cluster was considered significant by Scoary only if (1) it had a *q*-value less than 0.05 for Fisher's exact test, (2) the 'worst' *P* value from the pairwise comparison algorithm was < 0.05, and (3) the empirical (permutation-based) *P* value was < 0.05. These are very stringent criteria that yielded relatively few significant predictions. Odds ratios greater than or less than 1 in Scoary indicated plant-associated/root-associated and NPA/soil-associated proteins/domains, respectively.

See "URLs" for links to the code used for the gene enrichment tests. A description of additional assessment of plant-associated/NPA prediction robustness using validation genome datasets is presented in Supplementary Note 1.

Validation of predicted plant-associated, NPA, root-associated, and soil-associated genes using metagenomes. Metagenome samples ($n = 38$; Supplementary Table 16) were downloaded from NCBI and GOLD ("URLs"). The reads were translated into proteins, and proteins at least 40 amino acids long were aligned using HMMsearch[93] against the different protein references. The protein references included the predicted plant-associated, root-associated, soil-associated, and NPA proteins from OrthoFinder that were found to be significant by the different approaches. The normalization process is explained in Supplementary Figs. 12–16.

Principal coordinates analysis. To visualize the overall contribution of statistically significant enriched/depleted orthogroups to the differentiation of plant-associated and NPA genomes, we used principal coordinates analysis (PCoA) and logistic regression. For each of the nine taxa analyzed, we ran this analysis over a collection of matrices. The first matrix was the full pan-genome matrix, which depicted the distribution of all the orthogroups contained across all the genomes in a given taxon. The subsequent matrices represented subsets of the full pan-genome matrix; each of these matrices depicted the distribution of only the statistically significant orthogroups as called by one of the five different algorithms used to test for the genotype–phenotype association. A full description of this process is presented in Supplementary Note 1.

We used the function cmdscale from the R (v 3.3.1) stats package to run PCoA over all the matrices described above, using the Canberra distance as implemented in the vegdist function from the vegan (v 2.4-2) R package ("URLs"). Then, we took the first two axes output from the PCoA and used them as independent variables to fit a logistic regression over the labels of each genome (plant-associated, NPA). Finally, we computed the Akaike information criterion for each of the different models fitted. Briefly, the Akaike information criterion estimates how much information is lost when a model is applied to represent the true model of a particular dataset. See "URLs" for a link to the scripts used to perform the PCoA.

Validation of plant-associated genes in *Paraburkholderia kururiensis* M130 affecting rice root colonization. Growth and transformation details of *P. kururiensis* M130 are described in Supplementary Note 1.

Mutant construction. Internal fragments of 200–900 bp from each gene of interest were PCR-amplified with the primers listed in Supplementary Table 17c. Fragments were cloned in the pGem2T easy vector (Promega) and sequenced (GATC Biotech; Germany), then excised with *Eco*RI restriction enzyme and cloned in the corresponding site in pKNOCK-Km R[94]. These plasmids were then used as a suicide delivery system to create the knockout mutants and transferred to *P. kururiensis* M130 by triparental mating. All the mutants were verified by PCR with primers specific to the pKNOCK-Km vector and to the genomic DNA sequences upstream and downstream from the targeted genes.

Rhizosphere colonization experiments with P. kururiensis and mutant derivatives. Seeds of *Oryza sativa* (BALDO variety) were surface-sterilized and then left to germinate in sterile conditions at 30 °C in the dark for 7 d. Each seedling was then aseptically transferred into a 50-mL Falcon tube containing 35 mL of half-strength Hoagland solution semisolid substrate (0.4% agar). The tubes were then inoculated with 10^7 colony-forming units (cfu) of a *P. kururiensis* suspension. Plants were grown for 11 d at 30 °C (16/8-h light/dark cycles). For the determination of the bacterial counts, plants were washed under tap water for 1 min and then cut below the cotyledon to excise the roots. Roots were air-dried for 15 min, weighed, and then transferred to a sterile tube containing 5 mL of PBS. After vortexing, the suspension was serially diluted to 10^{-1} and 10^{-2} in PBS, and aliquots were plated on KB plates containing the appropriate antibiotic (rifampicin 50 µg/mL for the wild type, rifampicin 50 µg/mL and kanamycin 50 µg/mL for the mutants). After 3 d of incubation at 30 °C, we counted colony-forming units (CFU). Three replicates for each dilution from ten independent plantlets were used to determine the average CFU values.

Plant-mimicking plant-associated and root-associated proteins. Supplementary Fig. 21 summarizes the algorithm used to find plant-mimicking plant-associated and root-associated proteins. Pfam[25] version 30.0 metadata were downloaded. Protein domains that appeared in both Viridiplantae and bacteria and occurred at least two times more frequently in Viridiplantae than in bacteria were considered as plant-like domains ($n = 708$). In parallel, we scanned the set of significant plant-associated, root-associated, NPA, and soil-associated Pfam protein domains predicted by the five algorithms in the nine taxa. We compiled a list of domains that were significantly plant-associated/root-associated in at least four tests, and significantly NPA/soil-associated in up to two tests ($n = 1{,}779$). The overlapping domains between the first two sets were defined as PREPARADOs ($n = 64$). In parallel, we created two control sets of 500 random plant-like Pfam domains and 500 random plant-associated/root-associated Pfam domains. Enrichment of PREPARADOs integrated into plant NLR proteins in comparison to the domains in the control groups was tested by Fisher's exact test. To identify domains found in plant disease-resistance proteins, we retrieved all proteins from Phytozome and BrassicaDB ("URLs"). To identify domains in plant disease-resistance proteins, we used hmmscan to search protein sequences for the presence of NB-ARC (PF00931.20), TIR (PF01582.18), TIR_2 (PF13676.4), or RPW8 (PF05659.9) domains. Bacterial proteins carrying the PREPARADO domains were considered as having full-length identity to fungal, oomycete, or plant proteins on the basis of LAST alignments to all Refseq proteins of plants, fungi, and protozoa. "Full-length" is defined here as an alignment length of at least 90% of the length of both query and reference proteins. The threshold used for considering a high amino acid identity was 40%. An explanation of the prediction of secretion of proteins with PREPARADOs is presented in the Supplementary Information.

Prediction of plant-associated, NPA, root-associated, and soil-associated operons and their annotation as biosynthetic gene clusters. Significant plant-associated, NPA, root-associated, and soil-associated genes of each genome were clustered on the basis of genomic distance: genes sharing the same scaffold and strand that were up to 200 bp apart were clustered into the same predicted operon. We allowed up to one spacer gene, which is a non-significant gene, between each pair of significant genes within an operon. Operons were predicted for the genes in COG and OrthoFinder clusters using all five approaches. Operons were annotated as biosynthetic gene clusters if at least one of the constituent genes was part of a biosynthetic gene cluster from the IMG-ABC database[95].

***Jekyll* and *Hyde* analyses.** To find all homologs and paralogs of *Jekyll* and *Hyde* genes, we used IMG BLAST search with an e-value threshold of 1e−5 against all IMG isolates. We searched *Hyde1* homologs of *Acidovorax*, *Hyde1* homologs of *Pseudomonas*, *Hyde2*, and *Jekyll* genes using proteins of genes Aave_1071, A243_06583, Ga0078621_123530, and Ga0102403_10160 as the query sequence, respectively. Multiple sequence alignments were done with Mafft[96]. A phylogenetic

tree of *Acidovorax* species was produced with RaxML[97], based on concatenation of 35 single-copy genes[98].

Hyde1 toxicity assay. To verify the toxicity of Hyde1 and Hyde2 proteins to *E. coli*, we cloned genes encoding proteins Aave_0990 (Hyde2), Aave_0989 (Hyde1), and Aave_3191 (Hyde1), or GFP as a control, to the inducible pET28b expression vector via the LR reaction. The recombinant vectors were transformed into *E. coli* C41 competent cells by electroporation after sequencing validation. Five colonies were selected and cultured in LB liquid media supplemented with kanamycin with shaking overnight. The OD_{600} of the bacterial culture was adjusted to 1.0, and then the culture was diluted by 10^2, 10^4, 10^6, and 10^8 times successively. Bacteria culture gradients were spotted (5 µL) on LB plates with or without 0.5 mM IPTG to induce gene expression.

Construction of Δ5-Hyde1 strain. Details of the construction of the Δ5-Hyde1 strain are presented in Supplementary Note 1. *A. citrulli* strain AAC00-1 and its derived mutants were grown on nutrient agar medium supplemented with rifampicin (100 µg/ml). To delete a cluster of five *Hyde1* genes (Aave_3191–3195), we carried out a marker-exchange mutagenesis as previously described[99]. The marker-free mutant was designated as Δ1-Hyde1, and its genotype was confirmed by PCR amplification and sequencing. The marker-exchange mutagenesis procedure was repeated to delete four other *Hyde1* loci (Supplementary Fig. 28). The primers used are listed in Supplementary Table 25. The final mutant, with deletion of 9 out of 11 *Hyde1* genes (in five loci), was designated as Δ5-Hyde1 and was used for competition assay. The ΔT6SS mutant was provided by Ron Walcott's lab.

Competition assay of *Acidovorax citrulli* AAC00-1 against different strains. *Bacterial strains. E. coli* BW25113 pSEVA381 was grown aerobically in LB broth (5 g/L NaCl) at 37 °C in the presence of chloramphenicol. Naturally antibiotic-resistant bacterial leaf isolates[16] and *Acidovorax* strains were grown aerobically in NB medium (5 g/L NaCl) at 28 °C in the presence of the appropriate antibiotic. Antibiotic resistance and concentrations used in the competition assay are mentioned in Supplementary Table 25.

Competition assay. Competition assays were conducted similarly as described elsewhere[66,100]. Briefly, bacterial overnight cultures were harvested and washed in PBS (pH 7.4) to remove excess antibiotics, and resuspended in fresh NB medium to an optical density of 10. Predator and prey strains were mixed at a 1:1 ratio, and 5 µL of the mixture was spotted onto dry NB agar plates and incubated at 28 °C. As a negative control, the same volume of NB medium was mixed with prey cells instead of the predator strain. After 19 h of coincubation, bacterial spots were excised from the agar and resuspended in 500 µL of NB medium and then spotted on NB agar containing antibiotic selective for the prey strains. CFUs of recovered prey cells were determined after incubation at 28 °C. All assays were performed in at least three biological replicates.

Life sciences reporting summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. All new genomes (Supplementary Table 3) were submitted and are publicly available in at least one of the following databanks (see accessions in Supplementary Table 3):

1. IMG/M, https://img.jgi.doe.gov/cgi-bin/m/main.cgi
2. Genbank, https://www.ncbi.nlm.nih.gov/genbank/
3. ENA, http://www.ebi.ac.uk/ena
4. A dedicated website for the Dangl lab: http://labs.bio.unc.edu/Dangl/Resources/gfobap_website/index.html

The dedicated website contains nucleotide and amino acid FASTA files of all datasets used, protein/domain annotations (COG, KO, TiGRfam, Pfam), metadata, phylogenetic trees, OrthoFinder orthogroups, orthogroup hidden Markov models, full enrichment datasets, correlation between orthogroups, and predicted operons ("URLs").

Links to different scripts that were used in analysis are included in the "URLs" section. The full genome sequence, gene annotation, and metadata of each genome used can be found at the IMG website (https://img.jgi.doe.gov/). For example, the metadata of taxon ID 2558860101 can be found at https://img.jgi.doe.gov/cgi-bin/mer/main.cgi?section=TaxonDetail&page=taxonDetail&taxon_oid=2558860101.

References

79. Doty, S. L. et al. Diazotrophic endophytes of native black cottonwood and willow. *Symbiosis* **47**, 23–33 (2009).
80. Weston, D. J. et al. *Pseudomonas fluorescens* induces strain-dependent and strain-independent host plant responses in defense networks, primary metabolism, photosynthesis, and fitness. *Mol. Plant Microbe Interact.* **25**, 765–778 (2012).
81. Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
82. Beszteri, B., Temperton, B., Frickenhaus, S. & Giovannoni, S. J. Average genome size: a potential source of bias in comparative metagenomics. *ISME J.* **4**, 1075–1077 (2010).
83. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
84. Varghese, N. J. et al. Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* **43**, 6761–6771 (2015).
85. Kerepesi, C., Bánky, D. & Grolmusz, V. AmphoraNet: the webserver implementation of the AMPHORA2 metagenomic workflow suite. *Gene* **533**, 538–540 (2014).
86. Wu, M., Chatterji, S. & Eisen, J. A. Accounting for alignment uncertainty in phylogenomics. *PLoS One* **7**, e30288 (2012).
87. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
88. Sen, A. et al. Phylogeny of the class Actinobacteria revisited in the light of complete genomes. The orders 'Frankiales' and Micrococcales should be split into coherent entities: proposal of Frankiales ord. nov., Geodermatophilales ord. nov., Acidothermales ord. nov. and Nakamurellales ord. nov. *Int. J. Syst. Evol. Microbiol.* **64**, 3821–3832 (2014).
89. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
90. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
91. Wang, Z. & Wu, M. A phylum-level bacterial phylogenetic marker database. *Mol. Biol. Evol.* **30**, 1258–1262 (2013).
92. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
93. Finn, R. D. et al. HMMER web server: 2015 update. *Nucleic Acids Res.* **43**, W30–W38 (2015).
94. Alexeyev, M. F. The pKNOCK series of broad-host-range mobilizable suicide vectors for gene knockout and targeted DNA insertion into the chromosome of gram-negative bacteria. *Biotechniques* **26**, 824–826 (1999).
95. Hadjithomas, M. et al. IMG-ABC: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *MBio* **6**, e00932 (2015).
96. 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.* **30**, 3059–3066 (2002).
97. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* **57**, 758–771 (2008).
98. Finkel, O. M., Béjà, O. & Belkin, S. Global abundance of microbial rhodopsins. *ISME J.* **7**, 448–451 (2013).
99. Traore, S. M. *Characterization of Type Three Effector Genes of A. citrulli, the Causal Agent of Bacterial Fruit Blotch of Cucurbits*. (Virginia Polytechnic Institute and State University, Blacksburg, VA, 2014).
100. Basler, M., Ho, B. T. & Mekalanos, J. J. Tit-for-tat: type VI secretion system counterattack during bacterial cell-cell interactions. *Cell* **152**, 884–894 (2013).

nature research

Corresponding author(s): Jeffery L. Dangl

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Computational analysis is done based on sample size in the order of hunderds. Experiments were done with sample size of 3-30 biological replicates.

2. Data exclusions

Describe any data exclusions.

No data was excluded.

3. Replication

Describe whether the experimental findings were reliably reproduced.

All reported results in the paper was reliably reproduced in biological replicates.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Mostly irrelevant to the study. In the few cases where we used randomizations we made sure the randomized control group has similar features to the tested group. For example in the randomization done for PREPARADOs enrichment in planr NLR immune proteins we randomly sampled 500 PA/RA domains and 500 plant protein domains.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Irrelevant to the study.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	All software used is described in methods and provided to readers. We mention different R packages used in analysis.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	There are no restrictions on materials availability.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	Irrelevant to the study
b. Describe the method of cell line authentication used.	Irrelevant to the study
c. Report whether the cell lines were tested for mycoplasma contamination.	Irrelevant to the study
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	Irrelevant to the study

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	Irrelevant to the study

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Irrelevant to the study