MICROBIAL GENOMICS

RESEARCH ARTICLE

Siani *et al.*, *Microbial Genomics* 2021;7:000666

DOI 10.1099/mgen.0.000666

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Acidovorax pan-genome reveals specific functional traits for plant beneficial and pathogenic plant-associations

Roberto Siani[1,2], Georg Stabl[3], Caroline Gutjahr[3], Michael Schloter[1,2] and Viviane Radl[1,*]

Abstract

Beta-proteobacteria belonging to the genus *Acidovorax* have been described from various environments. Many strains can interact with a range of hosts, including humans and plants, forming neutral, beneficial or detrimental associations. In the frame of this study, we investigated the genomic properties of 52 bacterial strains of the genus *Acidovorax*, isolated from healthy roots of *Lotus japonicus,* with the intent of identifying traits important for effective plant-growth promotion. Based on single-strain inoculation bioassays with *L. japonicus*, performed in a gnotobiotic system, we distinguished seven robust plant-growth promoting strains from strains with no significant effects on plant-growth. We showed that the genomes of the two groups differed prominently in protein families linked to sensing and transport of organic acids, production of phytohormones, as well as resistance and production of compounds with antimicrobial properties. In a second step, we compared the genomes of the tested isolates with those of plant pathogens and free-living strains of the genus *Acidovorax* sourced from public repositories. Our pan-genomics comparison revealed features correlated with commensal and pathogenic lifestyle. We showed that commensals and pathogens differ mostly in their ability to use plant-derived lipids and in the type of secretion-systems being present. Most free-living *Acidovorax* strains did not harbour any secretion-systems. Overall, our data indicate that *Acidovorax* strains undergo extensive adaptations to their particular lifestyle by horizontal uptake of novel genetic information and loss of unnecessary genes.

DATA SUMMARY

The authors confirm all supporting data, code and protocols have been provided within the article or through supplementary data files. Novel genome assemblies are available in the European Nucleotide Archive under the accession number PRJEB37696. Other genome sequences employed in the study are available on NCBI-Genomes. A complementary R Markdown document containing the code used in this study is available on https://github.com/rsiani/AVX_PGC.

INTRODUCTION

Several soil-derived bacteria, which can colonize plant roots, directly or indirectly promote plants' health by facilitating nutrient mobilization and uptake [1], modulating plant hormone levels [2] and competing with pathogens [3, 4]. At the same time, soils also harbour a large repertoire of plant pathogens. It is not entirely clear what distinguishes non-harmful root-associated commensals from pathogens, as even closely related strains have been proven to promote plant-growth or, contrariwise, to negatively affect it [5]. Several studies [6, 7] link this to loss or acquisition of genomic features (e.g. virulence factor, pathogenicity islands), resulting from mutations and horizontal gene transfer [8]. Our existing knowledge in this field is still scarce and inconsistent, as in most cases only single strains were compared, which does not allow evolutionary patterns to be understood and a generalized model to be defined.

Received 09 February 2021; Accepted 05 August 2021; Published 10 December 2021

Author affiliations: [1]Helmholtz Center for Environmental Health, Institute for Comparative Microbiome Analysis, Ingolstaedter Landstr, Oberschleissheim, Germany; [2]Technical University of Munich, School of Life Sciences, Chair for Soil Science, Freising, Germany; [3]Technical University of Munich, School of Life Sciences, Plant Genetics, Freising, Germany.

***Correspondence:** Viviane Radl, viviane.radl@helmholtz-muenchen.de

Keywords: *Acidovorax*; plant pathogens; *Lotus japonicus*; plant growth promotion.

Abbreviations: ANI, average nucleotide identity; BGCs, biosynthesis gene clusters; BNS-AM, basal nutrient solution modified for arbuscular mycorrhiza; GABA, gamma-aminobutyric acid; HGT, horizontal gene transfer; kb, kilobases; KW, factorial Kruskal–Wallis test; LDA, linear discriminant analysis; LEfSe, Linear discriminant analysis Effect Size; PBPs, periplasmic binding proteins; PCA, Principal component analysis; Pfams, protein families; TRAPs, tripartite ATP-independant periplasmic transporters; TTTs, tripartite tricarboxylate transporters; W, pairwise Wilcoxon test.

Data statement: Five supplementary tables are available with the online version of this article.

Bacteria of the genus *Acidovorax* belong to the group of Gram-negative beta-proteobacteria [9] and form associations with a wide range of monocotyledonous and dicotyledonous plants [10]. *Acidovorax* has been mostly considered as biotrophic pathogen [11]. However, bacteria of this genus are also known as commensal species or plant beneficial bacteria, which produce secondary metabolites and hormones promoting plant growth, as well as competing with pathogens [12]. Most common species are *A. delafieldii* and *A. facilis*, whilst the most studied pathogenic species is *A. avenae*, the agent of corn (*Zea mays*), sugar cane (*Saccharum officinarum*) and rice (*Oryza sativa*) leaf blight, orchids brown spot disease and watermelon (*Citrullus lanatus*) fruit blotch. Thus, *Acidovorax* is a valuable model organism to investigate evolutionary patterns of plant-associated bacteria and to study which genes differentiate bacteria acting as biotrophic pathogens, as commensals or plant growth-promoting bacteria [13].

Pan-genomics, which is made possible by reductions in sequencing prices and improvement in the field of bioinformatics [14], allow dozens to hundreds of strains to be screened, also at lower taxonomic resolution, for genomics idiosyncrasies associated with traits of interest, such as pathogenic or beneficial plant-association [15]. Although several efforts to explore the genomes of pathogenic *Acidovorax* species have been undertaken [16–18], there has not yet been an attempt to reframe strain-specific findings at genus level and consolidate the understanding of the genomic basis of *Acidovorax* interaction with plants across multiple lifestyles and clades.

In the frame of this study, we compared the genomes of isolates classified as *Acidovorax,* which were obtained from roots of healthy *L. japonicus* ecotype Gifu plants. We tested the strains for their ability to promote plant growth of *L. japonicus* in sterilized sand and correlated differences in the genomes of the strains to the observed effects on plant growth. For a broader comparative analysis, we included additional genomes, available in public databases, belonging to strains from 15 different species of *Acidovorax* from different environments, including all major groups of plant pathogens from this genus. We studied the diversity in the pan-genome and critically evaluated enriched gene functions, secondary metabolites biosynthetic clusters in the different functional groups of *Acidovorax*.

Impact Statement

We still have a limited understanding of genomics basis of plant-bacteria associations and most of the existing knowledge stems from empirical evidences at strain-species level. Thus, we adopted an integrative approach to probe binary association between the model leguminous *L. japonicus* and the genus *Acidovorax*. Leveraging empirical evidence from *in planta* bioassays and the breadth of available genomics data, we were able to retrieve genomics biomarkers strongly associated with different behavioural phenotypes and train unsupervised models for accurate genotype to phenotype mapping. Furthermore, our findings shed light on the dynamics of *Acidovorax* specialization, via acquisition and loss of genomics traits. Our study increments the understanding of genomics constraints of plant-associated bacteria and identifies a wealth of putative biomarkers, which could be leveraged for future biotechnological applications in detecting and managing plant bacterial diseases and/or implementing ecologically sensible agricultural solutions.

METHODS

Origin of strains and genomes

Fifty-two strains of endophytic *Acidovorax* were isolated from the root systems of healthy *L. japonicus* ecotype Gifu B-129 (subsequently called *L. japonicus*) grown in natural soil in Cologne, Germany. DNA from all strains was extracted and sequenced using an Illumina HiSeq 2500 device (Illumina, USA). The isolation and sequencing procedure has been described in detail elsewhere [19]. Genome assemblies are available in the European Nucleotide Archive under the accession number PRJEB37696.

In addition, 54 more genomes where obtained from NCBI. The collection includes 15 species: *A. anthurii*, *A. avenae*, *A. carolinensis*, *A. cattleyae*, *A. citrulli*, *A. delafieldii*, *A. defluvii*, *A. ebreus*, *A. facilis*, *A. kalamii*, *A. konjaci*, *A. monticola*, *A. oryzae*, *A. radicis, A. varianellae* and *A. spp*. According to the database, all the selected bacterial genomes originate from plant, soil and water samples. Details on the selected genomes are summarized in Table S1, available in the online version of the article. The mean level of completion for all 106 genomes was calculated to 0.986.

Based on the available metadata, we assigned the genomes to three behavioural phenotypes: 62 genomes belong to commensal or beneficial plant-associated strains (thereby termed 'commensals'), 21 to plant pathogenic strains (thereby termed 'pathogens') and 23 to free-living strains (thereby termed 'free-living'). According to the results of our *in planta* bioassays (see below), we further classified the *Lotus*-isolated strains by their ability to promote *L. japonicus* growth.

Linear discriminant analysis effect sizes

We studied genomic features across the strains isolated from *L. japonicus* and correlated the obtained data with the effects detected by the *in planta* bioassays (see below). Therefore, we annotated the genomes using Anvi'o 6.2 microbial multi-omics platform [20], which computes *k*-mer frequencies and identifies coding sequences using Prodigal [21]. Afterwards, HMMER3 [22] was used to profile the genomes with hidden Markov models and to find homologous protein families in the Pfam database v33.1 [23]. Using the feature occurrence frequency matrix generated by Anvi'o 6.2, we calculated the effect sizes of the features over a linear discriminant analysis

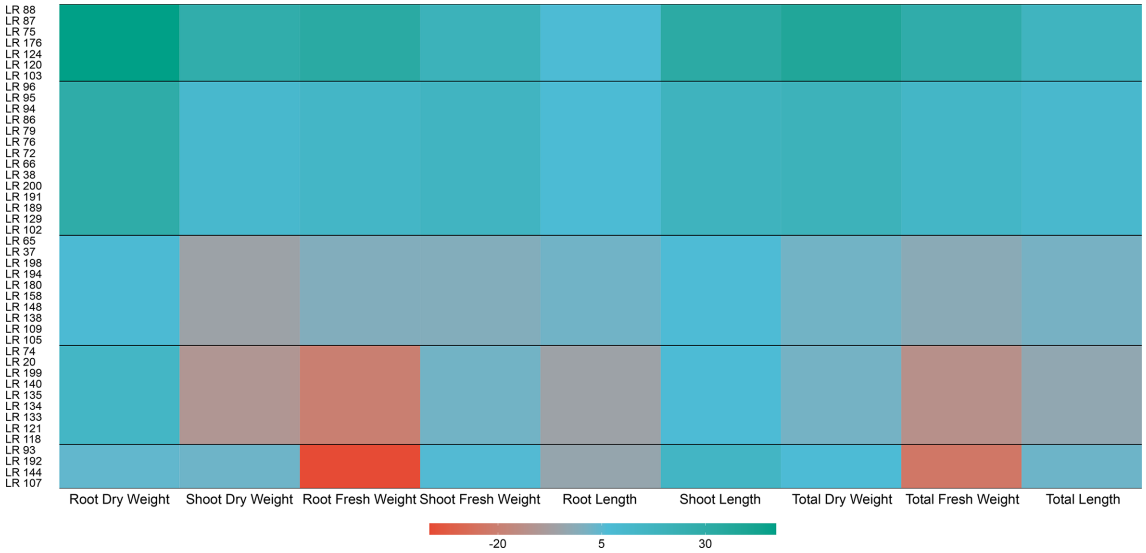

Fig. 1. Heat-map showing the effects of *Acidovorax* strains on *L. japonicus* growth parameters as observed in the bioassays. Strains were clustered in five groups according to the outcome across the replicates (n=25), assessed by pairwise Student's t-test. Results are shown here as the median, for each observed group, of the percentage change from the control.

(LEfSe [24]) between the groups of interest. We chose the default alpha values of 0.05 for the factorial Kruskal–Wallis test (KW) and the pairwise Wilcoxon test (W) and a threshold of 2.0 for the linear discriminant analysis (LDA) logarithmic scores.

Pan-genome analysis

We used Anvi'o 6.2 further to pre-process the genomes and reconstruct *Acidovorax* pan-genome [25]. We inferred the pan-genome with the following options: as suggested by the developers, DIAMOND was set to sensitive mode to calculate amino acid sequences similarity [26], minbit was kept at the default value of 0.5, minimum occurrence of gene clusters to 2 to exclude singletons from downstream processing and MCL inflation was increased to 7 to obtain a finer granularity of the clusters of orthologous genes (a default of 2 is suggested by the developers when comparing distantly related genomes, of 10 for closely related genomes). We separated gene clusters into occurrence frequency classes. Core gene clusters were defined as present in 100 to 106 genomes (more than 95% of the genomes). Shell gene clusters were defined as present in 7 to 99 genomes (between 5 and 95% of the genomes). Cloud gene clusters were defined as present in 2 to 6 genomes (below 5% of the genomes). Singleton gene clusters, present in only one genome, were excluded from the downstream processing to reduce computational load. We calculated functional and geometrical homogeneity for each gene cluster and genome average nucleotide identity using pyANI [27]. We calculated a function occurrence frequency matrix and function enrichment scores for gene clusters' functions by assigning each gene cluster to the closest Pfam.

Comparative genomics

We performed all downstream analysis in R (v4.02, code available on https://github.com/rsiani/AVX_PGC). We calculated total length, GC%, number of genes, gene clusters and singleton gene clusters, average gene length and number of genes per kb (gene density) for each of the genomes. We assessed the significance of differences by pairwise Student's t-test with Holm–Bonferroni's correction for multiple testing (P<0.05).

We converted the 'function frequency occurrence matrix' (3036 Pfams) to a binary presence/absence matrix and filtered zero variance (757 Pfams), near zero variance (1668 Pfams) and correlated variables (0 Pfams) to remove redundancy (resulting in 1368 Pfams). We performed a principal component analysis (PCA, R package: 'FactoMineR', v2.3) to ordinate the strains based on their functional similarity and tested the significance of the clustering by PERMANOVA (package: 'vegan', v2.5–7). We isolated the features responsible for most of the variance explained by extracting the highest contributors for each of the first three components.

From the enriched feature profiles predicted by Anvi'o 6,2, we manually curated a list of features putatively correlated with plant-association. Enrichment score and adjusted significance values were used to assess the relevance of the features across the groups.

We retrieved secondary metabolites biosynthesis gene clusters using AntiSMASH 5.0 [28]. We launched a local instance of the programme with no extra features at a 'relaxed' strictness level, running only the core detection modules. At this strictness level, well-defined clusters and partial clusters were detected and annotated against the antiSMASH database.

We built a classification model using neural networks with feature extraction (R package 'caret', v6.0–86), a model

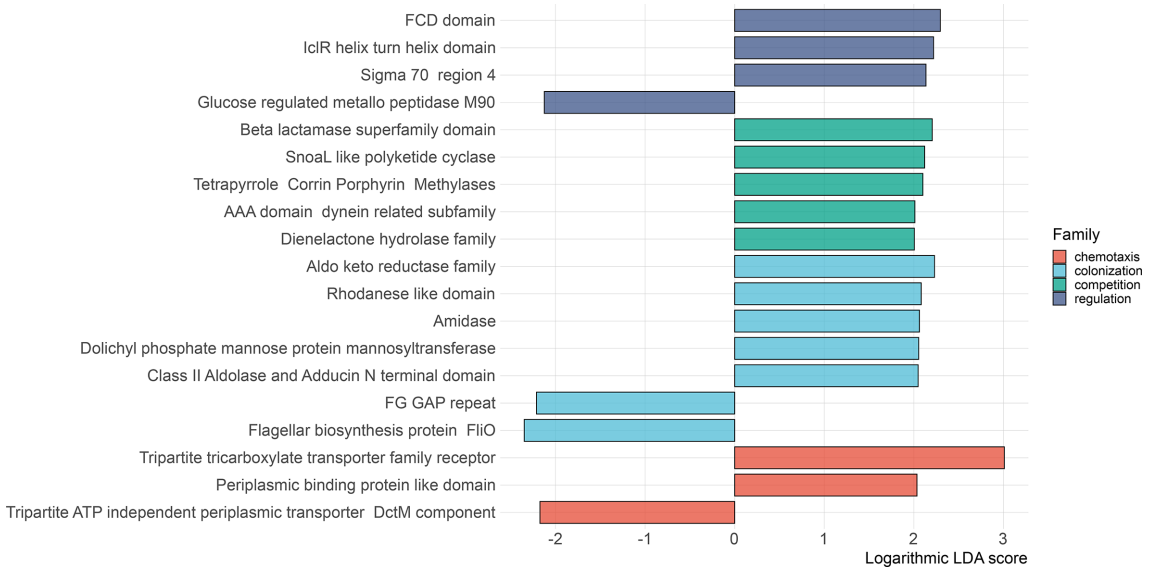

Fig. 2. Bar-plot displaying the discriminatory power, as calculated by LEfSe, of 19 Pfams in separating seven *Acidovorax* strains able to promote *L. japonicus* growth in the *in planta* bioassays from the remainder of the tested strains. The 19 Pfams were first isolated by factorial Kruskal–Wallis rank sum test and their relevance assessed by calculating the effect sizes in a linear discriminant analysis. We grouped the biomarkers to four broad, colour-coded, functional groups: chemotaxis, colonization, competition and regulation.

employing a preliminary feature extraction step to reduce redundancy, while preserving information and decreasing computational load [29]. We partitioned our genomes in a training set (90% of the genomes, *n*=96) and a testing set (10% of the genomes, *n*=10). We trained the model on the training set with repeated tenfold cross-validation over ten repeats. Briefly, the training set was split in ten partition and at each partition was held out in turn for testing the performance of the model. The partitioning has been repeated ten times with different randomly generated parameters for the model. The optimal parameters for the model have been selected by evaluating the average accuracy across the repeats. Finally, the fit of the model has been evaluated by predicting labels for the testing set and calculating a confusion matrix.

In planta bioassays

We prepared seeds of *L. japonicus* by sandpaper scarification and surface sterilization with a 10% DanKlorix Original (CP GABA GmbH, Germany) and 0.1% sodium dodecyl sulphate (SDS) solution. We germinated the seeds on a 0,8% water-agar plate for 3 days at 22 °C in the dark followed by 4 days at 22 °C in the light at 210 µM/m²s intensity (growth cabinet PK 520-LED, Polyklima).

We transferred plantlets to pots (Göttinger 7×7 x 8 cm, Hermann Meyer KG, Germany) filled with washed and autoclave sterilized quartz-sand (Casafino Quarzsand, fire-dried, 0.7–1.2 mm, BayWa AG, Germany) at a density of five plants per pot. Pots were supplied with a thin layer of synthetic cotton at the bottom to prevent the sand running out through the holes.

In total, 44 of the *Acidovorax* strains isolated from healthy *L. japonicus* plants were individually pre-grown in 50% TSB media at 30 °C overnight and diluted to an OD_{600} of 0.001 using 1/3 BNS-AM medium (Basal Nutrient Solution [30] modified for arbuscular mycorrhiza (0.025 mM $KH_2PO_{4)}$). Then, 25 ml of this bacterial suspension was applied to the pots. Finally, 20 ml of the sterile 1/3 BNS-AM-bacterial solution was added. Plants were grown at 60 % air humidity and light intensity of 150 µM/m²s. The photo-period was set to 16 h light/8 h dark long day with a temperature cycle of 24/22 °C. For each bacterial strain tested, four replicate pots were prepared.

After 9 weeks of growth, we harvested the plants and determined root and shoot length, wet weight and dry weight after freeze-drying (Alpha 1–2, Martin Christ Gefriertrocknungsanlagen GmbH, Germany). Statistical significance was assessed by pairwise Student's *t*-test adjusted for multiple comparisons using the Holm–Bonferroni method ($P<0.05$).

RESULTS

Acidovorax strains isolated from healthy *L. japonicus* roots diversely affect their host

We compared growth metrics of *L. japonicus* inoculated with different strains to evaluate the outcomes of the plant-association. Overall, 34 out of 44 tested strains showed significant effects ($P<0.05$; Fig. 1) when considering all the collected measurements of plants' growth (*n*=25). Twenty-one strains displayed a positive effect on at least one metric. Nine strains displayed a negative effect on plant growth, markedly on fresh weight measurements. Four strains displayed contrasting

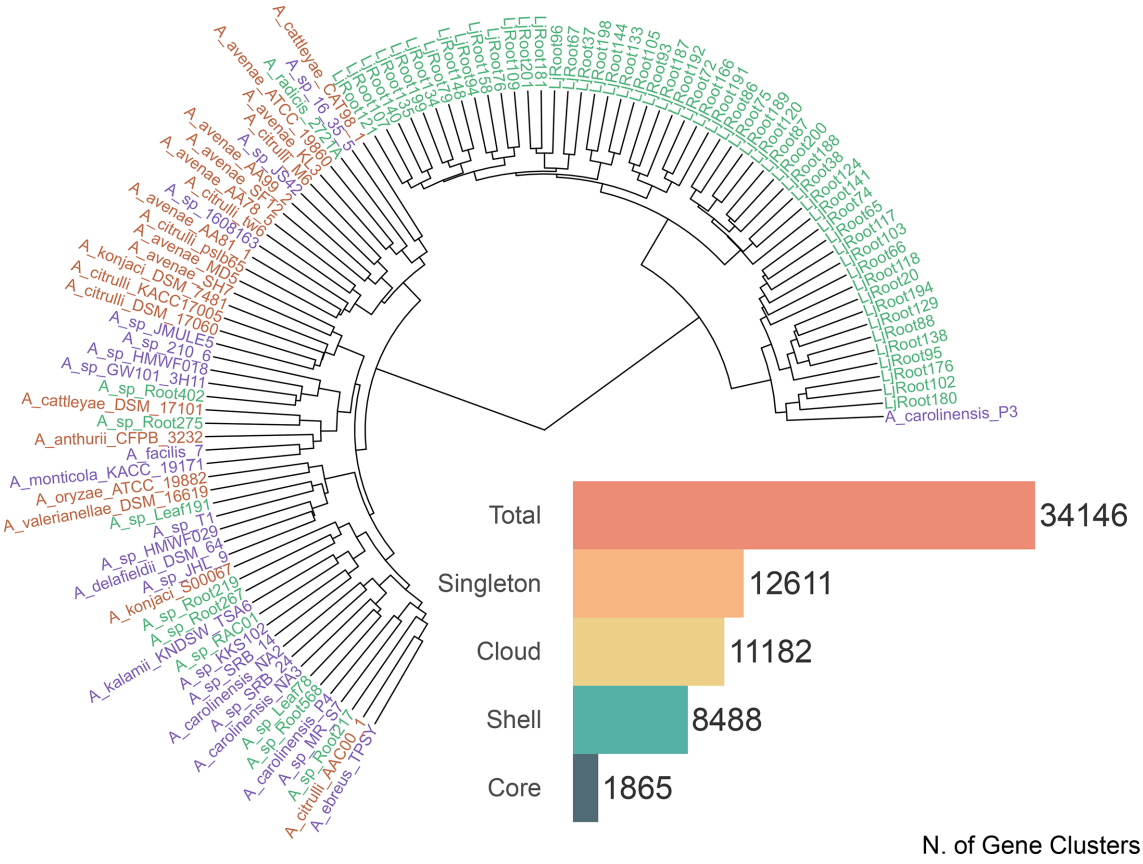

Fig. 3. Phylogenetic tree of the 106 *Acidovorax* strains included in the pan-genome. We calculated the average nucleotide identity between genomes using pyANI and performed a hierarchical clustering on euclidean distance with Ward metrics (R package: 'stats', v4.0.2). Genome sequences are attributed to three groups by the available metadata: pathogens (red), free-living (purple) and commensal (green). The bar-plot shows the total number of gene clusters retrieved in the pan-genome and the frequency of clusters across the genomes: singleton clusters are only found in one genome, cloud clusters are found in less than six genomes, shell clusters are found in 7 to 99 genomes and core clusters are present in more than 100 genomes.

effects on different metrics and ten had no significant effect on growth. From the 21 strains that resulted in improved plant-growth, seven strains displayed a more consistent effect across all the replicates on at least one of the considered metrics (P<0.001). We selected these robust growth-promoters for a follow-up genomics comparison.

Nineteen Pfams discriminate the robust plant-growth-promoting *Acidovorax* strains

To isolate genomic differences correlated with the outcomes of the *in planta* bioassays, we calculated linear discriminant analysis effect sizes on the function occurrence frequency table and identified 19 Pfams discriminating the seven robust growth-promoters from the remainder of the *Lotus*-isolated strains (KW P<0.05, W P<0.05, logarithmic LDA score >2.0, Fig. 2; Table S2). Out of those, 15 Pfams were enriched in the genomes of robust growth-promoters, related to four broad functional categories: chemotaxis by sensing and uptake of organic acids and sugars (tripartite tricarboxylate transporters, TTTs, and periplasmic binding proteins, PBPs), colonization through synthesis and detoxification of plant secondary metabolites (aldolase, aldo-keto reductase, rhodanase, amidase and mannosyltransferase), competition by synthesis, secretion and detoxification of antimicrobial compounds (beta-lactamase, polyketide cyclase, tetrapyrrole corrin-porphyrin methlyase, dynein-related domain and dienelactone hydrolase) and transcriptional regulation (FCD domain, IclR domain and sigma 70 factor). We found four Pfams, which were less abundant in the genomes of robust growth-promoters, also implicated in chemotaxis (Tripartite ATP-independent periplasmic transporters, TRAPs), colonization (FliO and the FG-GAP repeat) and regulation (metallo-peptidase M90).

Acidovorax has an open pan-genome

We reconstructed the *Acidovorax* pan-genome from 106 genomes (Fig. 3), representing 15 out of the 23 classified *Acidovorax* species [31], along with several yet unclassified strains, and capturing a large share of the genomic variability in the genus. The pan-genome contains a total of 523 555 genes grouped in 34 146 gene clusters, of which 5 % belonged to the core (1865 clusters present in 100 to 106 genomes), 25

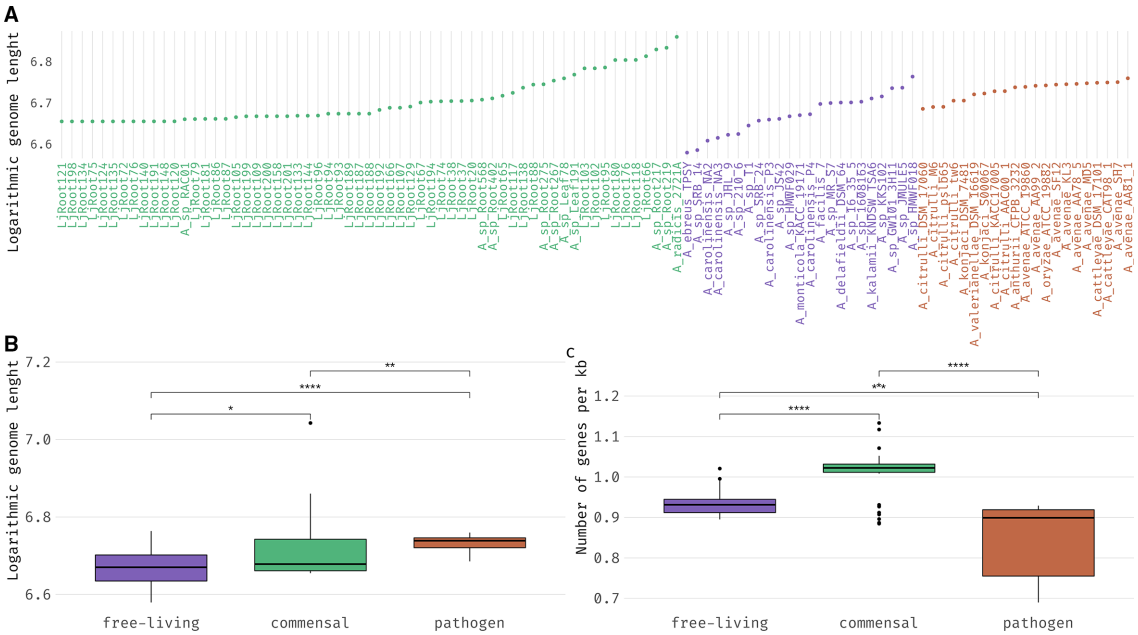

Fig. 4. (a) Dot-chart of logarithmic genome length for each of the genomes. (b) Box-plot comparing logarithmic genome length across the different behavioural groups. (c) Box-plot of each group gene density expressed as the number of genes per 1000 base-pairs. Significance was assessed by Student's *t*-test with Holm's correction ($P<0.05$). Genomes attributed to commensal strains are represented in green, genomes attributed to free-living and pathogenic strains are represented in purple and red, respectively.

% belonged to the shell (8488 clusters present in 16 to 99 of the genomes), 33 % belonged to the cloud (11 182 clusters present in 2 to 15 genomes) and 37 % were singleton (12 611 clusters present in one genome only). By fitting our results to a Heaps' power-law regression model [32] we obtained a $\gamma<1$ of 0.48, which defines *Acidovorax* pan-genome as 'open', with positive rates of novel gene discovery for every additional genome included.

Pathogenic *Acidovorax* strains have longer genomes but fewer genes

We calculated total length, GC%, number of genes, gene clusters and singleton gene clusters, average gene length and number of genes per kb (gene density) for each of the genomes (Fig. 4a, Table S3). On average, pathogens had longer genomes (5.38 mb, SD=0.27) than commensals (5.21 mb, SD=1.03, $P<0.01$) and free-living (4.72 mb, SD=0.53, $P<0.0001$, Fig. 4b). However, when testing for differences in gene density, pathogens had less genes per kb (0.86, SD=0.09) than both commensals (1.01, SD=0.05, $P<0.0001$) and free-living (0.94, SD=0.03, $P<0.001$, Fig. 4c).

Strains separate according to behavioural phenotype

We performed a PCA on the function presence/absence matrix and found that the first three dimensions explained 47 % of the variance in the matrix (24.3, 12.4 and 10.3 %). The individual strains clustered according to their observed behavioural phenotype ($P<0.05$), with pathogens grouping in a neatly separated cluster and a partial overlap between free-living and commensal strains (Fig. 5a,b). Notably, no clustering was visible when considering the original habitat of the strains. We extracted the highest contributing variables for the first three components (Fig. 5c, Table S4).

This approach retrieved a total of 14 Pfams displaying a strong variance across the groups. Two of these, the C-terminal domain of a secreted lipases and the mutagenesis inducer HIM1, contributed the most to the first component and were found in all of the commensals, 35 % of the free-living and none of the pathogens. A third Pfam, characterized by a WG repeat motif, was only retrieved in 40 % of the commensals and accounted for the high explanatory power of the second component. Finally, the highest contributor to the third component was a putative genomic island, encoding 11 elements of the type VI secretion-system, present in 95 % of the pathogens, 61 % of the commensals and only 13 % of the free-living strains.

Pathogenic and commensal *Acidovorax* strains have a distinctive set of features

We performed a function enrichment analysis to isolate features unique or more represented in the different groups. Overall, 1243 out of 3036 Pfams were significantly enriched in one or two groups (adjusted $q<0.05$). We further investigated enriched functions from plant-associated strains only: 371 Pfams strongly associated with commensals ($q<0.05$) and 303 Pfams strongly associated with pathogens ($q<0.05$) (Fig. 6a). From these, we manually curated a list of 54 features (Table S5) putatively involved in plant–microbe and

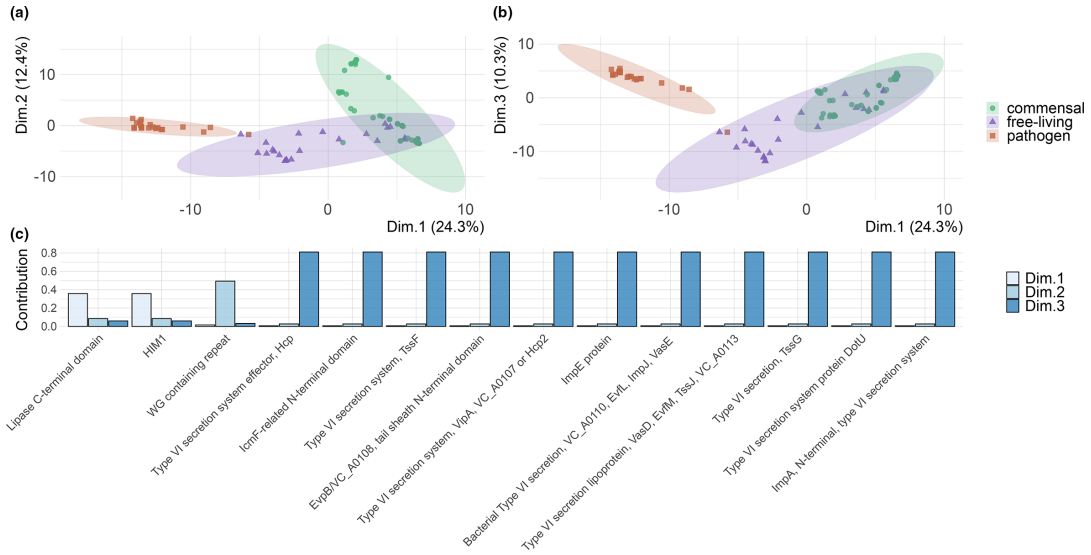

Fig. 5. (a,b) Ordination plots of the first and second (a) and first and third (b) components from a principal component analysis on the presence/absence patterns of gene functions in the pan-genome, showing the clustering of individual genomes according to their gene repertoire. Genomes attributed to commensal strains are represented as green circles, genomes attributed to free-living and pathogenic strains are represented as purple triangles and red squares, respectively. Ellipses are calculated assuming a multivariate normal distribution (R package: 'ggplot2', v3.3.3). (c) Bar-plot of the 14 Pfams with the highest contribution for the first (white, 2 Pfams), second (light blue, 1 Pfam) and third component of the PCA (blue, 11 Pfams).

microbe–microbe interactions (Fig. 6b). We assigned them to four families based on known performed function: 22 effectors, 13 hydrolytic enzymes, 12 motility functions and seven secondary metabolites.

Eighteen effectors and known virulence factors were enriched in pathogens, among which, notably, members of the HrpB gene of the type III secretion system, and its regulatory

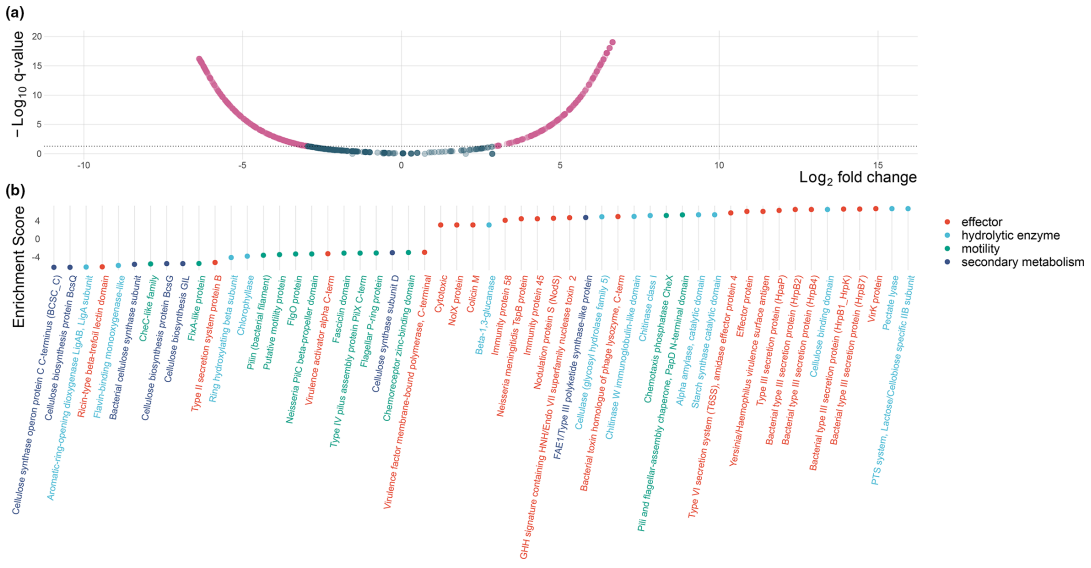

Fig. 6. (a) Volcano plot displaying all the differentially enriched Pfams of commensal (left) and pathogenic *Acidovorax* strains (right) as calculated by Anvi'o v6.2. On the *x* axis is represented Log10 fold change and on the *y* axis adjusted *q*-value. Dots are represented in red when crossing the significance threshold of 0.05. (b) Dot-plot of a curated list of 54 differentially enriched Pfams related to host–microbe or microbe–microbe interactions. The enrichment score indicated enrichment/depletion in the pathogen genomes. We assigned the Pfams to four groups according to their known functions: effectors (22 Pfams), hydrolytic enzymes (13 Pfams), motility (12 Pfams) and secondary metabolism (7 Pfams).

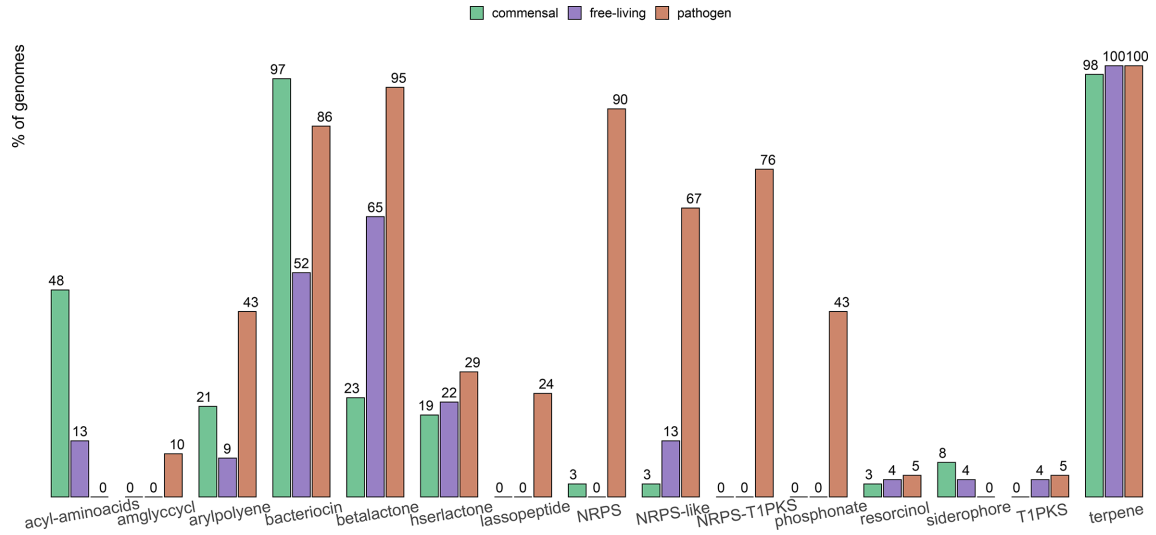

Fig. 7. Bar-plot of the prevalence in each group of 13 biosynthesis gene clusters retrieved by antiSMASH 5.0 in the genome collection. Prevalence is expressed as percentage of genomes per group carrying at least a single copy of each biosynthesis gene cluster. Genomes attributed to commensal strains are represented in green, genomes attributed to free-living and pathogenic strains are represented in purple and red, respectively.

element HpaP. Four effectors were also identified in commensals, including a type II secretion-system protein.

Among the hydrolytic enzymes, nine different Pfams were found to be enriched in pathogens, all related to plant and fungal polysaccharides degradation, including components of pectate lyases, amylases, chitinases, cellulases and glucanases. Four hydrolytic enzymes Pfams were enriched in commensals, but mainly responsible for the degradation of complex bioactive compounds, such as flavonoids, chlorophyll and aromatic-ring backbones.

Interestingly, ten Pfams related to diverse motility structures, namely flagella and pili, and chemotaxis were predominantly found in commensals. For comparison, just two motility related Pfams were enriched in pathogens.

Among those enriched in commensals, we also found six Pfams related to bacterial cellulose biosynthesis and one polyketide-synthetase Pfam enriched in pathogens.

Using AntiSMASH 5.0, we retrieved a repertoire of 15 biosynthesis gene clusters (BGCs) among the bacterial genomes (Fig. 7). The most commonly retrieved BGCs were those for bacteriocin and terpene production. Interestingly, phosphonate and all non-ribosomal peptide-synthetase BGCs (NRPS, NRPS-like, NRPS-T1PKS) were almost unique to pathogens and retrieved in more than half of the pathogenic strains.

Genotypes can predict the behaviour of the strains

We tested whether a model could predict a strain behavioural phenotype from its gene repertoire. We trained a neural network with feature extraction classifier on the gene function's presence/absence profiles of 90 % of our strains (n=96). The optimal model, which we selected according to accuracy value, registered a mean balanced accuracy=0.92 and an area under the curve (AUC)=0.98, with the selected final values of size=5 and decay=0.1. We tested the predictive power of the model against the remainder 10 % of the strains (n=10) and registered accuracy=1, sensitivity=1, specificity=1, kappa=1, 95 % confidence interval=0.69 to 1 and P=0.006. The good fit of the model could be partly explained by significant differences in the intra-group ANI (P<0.05, one-way ANOVA), with the commensal cluster being more homogeneous than both pathogens and free-living individuals (respectively, 0.79, 0.76 and 0.66 mean intra-group ANI), and by the presence of strongly correlated predictors. Thus, these results should be treated with caution until further testing can be conducted to identify the contribution of over-fitting to the reported performance.

DISCUSSION

Robust growth-promoting strains share traits, which allow a better use of phytochemicals

Our *in planta* bioassays showed a wide range of effects associated with the presence of diverse *Acidovorax* isolates obtained from healthy roots of *L. japonicus*. We compared the genomes of seven robust growth-promoters against the remainder of the isolates and retrieved 19 discriminant Pfams related to different aspects of plant–microbe and microbe–microbe interactions: chemotaxis towards root exudates, metabolism of plant secondary metabolites, antagonistic competition and transcriptional regulation.

Organic acids are a major component of root exudates [33]. Microbial transporters, such as ABC-type transport systems, TRAPs and TTTs, are essential to make use of these carbon-rich compounds and a number of studies confirmed that their presence is correlated to improved colonization in

both plant-growth promoters [34] and pathogens [35]. Many transporters are highly specific for individual substrates. For example, TTTs, which are enriched in robust-growth promoters in our study, show a higher affinity towards citrate compared to TRAPs. As shown by metabolic profiling [36], citrate is more abundant than other organic acids in *L. japonicus* root tissue. Improved sensing and uptake of citrate could translate in an evolutionary advantage of microbiota towards the colonization of this plant. Our robust growth-promoters also shared an enrichment of several features related to the metabolism of various plant-derived compounds, such as benzoic acids, carbonyls, cyanide and rare sugars. Some of these traits have been already correlated to improved plant-association in taxonomically diverse bacteria [13, 37, 38]. Furthermore, we observed the enrichment of an amidase involved in the biosynthesis of indole-3-acetic acid and of a known regulator of GABA uptake [39]. The role of auxin in plant-growth promotion by bacteria has been thoroughly characterized [40] and GABA has been more recently discovered to act as signal from plants to their associated microbiota [41, 42]. Finally, we found enriched components of diverse regulatory pathways and of both antibiotic synthesis and degradation, which could help the robust growth-promoters to not only better colonize the plant by degrading plant-derived phytochemicals with antimicrobial properties, but also to gain competitive advantage over other microbiota.

Overall, we found evidence of genomic differences among robust growth-promoting bacteria and the remainder of *Acidovorax* isolates, suggesting improved chemotaxis, competitive traits and interaction with the plant metabolism and hormonal balance, possibly leading to a stronger association between host and colonizers and explaining the better growth outcomes observed *in planta*.

The evolutionary trajectory towards pathogenic plant-association

Acidovorax strains have been naturally found to occupy widely diverse niches. In the frame of this study, we reconstructed their pan-genome to explore the importance of genomic features across the complete behavioural spectrum. Pan-genome analyses have become a staple to estimate the complete gene repertoire accessible to an organism and to understand genotype variations and evolution in a broader context [15, 43]. We found that the *Acidovorax* pan-genome is largely based on accessory or unique gene clusters (95 %) and predisposed to a continuous inflation, a predictive feature of functionally flexible organisms [44], with each occupied niche being a potential source of novel genomic traits.

Specialization has been associated with smaller genomes and previously reported in pathogenic strains [45, 46]. Therefore, we studied *Acidovorax* genome size and density, expecting to observe evidence of reductive evolution. For example, Merhej and colleagues analysed a comprehensive collection of 317 bacterial genomes to trace the evolutionary path from free-living to specialized intracellular strains [14], uncovering 'massive gene loss' as a consequence of more gene loss than horizontal gene transfer (HGT) events. Mainly in isolated niches like the interior of roots, HGT rates are considered as low due to the missing interaction with other microbiota. Surprisingly, in our study, the genomes of pathogenic *Acidovorax* were the largest, yet showed a significantly lower gene density than both free-living and commensal strains. Similar data was also described for *Rickettsia prowazekii* [45], which features more non-coding sequences than any of its close, non-pathogenic, relatives. In their review on pathogenomics, Georgiades and Raoult argued that true bacterial speciation is only observed in the contest of segregated niche, such in the case of obligated parasitism [47]. For Acidovorax, Fegan [10] listed 28 known Graminaceae hosts for *Acidovorax avenae* subsp. a*venae* and ten Cucurbitaceae natural hosts and two Solenaceae for *Acidovorax avenae* subsp. c*itrulli*. Thus, *Acidovorax* pathogens seem still far from true speciation, as suggested by their wide choice of hosts, and, likely, have already endured evolutionary driven gene losses, whereas extensive genome reductions still have to occur.

Major functional differences across the groups

Principal component analysis revealed that pathogens differ in their genomes from commensal and free-living strains. Among the differences that contributed the most to the separation, we found that the lipase C-terminal domain and the mutagenesis inducer HIM1 were both enriched in commensal strains. Assis and colleagues investigated the phylogeny of lipases in bacteria and found orthologous of the same secreted lipase ubiquitous among plant-associated strains, including *Acidovorax* [48]. Pathogens may be less dependent on lipases, as, during infection, plant-derived lipids may be of minor value to them, as they preferentially hydrolyse carbohydrates to cover their nutritional needs.

The presence of type VI secretion-systems has been strongly correlated with plant-association in Gram-negative bacteria [49] and has been in observed in pathogenic and commensal strains alike. Similarly, in *Acidovorax*, we retrieved a putative type VI secretion-system genomic island in both groups of plant-associated bacteria (commensals and pathogens, but only in 13 % of the free-living strains). For the pathogens, we also identified an operon encoding components of the type III secretion-system, absent from the other groups.

The distribution of these Pfams across the groups suggests the recruitment of commensals from the larger pool of free-living strains through horizontal gene transfer of a single genomic island encoding elements of the type VI secretion-system and further specialization of commensals into pathogenic strains, through gene losses and acquisition of type III secretion-systems.

We observed in the commensals an enrichment of several motility-associated domains. Pallen and Wren [50] postulated the loss of motility as a common side-effect of intracellular endosymbiosis, a phenomena recently witnessed in several emerging pathogens. It stands to be clarified whether this adaptation occurs for disuse, to prevent recognition by the host immune system or both. For commensals however,

which are depending on chemotaxis to make use of the plant-derived phytochemicals, motility is an essential function.

Finally, pathogens showed an enrichment of polyketide synthases and non-ribosomal peptide synthases, which have been both extensively researched due to their promising role in pharmaceutical applications for the production of diverse antimicrobial, immunosuppressive and cytostatic compounds [51, 52]. Pathogenic *Acidovorax*, thus, have access to a wider array of secondary metabolites, likely necessary for microbe–microbe competition [53] and possibly to interact with the host.

CONCLUSIONS

The focus of this research study has been to assess the role of genomic traits across the *Acidovorax* behavioural spectrum and to evaluate which genomic features can discriminate plant-pathogenic from commensal and plant-growth promoting strains. We have reported many discriminant traits through association of plant-growth data with *in silico* pan-genome-wide comparisons. The robustness of our analysis was also supported by our neural network classifier, which accurately matched each individual to its observed phenotype by using the information encoded in the pan-genome, validating our hypothesis. Researchers in the field of genomics and microbiology have been investigating the potential of machine-learning applications to extract meaningful patterns from the high-dimensional data generated by next-generation sequencing [54, 55]. Even though our method was tested on a single genus, it shows promise for the generalization and advancement of genotype-based, high-throughput phenotyping and monitoring of emerging pathogens.

However, our data is based on genomic predictions and the importance of the described Pfams, both for plant-growth-promoting commensals as well as plant-pathogens, needs to be verified firstly by analysing bacterial transcriptomes during the interaction with the plant, followed by targeted mutagenesis and analysis of the impact of bacterial mutants on survival in the rhizosphere and plant growth.

Funding information
Michael Schloter and Caroline Gutjahr were supported by grants from the Deutsche Forschungsgemeinschaft [DFG; SCHL446/38-1 and GU1423/3-1] respectively the frame of the SPP2125 'Deconstruction and reconstruction of the plant microbiota [DECRyPT]'.

Acknowledgement
We thank Ruben Garrido-Oter for making the isolates available for the *L. japonicus* bioassays and the respective sequenced genomes for the genomics comparison We highly appreciate the discussions with Martin Parniske and Corinna Dawid. We thank M.H.H. Nguyen for help with the *L. japonicus* bioassays.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Harbort CJ, Hashimoto M, Inoue H, Niu Y, Guan R, *et al*. Root-secreted coumarins and the microbiota interact to improve iron nutrition in arabidopsis. *Cell Host Microbe* 2020;28:825–837.
2. Finkel OM, Salas-González I, Castrillo G, Conway JM, Law TF. A single bacterial genus maintains root growth in a complex microbiome. *Nature* 2020;587:103–108.
3. Berg G, Grube M, Schloter M, Smalla K. The plant microbiome and its importance for plant and human health. *Front Microbiol* 2014;5:1.
4. Durán P, Thiergart T, Garrido-Oter R, Agler M, Kemen E. microbial interkingdom interactions in roots promote arabidopsis survival. *Cell* 2018;175:e14:973–983..
5. Rodriguez PA, Rothballer M, Chowdhury SP, Nussbaumer T, Gutjahr C. Systems biology of plant-microbiome interactions. *Mol Plant* 2019;12:804–821.
6. Idnurm A, Howlett BJ. Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete Leptosphaeria maculans. *Fungal Genet Biol* 2003;39:31–37.
7. Kanda A, Ohnishi S, Tomiyama H, Hasegawa H, Yasukohchi M, *et al*. Type III secretion machinery-deficient mutants of Ralstonia solanacearum lose their ability to colonize resulting in loss of pathogenicity. *J Gen Plant Pathol* 2003;69:250–257.
8. Melnyk RA, Hossain SS, Haney CH. Convergent gain and loss of genomic islands drive lifestyle changes in plant-associated Pseudomonas. *ISME J* 2019;13:1575–1588.
9. Willems A, Acidovorax GM. *Bergey's Manual of Systematics of Archaea and Bacteria*. New York: Wiley; 2015. pp. 1–16.
10. Fegan M. *Plant pathogenic members of the genera Acidovorax and Herbaspirillum*. Plant-Associated Bacteria. Springer Netherlands, 2006. pp. 671–702.
11. Giordano PR, Chaves AM, Mitkowski NA, Vargas JM. Identification, characterization, and distribution of Acidovorax avenae subsp. avenae associated with creeping bentgrass etiolation and decline. *Plant Dis* 2012;96:1736–1742.
12. Han S, Li D, Trost E, Mayer KF, Corina V. Systemic responses of barley to the 3-hydroxy-decanoyl-homoserine lactone producing plant beneficial endophyte acidovorax radicis N35. *Front Plant Sci* 2016;7.
13. Levy A, Gonzalez S, Mittelviefhaus M, Clingenpeel S, Herrera Paredes S. Genomic features of bacterial adaptation to plants. *Nat Genet* 2018;50:138–150.
14. Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D. Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biol Direct* 2009;4.
15. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. *Curr Opin Microbiol* 2015;23:148–154.
16. Eckshtain-Levi N, Munitz T, Živanović M, Traore SM, Spröer C, *et al*. Comparative analysis of type III secreted effector genes reflects divergence of Acidovorax citrulli strains into three distinct lineages. *Phytopathology* 2014;104:1152–1162.
17. Eckshtain-Levi N, Shkedy D, Gershovits M, Da Silva GM, Tamir-Ariel D, *et al*. Insights from the genome sequence of Acidovorax citrulli M6, a group I strain of the causal agent of bacterial fruit blotch of cucurbits. *Front Microbiol* 2016;7.
18. Zeng Q, Wang J, Bertels F, Giordano PR, Chilvers MI, *et al*. Recombination of virulence genes in divergent Acidovorax avenae strains that infect a common host. *Mol Plant Microbe Interact* 2017;30:813–828.
19. Wippel K, Tao K, Niu Y, Zgadzaj R, Guan R. Host preference and invasiveness of commensals in the Lotus and Arabidopsis root microbiota. *BioRxiv* 2021;2021:25–44.
20. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, *et al*. Anvi'o: an advanced analysis and visualization platform for omics data. *PeerJ* 2015;3:e1319.
21. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11.
22. Eddy SR. A new generation of homology search tools based on probabilistic inference. *Genome Inform* 2009;23:205–211.

23. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A. The Pfam protein families database in 2019. *Nucleic Acids Res* 2019;47:D427–D432.
24. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L. Metagenomic biomarker discovery and explanation. *Genome Biol* 2011;12.
25. Delmont TO, Eren EM. Linking pangenomes and metagenomes: The Prochlorococcus metapangenome. *PeerJ* 2018;6.
26. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;12:59–60.
27. Pritchard L, Cock P, Esen Ö. pyani v0.2.8: average nucleotide identity (ANI) and related measures for whole genome comparisons. *Epub ahead of print* 2019.
28. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* 2019;47:W81–W87.
29. Khalid S, Khalil T, Nasreen S. Proceedings of 2014 Science and Information Conference, SAI 2014. Institute of Electrical and Electronics Engineers Inc. In: *A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning*. 2014. pp. 372–378.
30. Conn SJ, Hocking B, Dayod M, Xu B, Athman A, *et al*. Protocol: Optimising hydroponic growth systems for nutritional and physiological analysis of arabidopsis thaliana and other plants. *Plant Methods* 2013;9:1–11.
31. Schoch CL, Ciufo S, Domrachev M. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* 2020;2020:baaa062.
32. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* 2008;11:472–477.
33. Jones DL. Organic acids in the rhizosphere - A critical review. *Plant Soil* 1998;205:25–44.
34. Kelly DJ, Thomas GH. The tripartite ATP-independent periplasmic (TRAP) transporters of bacteria and archaea. *FEMS Microbiol Rev* 2001;25:405–424.
35. Rosa LT, Bianconi ME, Thomas GH, Kelly DJ. Tripartite ATP-independent periplasmic (TRAP) transporters and Tripartite Tricarboxylate Transporters (TTT): From uptake to pathogenicity. *Front Cell Infect Microbiol* 2018;8.
36. Desbrosses GG, Kopka J, Udvardi MK. Lotus japonicus metabolic profiling. Development of gas chromatography-mass spectrometry resources for the study of plant-microbe interactions. In: *Plant Physiology*, Vol. 137. American Society of Plant Biologists, 2005. pp. 1302–1318.
37. Liu CF, Tonini L, Malaga W, Beau M, Stella A. Bacterial protein-O-mannosylating enzyme is crucial for virulence of Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* 2013;110:6560–6565.
38. Fernández-Álvarez A, Elias-Villalobos A, Ibeas JI. The O-mannosyltransferase PMT4 is essential for normal appressorium formation and penetration in Ustilago maydis. *Plant Cell* 2009;21:3397–3412.
39. Planamente S, Mondy S, Hommais F, Vigouroux A, Moréra S. Structural basis for selective GABA binding in bacterial pathogens. *Mol Microbiol* 2012;86:1085–1099.
40. Li M, Guo R, Yu F, Chen X, Zhao H, *et al*. Indole-3-acetic acid biosynthesis pathways in the plant-beneficial bacterium arthrobacter pascens ZZ21. *Int J Mol Sci* 2018;19:443.
41. Chevrot R, Rosen R, Haudecoeur E, Cirou A, Shelp BJ. GABA controls the level of quorum-sensing signal in Agrobacterium tumefaciens. *Proc Natl Acad Sci USA* 2006;103:7460–7464.
42. Park DH, Mirabella R, Bronstein PA, Preston GM, Haring MA. Mutations in γ-aminobutyric acid (GABA) transaminase genes in plants or Pseudomonas syringae reduce bacterial virulence. *Plant J* 2010;64:318–330.
43. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* 2008;11:472–477.
44. Moran NA. Microbial minimalism: Genome reduction in bacterial pathogens. *Cell* 2002;108:583–586.
45. Wixon J. Reductive evolution in bacteria: Buchnera sp., Rickettsia prowazekii and Mycobacterium leprae. In: *Comparative and Functional Genomics*, Vol. 2. 2001. pp. 44–48.
46. Andersson SGE, Zomorodipour A, Andersson JO, Sicheritz-Pontén T, Alsmark UCM, *et al*. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature* 1998;396:133–140.
47. Georgiades K, Raoult D. Defining pathogenic bacterial species in the genomic era. *Front Microbiol* 2011;1.
48. Assis R, Polloni LC, Patané JSL, Thakur S, ÉB F. Identification and analysis of seven effector protein families with different adaptive and evolutionary histories in plant-associated members of the Xanthomonadaceae. *Sci Rep* 2017;7.
49. Bernal P, Llamas MA, Filloux A. Type VI secretion systems in plant-associated bacteria. *Environ Microbiol* 2018;20:1–15.
50. Pallen MJ, Wren BW. Bacterial pathogenomics. *Nature* 2007;449:835–842.
51. Weissman KJ, Leadlay PF. Combinatorial biosynthesis of reduced polyketides. *Nat Rev Microbiol* 2005;3:925–936.
52. Grünewald J, Marahiel MA. Chemoenzymatic and template-directed synthesis of bioactive macrocyclic peptides. *Microbiol Mol Biol Rev* 2006;70:121–146.
53. Khalid S, Baccile JA, Spraker JE, Tannous J, Imran M. NRPS-derived isoquinolines and lipopetides mediate antagonism between plant pathogenic fungi and bacteria. *ACS Chem Biol* 2018;13:171–179.
54. Qu K, Guo F, Liu X, Lin Y, Zou Q. Application of machine learning in microbiology. *Front Microbiol* 2019;10:827.
55. Tarca AL, Carey VJ, wen CX, Romero R, Drăghici S. Machine learning and its applications to biology. *PLoS Comput Biol* 2007;3.