

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

Roberto Siani^{1,2}, Georg Stabl³, Caroline Gutjahr³, Michael Schlöter^{1,2}, Viviane Radl^{1*}

¹ Helmholtz Center for Environmental Health, Institute for Comparative Microbiome Analysis; Ingolstaedter Landstr. 1; 85758 Oberschleissheim, Germany

² Technical University of Munich, School of Life Sciences, Chair for Soil Science; Emil-Ramann-Strasse 2; 85354 Freising, Germany

³ Technical University of Munich, School of Life Sciences, Plant Genetics; Emil-Ramann-Strasse 4; 85354 Freising, Germany

*Corresponding author:

viviane.radl@helmholtz-muenchen.de

phone +49 89 3187 2903

Keywords:

Acidovorax, plant growth promotion, plant pathogens, *Lotus japonicus*, comparative genome analysis

19 Abstract

20 Beta-proteobacteria belonging to the genus *Acidovorax* have been described from various
 21 environments. Many strains can interact with a range of hosts, including humans and plants,
 22 forming neutral, beneficial or detrimental associations. In the frame of this study we investigated
 23 genomic properties of 52 bacterial strains of the genus *Acidovorax*, isolated from healthy roots
 24 of *Lotus japonicus*, with the intent of identifying traits important for effective plant-growth
 25 promotion. Based on single-strain inoculation bioassays with *L. japonicus*, performed in a
 26 gnotobiotic system, we could distinguish 7 robust plant-growth promoting strains from strains
 27 with no significant effects on plant-growth. We could show that the genomes of the two groups
 28 differed prominently in protein families linked to sensing and transport of organic acids,
 29 production of phytohormones, as well as resistance and production of compounds with
 30 antimicrobial properties. In a second step, we compared the genomes of the tested isolates with
 31 those of plant pathogens and free-living strains of the genus *Acidovorax* sourced from public
 32 repositories. Our pan-genomics comparison revealed features correlated with commensal and
 33 pathogenic lifestyle. We could show that commensals and pathogens differ mostly in their ability
 34 to use plant-derived lipids and in the type of secretion-systems being present. Most free-living
 35 *Acidovorax* strains did not harbour any secretion-systems. Overall, our data indicated that
 36 *Acidovorax* strains undergo extensive adaptations to their particular lifestyle by horizontal
 37 uptake of novel genetic information and loss of unnecessary genes.

38 Introduction

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

49 Bacteria of the genus *Acidovorax* belong to the group of Gram-negative beta-proteobacteria [9]
 50 and form associations with a wide range of monocotyledonous and dicotyledonous plants [10].
 51 *Acidovorax* has been mostly considered as biotrophic pathogen [11]. However, bacteria of this
 52 genus are also known as commensal species or plant beneficial bacteria, which produce
 53 secondary metabolites and hormones promoting plant growth, as well as competing with
 54 pathogens [12]. Most common species are *A. delafieldii* and *A. facilis*, whilst the most studied
 55 pathogenic species is *A. avenae*, the agent of corn (*Zea mays*), sugar cane (*Saccharum*
 56 *officinarum*) and rice (*Oryza sativa*) leaf blight, orchids brown spot disease and watermelon
 57 (*Citrullus lanatus*) fruit blotch. Thus, *Acidovorax* is a valuable model organism to investigate
 58 evolutionary patterns of plant-associated bacteria and to study which genes differentiate
 59 bacteria acting as biotrophic pathogens, as commensals or plant growth-promoting bacteria
 60 [13].

61 In the frame of this study, we compared the genomes of isolates classified as *Acidovorax* which
 62 were obtained from roots of healthy *Lotus japonicus* ecotype Gifu plants. We tested the strains
 63 for their ability to promote plant growth of *L. japonicus* in sterilized sand and correlated
 64 differences in the genomes of the strains to the observed effects on plant growth. For a broader
 65 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*.

70

71 **Material and methods**

72 **Origin of strains and genomes**

Fifty-two strains of endophytic *Acidovorax* were isolated from the root systems of healthy *Lotus 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 is described in detail elsewhere [14]. Genomes assemblies are available in the European Nucleotide Archive under the accession number PRJEB37696.

In addition, 54 other genomes were 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. radialis*, *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 Supplemental Material 1. The mean level of completion for all 106 genomes was calculated to 0.986.

Based on the available metadata, we assigned the genomes to 3 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 *Lotus japonicus* growth.

91

92 **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 [15] which computes *k*-mer frequencies and identifies coding sequences using Prodigal [16]. Afterwards, HMMER3 [17] was used to profile the genomes with hidden Markov models and to find homologous protein families in the Pfam database v33.1 [18]. Using the features occurrence frequency matrix generated by Anvi'o 6.2, we calculated the effect sizes of the features over a linear discriminant analysis (LEfSe [19]) 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 [20]. We inferred the pan-genome with the following options: DIAMOND in sensitive mode to calculate amino-acids sequences similarity, Minbit parameter to 0.5, minimum occurrence of gene clusters to 2, MCL inflation parameter to 7. We separated gene clusters into occurrence frequency classes. Kernel gene clusters were defined as present in 106 genomes. Core gene clusters were defined as present in 100 to 105 genomes. Shell gene clusters were defined as present in 16 to 99 genomes. Cloud gene clusters were defined as present in 2 to 15 genomes. Singleton gene clusters were defined as present in only 1 genome and 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 [21]. We calculated a functions occurrence frequency matrix and functions 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 3 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 value were used to assess the relevance of the features across the groups.

We retrieved secondary metabolites biosynthesis gene clusters using AntiSMASH 5.0 [22]. We launched a local instance of the program 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 are 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 employing a preliminary feature extraction step to reduce redundancy, while preserving information and decreasing computational load [23]. We trained the model over 90 % of the strains with "leave-one-out" cross-validation over 1000 repeats, using accuracy scores to select the optimal parameters. The fit of the model was evaluated by predicting labels for the remaining 10 % of the strains and again against the whole dataset. Results were assessed by calculating confusions matrix for both predictions.

146

147 *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 sulfate (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 $\mu\text{M}/\text{m}^2\text{s}$ intensity (growth cabinet PK 520-LED, Polyklima).

We transferred plantlets to pots (Göttinger 7 x 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 5 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.

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 [24] modified for arbuscular mycorrhiza (0.025 mM KH_2PO_4)). 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 $\mu\text{M}/\text{m}^2\text{s}$. 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 ($p < 0.05$; Figure 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 effects on different metrics and 10 had no significant effect on growth. From the 21 strains that resulted in improved plant-growth, 7 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.

181

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

183 To isolate genomic differences correlated with the outcomes of the *in planta* bioassays, we
184 calculated linear discriminant analysis effect sizes on the function occurrence frequency table
185 and identified 19 Pfams discriminating the 7 robust growth-promoters from the remainder of the
186 *Lotus*-isolated strains (KW $p < 0.05$, W $p < 0.05$, logarithmic LDA score > 2.0 , Figure 2; Table
187 1). Out of those, 15 Pfams were enriched in the genomes of robust growth-promoters, related to
188 4 broad functional categories: chemotaxis by sensing and uptake of organic acids and sugars
189 (tripartite tricarboxylate transporters, TTTs, and periplasmic binding proteins, PBPs),
190 colonization through synthesis and detoxification of plant secondary metabolites (aldolase, aldo-
191 keto reductase, rhodanase, amidase and mannosyltransferase), competition by synthesis,
192 secretion and detoxification of antimicrobial compounds (beta-lactamase, polyketide cyclase,
193 tetrapyrrole corrin-porphyrin methylease, dynein-related domain and dienelactone hydrolase) and
194 transcriptional regulation (FCD domain, IclR domain and sigma 70 factor). We found 4 Pfams
195 which were less abundant in the genomes of robust growth-promoters, also implicated in
196 chemotaxis (Tripartite ATP-independent periplasmic transporters, TRAPs), colonization (FliO
197 and the FG-GAP repeat) and regulation (metallo-peptidase M90).

198

199 ***Acidovorax* has an open pan-genome**

200 We reconstructed the *Acidovorax* pan-genome from 106 genomes (Figure 3), representing 15
201 out of the 23 classified *Acidovorax* species [51], along with several yet unclassified strains, and
202 capturing a large share of the genomic variability in the genus. The pan-genome contains a total
203 of 523555 genes grouped in 34146 gene clusters, of which 2 % belonged to the kernel (present

in all of the genomes), 3% belonged to the core (present in 100 to 105 genomes), 13% belonged to the shell (present in 16 to 99 of the genomes), 45 % belonged to the cloud (present in 2 to 15 genomes) and 37 % were singleton (present in one genome only). Ninety-five percent of the gene clusters were classified as accessory, while 5 % are shared by all or most of the genomes. By fitting our results to a Heaps' Power-law regression model [52] we obtained a γ of 0.48, which defines *Acidovorax* pan-genome as "open", with positive rates of novel genes discovery for every additional genome included.

Pathogenic *Acidovorax* strains have longer genomes but less 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 (Figure 4a, Supplemental Material 2). 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, $p < 0.0001$, Figure 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, $p < 0.001$, Figure 4c).

Strains separate according to behavioural phenotype

We performed a PCA on the function presence/absence matrix and found that the first 3 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 (Figure 5ab). Notably, no clustering was visible when considering the original habitat of the strains. We extracted the highest contributing variables for the first 3 components (Figure 5c, Table 2).

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 a high proportion of the second component explanatory power. 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 (Supplemental Material 3). 1243 out of 3036 Pfams were significantly enriched in 1 or 2 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 ($q < 0.05$) (Figure 6a). From these, we manually curated a list of 54 features (Table 3) putatively involved in plant-microbe and microbe-microbe interactions (Figure 6b). We assigned them to 4 families based on known performed function: 22 effectors, 13 hydrolytic enzymes, 12 motility functions and 7 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 element HpaP. Four effectors were also identified in commensals, including a type II secretion-system protein.

Among the hydrolytic enzymes, 9 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, 10 Pfams related to diverse motility structures, namely flagella and pili, and chemotaxis were predominantly found in commensals. For comparison, just 2 motility related Pfams were enriched in pathogens.

259 Among those enriched in commensals, we also found 6 Pfams related to bacterial cellulose
 260 biosynthesis and 1 polyketide-synthetase Pfams enriched in pathogens.
 261 Using AntiSMASH 5.0, we retrieved a repertoire of 15 biosynthesis gene clusters (BGCs)
 262 among the bacterial genomes (Figure 7). The most commonly retrieved BGCs were those for
 263 bacteriocin and terpene production. Interestingly, phosphonate and all non-ribosomal peptide-
 264 synthetase BGCs (NRPS, NRPS-like, NRPS-T1PKS) were almost unique to pathogens and
 265 retrieved in more than half of the pathogenic strains.

266

267 **Genotypes can predict the behaviour of the strains**

268 We tested whether a model could predict a strain behavioural phenotype from its gene
 269 repertoire. We trained a neural network with feature extraction classifier on the gene function's
 270 presence/absence profiles of 90% of our strains (n = 96). The optimal model, which we selected
 271 according to accuracy value, registered a mean balanced accuracy = 0.93 and an area under
 272 the curve (AUC) = 0.97, with the selected final values of size = 3 and decay = 0.1. We tested
 273 the predictive power of the model against the remainder 10% of the strains (n = 10) and
 274 registered accuracy = 1, sensitivity = 1, specificity = 1, kappa = 1, 95% confidence interval =
 275 0.69 to 1 and $p = 0.006$. When testing against the whole collection (n = 106), all metrics
 276 remained unchanged but for the 95% confidence interval = 0.97 to 1 and $p = 2.2e-16$.

277

278 **Discussion**

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

280 Our *in planta* bioassays showed a wide range of effects associated with the presence of diverse
 281 *Acidovorax* isolates obtained from healthy roots of *L. japonicus*. We compared the genomes of
 282 7 robust growth-promoters against the remainder of the isolates and retrieved 19 discriminant
 283 Pfams related to different aspects of plant-microbe and microbe-microbe interactions:
 284 chemotaxis towards root exudates, metabolism of plant secondary metabolites, antagonistic
 285 competition and transcriptional regulation.

Organic acids are a major component of root exudates [110]. 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 [111] and pathogens [32]. 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 [112], citrate is more abundant than other organic acids in *L. japonicus* root tissue and 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, 113, 114]. 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 [115]. The role of auxin in plant-growth promotion by bacteria has been thoroughly characterized [116] and GABA has been more recently discovered to act as signal from plants to their associated microbiota [117, 118]. 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 microbial populations.

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 are 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 whole 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 [119, 120]. We found that the *Acidovorax* pan-genome is largely made up by accessory or unique gene clusters (95 %) and predisposed to a continuous inflation, a predictive feature of functionally flexible organisms [121], 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 [121, 122]. 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 [123], 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 commensals strains. Similar data was also observed for *Rickettsia prowazekii* [124], 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 context of segregated niche, such in the case of obligated parasitism [125]. For *Acidovorax*, Fegan [10] listed 28 known Graminaceae hosts for *Acidovorax avenae* subsp. *avenae* and 10 Cucurbitaceae natural hosts and 2 Solenaceae for *Acidovorax avenae* subsp. *citrulli*. Thus, *Acidovorax* pathogens seems 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* [126]. 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 [127] and has been 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 [128] 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 [129, 130]. Pathogenic *Acidovorax*, thus, have access to a wider array of secondary metabolites, likely necessary for microbe-microbe competition [131] 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 [132, 133, 134, 135]. Even though our method was tested on a single genus, it shows promises for generalization and the 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.

Authors statement

The authors declare no financial conflicts of interest.

Acknowledgement

Michael Schlöter 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]”. 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.

Bibliography

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.e6. doi: 10.1016/j.chom.2020.09.006
2. Finkel OM, Salas-González I, Castrillo G, Conway JM, Law TF, et al. A single bacterial genus maintains root growth in a complex microbiome. *Nature* 2020;587:103–108. doi: 10.1038/s41586-020-2778-7
3. Berg G, Grube M, Schlöter M, Smalla K. The plant microbiome and its importance for plant and human health. *Front Microbiol* 2014;5:1. doi: 10.3389/fmicb.2014.00491
4. Durán P, Thiergart T, Garrido-Oter R, Agler M, Kemen E, et al. Microbial Interkingdom Interactions in Roots Promote Arabidopsis Survival. *Cell* 2018;175:973-983.e14. doi: 10.1016/j.cell.2018.10.020
5. Rodriguez PA, Rothballer M, Chowdhury SP, Nussbaumer T, Gutjahr C, et al. Systems Biology of Plant-Microbiome Interactions. *Mol Plant* 2019;12:804–821. doi: 10.1016/j.molp.2019.05.006
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. doi: 10.1016/S1087-1845(02)00588-1
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. doi: 10.1016/S1087-1845(02)00588-1

423 8. Melnyk RA, Hossain SS, Haney CH. Convergent gain and loss of genomic islands drive lifestyle
424 changes in plant-associated *Pseudomonas*. ISME J 2019;13:1575–1588. doi: 10.1016/S1087-
425 1845(02)00588-1

426 9. Willems A, Gillis M. *Acidovorax*. In: Bergey's Manual of Systematics of Archaea and Bacteria.
427 John Wiley & Sons, Ltd; 2015. pp. 1–16. doi: 10.1099/00207713-42-1-107

428 10. Fegan M. Plant pathogenic members of the genera *Acidovorax* and *Herbaspirillum*. In: Plant-
429 Associated Bacteria. Springer Netherlands; 2006. pp. 671–702. doi: 10.1007/978-1-4020-4538-
430 7_18

431 11. Giordano PR, Chaves AM, Mitkowski NA, Vargas JM. Identification, characterization, and
432 distribution of *Acidovorax avenae* subsp. *avenae* associated with creeping bentgrass etiolation
433 and decline. Plant Dis 2012;96:1736–1742. doi: 10.1007/978-1-4020-4538-7_18

434 12. Han S, Li D, Trost E, Mayer KF, Corina V, et al. Systemic responses of barley to the 3-
435 hydroxy-decanoyl-homoserine lactone producing plant beneficial endophyte *acidovorax radialis*
436 N35. Front Plant Sci 2016;7:1868. doi: 10.3389/fpls.2016.01868

437 13. Levy A, Salas Gonzalez I, Mittelviefhaus M, Clingenpeel S, Herrera Paredes S, et al. Genomic
438 features of bacterial adaptation to plants. Nat Genet 2018;50:138–150. doi: 10.1038/s41588-017-
439 0012-9

440 14. Wippel K, Tao K, Niu Y, Zgadzaj R, Guan R, et al. Host preference and invasiveness of
441 commensals in the Lotus and Arabidopsis root microbiota. BioRxiv 2021;2021.01.12.426357. doi:
442 10.1101/2021.01.12.426357

443 15. Eren AM, Esen OC, Quince C, Vineis JH, Morrison HG, et al. Anvi'o: An advanced analysis
444 and visualization platform for 'omics data. PeerJ 2015;2015:e1319. doi: 10.7717/peerj.1319

445 16. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, et al. Prodigal: Prokaryotic gene
446 recognition and translation initiation site identification. BMC Bioinformatics 2010;11:119. doi:
447 10.1186/1471-2105-11-119

448 17. Eddy SR. A new generation of homology search tools based on probabilistic inference.
449 Genome Inform 2009;23:205–211. doi: 10.1142/9781848165632_0019

450 18. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, et al. The Pfam protein families
451 database in 2019. *Nucleic Acids Res* 2019;47:D427–D432. doi: 10.1093/nar/gky995

452 19. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, et al. Metagenomic biomarker
453 discovery and explanation. *Genome Biol* 2011;12:R60. doi: 10.1186/gb-2011-12-6-r60

454 20. Delmont TO, Eren EM. Linking pangenomes and metagenomes: The *Prochlorococcus*
455 metapangenome. *PeerJ* 2018;2018:e4320. doi: 10.7717/peerj.4320

456 21. Pritchard L, Cock P, Esen Ö. pyani v0.2.8: average nucleotide identity (ANI) and related
457 measures for whole genome comparisons. Epub ahead of print 5 March 2019. doi:
458 10.5281/ZENODO.2584238.

459 22. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, et al. antiSMASH 5.0: updates to the
460 secondary metabolite genome mining pipeline. *Nucleic Acids Res* 2019;47:W81–W87. doi:
461 10.1093/nar/gkz310

462 23. Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques
463 in machine learning. In: *Proceedings of 2014 Science and Information Conference, SAI 2014*.
464 Institute of Electrical and Electronics Engineers Inc.; 2014. pp. 372–378. doi:
465 10.1109/SAI.2014.6918213

466 24. Conn SJ, Hocking B, Dayod M, Xu B, Athman A, et al. Protocol: Optimising hydroponic growth
467 systems for nutritional and physiological analysis of *Arabidopsis thaliana* and other plants. *Plant*
468 *Methods* 2013;9:1–11. doi: 10.1186/1746-4811-9-4

469 25. Kelly DJ, Thomas GH. The tripartite ATP-independent periplasmic (TRAP) transporters of
470 bacteria and archaea. *FEMS Microbiol Rev* 2001;25:405–424. doi: 10.1111/j.1574-
471 6976.2001.tb00584.x

472 26. Simpson DA, Ramphal R, Lory S. Characterization of *Pseudomonas aeruginosa* fliO, a gene
473 involved in flagellar biosynthesis and adherence. *Infect Immun* 1995;63:2950–2957. doi:
474 10.1128/iai.63.8.2950-2957.1995

475 27. Barker CS, Meshcheryakova I V., Kostyukova AS, Samatey FA. FliO Regulation of FlhP in the
476 Formation of the *Salmonella enterica* Flagellum. *PLoS Genet* 2010;6:e1001143. doi:
477 10.1371/journal.pgen.1001143

28. Absalon C, Van Dellen K, Watnick PI. A Communal Bacterial Adhesin Anchors Biofilm and Bystander Cells to Surfaces. *PLoS Pathog* 2011;7:e1002210. doi: 10.1371/journal.ppat.1002210

29. Göhler AK, Staab A, Gabor E, Homann K, Klang E, et al. Characterization of MtfA, a novel regulatory output signal protein of the glucose-phosphotransferase system in *Escherichia coli* K-12. *J Bacteriol* 2012;194:1024–1035. doi: 10.1128/JB.06387-11

30. Xu Q, Göhler AK, Kosfeld A, Carlton D, Chiu HJ, et al. The structure of MLC titration factor a (MtfA/Yeel) reveals a Prototypical zinc metallopeptidase related to anthrax lethal factor. *J Bacteriol* 2012;194:2987–2999. doi: 10.1128/JB.00038-12

31. Rosa LT, Bianconi ME, Thomas GH, Kelly DJ. Tripartite ATP-independent periplasmic (TRAP) transporters and Tripartite Tricarboxylate Transporters (TTT): From uptake to pathogenicity. *Frontiers in Cellular and Infection Microbiology* 2018;8:33. doi: 10.3389/fcimb.2018.00033

32. Cangelosi GA, Ankenbauer RG, Nestert EW. Sugars induce the *Agrobacterium* virulence genes through a periplasmic binding protein and a transmembrane signal protein (crown gap1/signal transduction/ChvE/VirA/galactose-binding proteins). 1990;87:6708-6712 doi: <http://www.jstor.org/stable/2355369>.

33. Rodionova IA, Li X, Thiel V, Stolyar S, Stanton K, et al. Comparative genomics and functional analysis of rhamnose catabolic pathways and regulons in bacteria. *Front Microbiol* 2013;4:407. doi: 10.3389/fmicb.2013.00407

34. 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* 2011;286:6999–7009. doi: 10.1074/jbc.M110.202226

35. Cipollone R, Ascenzi P, Tomao P, Imperi F, Visca P. Enzymatic detoxification of cyanide: Clues from *Pseudomonas aeruginosa* rhodanese. *J Mol Microbiol Biotechnol* 2008;15:199–211. doi: 10.1159/000121331

36. Pollmann S, Müller A, Weiler EW. Many Roads Lead to ‘Auxin’: of Nitrilases, Synthases, and Amidases. *Plant Biol* 2006;8:326–333. doi: 10.1055/s-2006-924075

504 37. Lengeler KB, Tielker D, Ernst JF. Protein-O-mannosyltransferases in virulence and
505 development. *Cellular and Molecular Life Sciences* 2008;65:528–544. doi: 10.1007/s00018-007-
506 7409-z

507 38. Zowawi HM, Harris PNA, Roberts MJ, Tambyah PA, Schembri MA, et al. The emerging threat
508 of multidrug-resistant Gram-negative bacteria in urology. *Nat Rev Urol* 2015;12:570–584. doi:
509 10.1038/nrurol.2015.199

510 39. van den Bosch TJM, Tan K, Joachimiak A, Welte CU. Functional profiling and crystal
511 structures of isothiocyanate hydrolases found in gut-associated and plant-pathogenic bacteria.
512 *Appl Environ Microbiol* 2018;84:2020. doi: 10.1128/AEM.00478-18

513 40. Sultana A, Kallio P, Jansson A, Wang JS, Niemi J, et al. Structure of the polyketide cyclase
514 SnoaL reveals a novel mechanism for enzymatic aldol condensation. *EMBO J* 2004;23:1911–
515 1921. doi: 10.1038/sj.emboj.7600201

516 41. Raux E, Schubert HL, Warren MJ. Biosynthesis of cobalamin (vitamin B12): A bacterial
517 conundrum. 2000. Epub ahead of print 2000. doi: 10.1007/PL00000670.

518 42. Schoenhofen IC, Li G, Strozen TG, Howard SP. Purification and characterization of the N-
519 terminal domain of ExeA: A novel ATPase involved in the type II secretion pathway of *Aeromonas*
520 *hydrophila*. *J Bacteriol* 2005;187:6370–6378. doi: 10.1128/JB.187.18.6370-6378.2005

521 43. Schlomann M, Schmidt E, Knackmuss HJ. Different types of diene lactone hydrolase in 4-
522 fluorobenzoate-utilizing bacteria. *J Bacteriol* 1990;172:5112–5118. doi: 10.1128/jb.172.9.5112-
523 5118.1990

524 44. Schlomann M, Ngai KL, Ornston LN, Knackmuss HJ. Diene lactone hydrolase from
525 *Pseudomonas cepacia*. *J Bacteriol* 1993;175:2994–3001. doi: 10.1128/jb.175.10.2994-3001.1993

526 45. Lord DM, Uzgoren Baran A, Soo VWC, Wood TK, Peti W, et al. McbR/YncC: Implications for
527 the mechanism of ligand and DNA binding by a bacterial gntR transcriptional regulator involved in
528 biofilm formation. *Biochemistry* 2014;53:7223–7231. doi: 10.1021/bi500871a

529 46. Planamente S, Mondy S, Hommais F, Vigouroux A, Moréra S, et al. Structural basis for
530 selective GABA binding in bacterial pathogens. *Mol Microbiol* 2012;86:1085–1099. doi:
531 10.1111/mmi.12043

532 47. Thomson NR, Nasser W, McGowan S, Sebahia M, Salmond GPC. *Erwinia carotovora* has two
533 KdgR-like proteins belonging to the IclR family of transcriptional regulators: Identification and
534 characterization of the RexZ activator and the KdgR repressor of pathogenesis. *Microbiology*
535 1999;145:1531–1545. doi: 10.1099/13500872-145-7-1531

536 48. Molina-Henares AJ, Krell T, Eugenia Guazzaroni M, Segura A, Ramos JL. Members of the
537 IclR family of bacterial transcriptional regulators function as activators and/or repressors. *FEMS*
538 *Microbiology Reviews* 2006;30:157–186. doi: 10.1111/j.1574-6976.2005.00008.x

539 49. Paget MSB, Helmann JD. The $\sigma 70$ family of sigma factors. *Genome Biology* 2003;4:203. doi:
540 10.1186/gb-2003-4-1-203

541 50. Kazmierczak MJ, Wiedmann M, Boor KJ. Alternative Sigma Factors and Their Roles in
542 Bacterial Virulence. *Microbiol Mol Biol Rev* 2005;69:527–543. doi: 10.1128/mmbr.69.4.527-
543 543.2005

544 51. <https://www.ncbi.nlm.nih.gov/taxonomy>

545 52. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome.
546 *Current Opinion in Microbiology* 2008;11:472–477. doi: 10.1016/j.mib.2008.09.006

547 53. Chen CKM, Lee GC, Ko TP, Guo RT, Huang LM, et al. Structure of the
548 Alkalohyperthermophilic *Archaeoglobus fulgidus* Lipase Contains a Unique C-Terminal Domain
549 Essential for Long-Chain Substrate Binding. *J Mol Biol* 2009;390:672–685. doi:
550 10.1016/j.jmb.2009.05.017

551 54. Flaugnatti N, Le TTH, Canaan S, Aschtgen MS, Nguyen VS, et al. A phospholipase A1
552 antibacterial Type VI secretion effector interacts directly with the C-terminal domain of the VgrG
553 spike protein for delivery. *Mol Microbiol* 2016;99:1099–1118. doi: 10.1111/mmi.13292

554 55. Kelberg EP, Kovaltsova S V., Alekseev SY, Fedorova I V., Gracheva LM, et al. HIM1, a new
555 yeast *Saccharomyces cerevisiae* gene playing a role in control of spontaneous and induced
556 mutagenesis. *Mutat Res - Fundam Mol Mech Mutagen* 2005;578:64–78. doi:
557 10.1016/j.mrfmmm.2005.03.003

558 56. Karlowski WM, Zielezinski A, Carrère J, Pontier D, Lagrange T, et al. Genome-wide
559 computational identification of WG/GW Argonaute-binding proteins in Arabidopsis. *Nucleic Acids*
560 *Res* 2010;38:4231–4245. doi: 10.1093/nar/gkq162

561 57. Filloux A, Hachani A, Bleves S. The bacterial type VI secretion machine: yet another player for
562 protein transport across membranes. *Microbiology* 2008;154:1570–1583. doi:
563 10.1099/mic.0.2008/016840-0

564 58. Kapitein N, Bönemann G, Pietrosiuk A, Seyffer F, Hausser I, et al. ClpV recycles VipA/VipB
565 tubules and prevents non-productive tubule formation to ensure efficient type VI protein secretion.
566 *Mol Microbiol* 2013;87:1013–1028. doi: 10.1111/mmi.12147

567 59. English G, Byron O, Cianfanelli FR, Prescott AR, Coulthurst SJ. Biochemical analysis of TssK,
568 a core component of the bacterial Type VI secretion system, reveals distinct oligomeric states of
569 TssK and identifies a TssK-TssFG subcomplex. *Biochem J* 2014;461:291–304. doi:
570 10.1042/BJ20131426

571 60. Bingle LE, Bailey CM, Pallen MJ. Type VI secretion: a beginner's guide. *Current Opinion in*
572 *Microbiology* 2008;11:3–8. doi: 10.1016/j.mib.2008.01.006

573 61. Mintz KP, Fives-Taylor PM. *impA*, a gene coding for an inner membrane protein, influences
574 colonial morphology of *Actinobacillus actinomycetemcomitans*. *Infect Immun* 2000;68:6580–6586.
575 doi: 10.1128/IAI.68.12.6580-6586.2000

576 62. Suarez G, Sierra JC, Sha J, Wang S, Erova TE, et al. Molecular characterization of a
577 functional type VI secretion system from a clinical isolate of *Aeromonas hydrophila*. *Microb Pathog*
578 2008;44:344–361. doi: 10.1016/j.micpath.2007.10.005

579 63. Bladergroen MR, Badelt K, Spaink HP. Infection-blocking genes of a symbiotic *Rhizobium*
580 *leguminosarum* strain that are involved in temperature-dependent protein secretion. *Mol Plant-*
581 *Microbe Interact* 2003;16:53–64. doi: 10.1094/MPMI.2003.16.1.53

582 63. Zhou Y, Tao J, Yu H, Ni J, Zeng L, et al. Hcp family proteins secreted via the type VI secretion
583 system coordinately regulate *Escherichia coli* K1 interaction with human brain microvascular
584 endothelial cells. *Infect Immun* 2012;80:1243–1251. doi: 10.1128/IAI.05994-11

585 64. Cunnac S, Boucher C, Genin S. Characterization of the cis-Acting Regulatory Element
586 Controlling HrpB-Mediated Activation of the Type III Secretion System and Effector Genes in
587 *Ralstonia solanacearum*. J Bacteriol 2004;186:2309–2318. doi: 10.1128/JB.186.8.2309-2318.2004
588 65. Furutani A, Takaoka M, Sanada H, Noguchi Y, Oku T, et al. Identification of novel type III
589 secretion effectors in *Xanthomonas oryzae* pv. *oryzae*. Mol Plant-Microbe Interact 2009;22:96–
590 106. doi: 10.1094/MPMI-22-1-0096
591 66. van Gijsegem F, Gough C, Zischek C, Niqueux E, Arlat M, et al. The hrp gene locus of
592 *Pseudomonas solanacearum*, which controls the production of a type III secretion system,
593 encodes eight proteins related to components of the bacterial flagellar biogenesis complex. Mol
594 Microbiol 1995;15:1095–1114. doi: 10.1111/j.1365-2958.1995.tb02284.x
595 67. Baruch K, Gur-Arie L, Nadler C, Koby S, Yerushalmi G, et al. Metalloprotease type III effectors
596 that specifically cleave JNK and NF- κ B. EMBO J 2011;30:221–231. doi: 10.1038/emboj.2010.297
597 68. Zhang H, Gao ZQ, Wang WJ, Liu GF, Xu JH, et al. Structure of the type vi effector-immunity
598 complex (Tae4-Tai4) provides novel insights into the inhibition mechanism of the effector by its
599 immunity protein. J Biol Chem 2013;288:5928–5939. doi: 10.1074/jbc.M112.434357
600 69. López-Lara IM, Kafetzopoulos D, Spaink HP, Thomas-Oates JE. Rhizobial NodL O-acetyl
601 transferase and nodS N-methyl transferase functionally interfere in production of modified nod
602 factors. J Bacteriol 2001;183:3408–3416. doi: 10.1128/JB.183.11.3408-3416.2001
603 70. Müller MG, Moe NE, Richards PQ, Moe GR. Resistance of *Neisseria meningitidis* to human
604 serum depends on T and B cell stimulating protein B. Infect Immun 2015;83:1257–1264. doi:
605 10.1128/IAI.03134-14
606 71. Carr S, Walker D, James R, Kleanthous C, Hemmings AM. Inhibition of a ribosome-
607 inactivating ribonuclease: The crystal structure of the cytotoxic domain of colicin E3 in complex
608 with its immunity protein. Structure 2000;8:949–960. doi: 10.1016/S0969-2126(00)00186-6
609 72. Krishnan HB. NolX of *Sinorhizobium fredii* USDA257, a type III-secreted protein involved in
610 host range determination, is localized in the infection threads of cowpea (*Vigna unguiculata* [L.]
611 Walp) and soybean (*Glycine max* [L.] Merr.) nodules. J Bacteriol 2002;184:831–839. doi: 10.1128/
612 JB.184.3.831-839.2002

73. El Ghachi M, Bouhss A, Barreteau H, Touzé T, Auger G, et al. Colicin M exerts its bacteriolytic effect via enzymatic degradation of undecaprenyl phosphate-linked peptidoglycan precursors. *J Biol Chem* 2006;281:22761–22772. doi: 10.1074/jbc.M602834200

74. Hattori Y, Iwata K, Suzuki K, Uraji M, Ohta N, et al. Sequence characterization of the vir region of a nopaline type Ti plasmid, pTi-SAKURA. *Genes Genet Syst* 2001;76:121–130. doi: 10.1266/ggs.76.121

75. Fields KA, Nilles ML, Cowan C, Straley SC. Virulence Role of V Antigen of *Yersinia pestis* at the Bacterial Surface. *Infect Immun* 1999;67:5395–5408. doi: 0019-9567/99/\$04.000

76. Zhang D, de Souza RF, Anantharaman V, Iyer LM, Aravind L. Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. *Biol Direct*;7. Epub ahead of print 25 June 2012. doi: 10.1186/1745-6150-7-18.

77. Patzer SI, Albrecht R, Braun V, Zeth K. Structural and mechanistic studies of pesticin, a bacterial homolog of phage lysozymes. *J Biol Chem* 2012;287:23381–23396. doi: 10.1074/jbc.M112.362913

78. Guyon K, Balagué C, Roby D, Raffaele S. Secretome analysis reveals effector candidates associated with broad host range necrotrophy in the fungal plant pathogen *Sclerotinia sclerotiorum*. *BMC Genomics* 2014;15:1–19. doi: 10.1186/1471-2164-15-336

79. Korotkov K V., Sandkvist M, Hol WGJ. The type II secretion system: Biogenesis, molecular architecture and mechanism. *Nat Rev Microbiol* 2012;10:336–351. doi: 10.1038/nrmicro2762

80. Lohrke SM, Nechaev S, Yang H, Severinov K, Jin SJ. Transcriptional activation of *Agrobacterium tumefaciens* virulence gene promoters in *Escherichia coli* requires the *A. tumefaciens* rpoA gene, encoding the alpha subunit of RNA polymerase. *J Bacteriol* 1999;181:4533–4539. doi: 10.1128/jb.181.15.4533-4539.1999

81. Kadioglu A, Weiser JN, Paton JC, Andrew PW. The role of *Streptococcus pneumoniae* virulence factors in host respiratory colonization and disease. *Nat Rev Microbiol* 2008;6:288–301. doi: 10.1038/nrmicro1871

- 640 82. Charkowski AO, Alfano JR, Preston G, Yuan J, He SY, et al. The *Pseudomonas syringae* pv.
641 tomato HrpW protein has domains similar to harpins and pectate lyases and can elicit the plant
642 hypersensitive response and bind to pectate. *J Bacteriol* 1998;180:5211–5217. doi:
643 10.1128/jb.180.19.5211-5217.1998
- 644 83. Kotrba P, Inui M, Yukawa H. Bacterial phosphotransferase system (PTS) in carbohydrate
645 uptake and control of carbon metabolism. *Journal of Bioscience and Bioengineering* 2001;92:502–
646 517. doi: 10.1016/S1389-1723(01)80308-X
- 647 84. Johnson PE, Joshi MD, Tomme P, Kilburn DG, McIntosh LP. Structure of the N-terminal
648 cellulose-binding domain of *Cellulomonas fimi* CenC determined by nuclear magnetic resonance
649 spectroscopy. *Biochemistry* 1996;35:14381–14394. doi: 10.1021/bi961612s
- 650 85. Larson SB, Greenwood A, Cascio D, Day J, McPherson A. Refined molecular structure of pig
651 pancreatic α -amylase at 2.1 Å resolution. *J Mol Biol* 1994;235:1560–1584. doi:
652 10.1006/jmbi.1994.1107
- 653 86. Rodríguez-Sanoja R, Oviedo N, Sánchez S. Microbial starch-binding domain. *Current Opinion*
654 *in Microbiology* 2005;8:260–267. doi: 10.1016/j.mib.2005.04.013
- 655 87. Gijzen M, Kuflu K, Qutob D, Chernys JT. A class I chitinase from soybean seed coat. *J Exp*
656 *Bot* 2001;52:2283–2289. doi: 10.1093/jexbot/52.365.2283
- 657 88. Itoh T, Hibi T, Suzuki F, Sugimoto I, Fujiwara A, et al. Crystal structure of chitinase ChiW from
658 *Paenibacillus* sp. str. FPU-7 reveals a novel type of bacterial cell-surface-expressed multi-modular
659 enzyme machinery. *PLoS One*;11. Epub ahead of print 1 December 2016. doi:
660 10.1371/journal.pone.0167310.
- 661 89. Py B, Isabelle BG, Haiech J, Chippaux M, Barras F. Cellulase egz of *Erwinia chrysanthemi*:
662 Structural organization and importance of his98 and glu133 residues for catalysis. *Protein Eng*
663 *Des Sel* 1991;4:325–333. doi: 10.1093/protein/4.3.325
- 664 90. Palumbo JD, Sullivan RF, Kobayashi DY. Molecular characterization and expression in
665 *Escherichia coli* of three β -1,3-glucanase genes from *Lysobacter enzymogenes* strain N4-7. *J*
666 *Bacteriol* 2003;185:4362–4370. doi: 10.1128/JB.185.15.4362-4370.2003

667 91. Sugimoto K, Senda T, Aoshima H, Masai E, Fukuda M, et al. Crystal structure of an aromatic
668 ring opening dioxygenase LigAB, a protocatechuate 4,5-dioxygenase, under aerobic conditions.
669 Structure 1999;7:953–965. doi: 10.1016/S0969-2126(99)80122-1

670 92. Stehr M, Diekmann H, Smau L, Seth O, Ghisla S, et al. A hydrophobic sequence motif
671 common to N-hydroxylating enzymes. Trends Biochem Sci 1998;23:56–57. doi: 10.1016/S0968-
672 0004(97)01166-3

673 93. Kauppi B, Lee K, Carredano E, Parales RE, Gibson DT, et al. Structure of an aromatic-ring-
674 hydroxylating dioxygenasenaphthalene 1,2-dioxygenase. Structure 1998;6:571–586. doi: 10.1016/
675 S0969-2126(98)00059-8

676 94. Tsuchiya T, Ohta H, Okawa K, Iwamatsu A, Shimada H, et al. Cloning of chlorophyllase, the
677 key enzyme in chlorophyll degradation: Finding of a lipase motif and the induction by methyl
678 jasmonate. Proc Natl Acad Sci U S A 1999;96:15362–15367. doi: 10.1073/pnas.96.26.15362

679 95. Park SY, Chao X, Gonzalez-Bonet G, Beel BD, Bilwes AM, et al. Structure and function of an
680 unusual family of protein phosphatases: The bacterial chemotaxis proteins CheC and CheX. Mol
681 Cell 2004;16:563–574. doi: 10.1016/j.molcel.2004.10.018

682 96. Holmgren A, Bränden CL. Crystal structure of chaperone protein PapD reveals an
683 immunoglobulin fold. Nature 1989;342:248–251. doi: 10.1038/342248a0

684 97. Draper J, Karplus K, Ottemann KM. Identification of a chemoreceptor zinc-binding domain
685 common to cytoplasmic bacterial chemoreceptors. J Bacteriol 2011;193:4338–4345. doi: 10.1128/
686 JB.05140-11

687 98. Szurmant H, Ordal GW. Diversity in Chemotaxis Mechanisms among the Bacteria and
688 Archaea. Microbiol Mol Biol Rev 2004;68:301–319. doi: 10.1128/mmbr.68.2.301-319.2004

689 99. Ide N, Kutsukake K. Identification of a novel Escherichia coli gene whose expression is
690 dependent on the flagellum-specific sigma factor, FliA, but dispensable for motility development.
691 Gene 1997;199:19–23. doi: 10.1016/S0378-1119(97)00233-3

692 100. Pelicic V. Type IV pili: e pluribus unum? Mol Microbiol 2008;68:827–837. doi: 10.1111/j.1365-
693 2958.2008.06197.x

694 101. Serizawa M, Yamamoto H, Yamaguchi H, Fujita Y, Kobayashi K, et al. Systematic analysis of
695 SigD-regulated genes in *Bacillus subtilis* by DNA microarray and Northern blotting analyses. *Gene*
696 2004;329:125–136. doi: 10.1016/j.gene.2003.12.024

697 102. Martinez RM, Dharmasena MN, Kirn TJ, Taylor RK. Characterization of two outer membrane
698 proteins, FlgO and FlgP, that influence *Vibrio cholerae* motility. *J Bacteriol* 2009;191:5669–5679.
699 doi: 10.1128/JB.00632-09

700 103. Bäckman M, Källström H, Jonsson AB. The phase-variable pilus-associated protein PilC is
701 commonly expressed in clinical isolates of *Neisseria gonorrhoeae*, and shows sequence variability
702 among strains. *Microbiology* 1998;144:149–156. doi: 10.1099/00221287-144-1-149

703 104. Kawamoto T, Noshiro M, Shen M, Nakamasu K, Hashimoto K, et al. Structural and
704 phylogenetic analyses of RGD-CAP/βig-h3, a fasciclin- like adhesion protein expressed in chick
705 chondrocytes. *Biochim Biophys Acta - Gene Struct Expr* 1998;1395:288–292. doi: 10.1016/S0167-
706 4781(97)00172-3

707 105. Alm RA, Hallinan JP, Watson AA, Mattick JS. Fimbrial biogenesis genes of *Pseudomonas*
708 *aeruginosa*: pilW and pilX increase the similarity of type 4 fimbriae to the GSP protein-secretion
709 systems and pilY1 encodes a gonococcal PilC homologue. *Mol Microbiol* 1996;22:161–173. doi:
710 10.1111/j.1365-2958.1996.tb02665.x

711 106. Nambu T, Kutsukake K. The *Salmonella* FlgA protein, a putative periplasmic chaperone
712 essential for flagellar P ring formation. *Microbiology* 2000;146:1171–1178. doi: 10.1099/00221287-
713 146-5-1171

714 107. Funa N, Ohnishi Y, Ebizuka Y, Horinouchi S. Alteration of reaction and substrate specificity of
715 a bacterial type III polyketide synthase by site-directed mutagenesis. *Biochem J* 2002;367:781–
716 789. doi: 10.1042/BJ20020953

717 108. Ross P, Mayer R, Benziman M. Cellulose biosynthesis and function in bacteria. *Microbiol*
718 *Rev*;55. Epub ahead of print 1991. doi: 10.1128/mmbr.55.1.35-58.1991.

719 109. Matthysse AG. Role of bacterial cellulose fibrils in *Agrobacterium tumefaciens* infection. *J*
720 *Bacteriol* 1983;154:906–915. doi: 10.1128/jb.154.2.906-915.1983

110. Jones DL. Organic acids in the rhizosphere - A critical review. *Plant Soil* 1998;205:25–44.
doi: 10.1023/A:1004356007312
111. de Souza RSC, Armanhi JSL, Damasceno N de B, Imperial J, Arruda P. Genome Sequences of a Plant Beneficial Synthetic Bacterial Community Reveal Genetic Features for Successful Plant Colonization. *Front Microbiol* 2019;10:1779. doi: 10.3389/fmicb.2019.01779
112. 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*. American Society of Plant Biologists; 2005. pp. 1302–1318. doi: 10.1104/pp.104.054957
113. Liu CF, Tonini L, Malaga W, Beau M, Stella A, et al. Bacterial protein-O-mannosylating enzyme is crucial for virulence of Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* 2013;110:6560–6565. doi: 10.1073/pnas.1219704110
114. 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. doi: 10.1105/tpc.109.065839
115. Planamente S, Mondy S, Hommais F, Vigouroux A, Moréra S, et al. Structural basis for selective GABA binding in bacterial pathogens. *Mol Microbiol* 2012;86:1085–1099. doi: 10.1111/mmi.12043
116. 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. doi: 10.3390/ijms19020443
117. Chevrot R, Rosen R, Haudecoeur E, Cirou A, Shelp BJ, et al. GABA controls the level of quorum-sensing signal in Agrobacterium tumefaciens. *Proc Natl Acad Sci U S A* 2006;103:7460–7464. doi: 10.1073/pnas.0600313103
118. Park DH, Mirabella R, Bronstein PA, Preston GM, Haring MA, et al. Mutations in γ -aminobutyric acid (GABA) transaminase genes in plants or Pseudomonas syringae reduce bacterial virulence. *Plant J* 2010;64:318–330. doi: 10.1111/j.1365-313X.2010.04327.x

748 119. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Current
749 Opinion in Microbiology 2015;23:148–154. doi: 10.1016/j.mib.2014.11.016

750 120. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome.
751 Current Opinion in Microbiology 2008;11:472–477. doi: 10.1016/j.gde.2005.09.006

752 121. Moran NA. Microbial minimalism: Genome reduction in bacterial pathogens. Cell
753 2002;108:583–586. doi: 10.1016/S0092-8674(02)00665-7

754 122. Wixon J. Reductive evolution in bacteria: *Buchnera* sp., *Rickettsia prowazekii* and
755 *Mycobacterium leprae*. 2001. Epub ahead of print 2001. doi: 10.1002/cfg.70.

756 123. Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D. Massive comparative genomic analysis
757 reveals convergent evolution of specialized bacteria. Biol Direct 2009;4:13. doi: 10.1186/1745-
758 6150-4-13

759 124. Andersson SGE, Zomorodipour A, Andersson JO, Sicheritz-Ponté T, Cecilia U, et al. The
760 genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. 1998;396. doi:
761 10.1038/2409

762 125. Georgiades K, Raoult D. Defining pathogenic bacterial species in the genomic era. Front
763 Microbiol 2011;1:151. doi: 10.3389/fmicb.2010.00151

764 126. Assis RDAB, Polloni LC, Patané JSL, Thakur S, Felestrino ÉB, et al. Identification and
765 analysis of seven effector protein families with different adaptive and evolutionary histories in
766 plant-associated members of the Xanthomonadaceae. Sci Rep;7. Epub ahead of print 1
767 December 2017. doi: 10.1038/s41598-017-16325-1.

768 127. Bernal P, Llamas MA, Filloux A. Type VI secretion systems in plant-associated bacteria.
769 Environ Microbiol 2018;20:1–15. doi: 10.1111/1462-2920.13956

770 128. Pallen MJ, Wren BW. Bacterial pathogenomics. Nature 2007;449:835–842. doi:
771 10.1038/nature06248

772 129. Weissman KJ, Leadlay PF. Combinatorial biosynthesis of reduced polyketides. Nature
773 Reviews Microbiology 2005;3:925–936. doi: 10.1038/nrmicro1287

774 130. Grünewald J, Marahiel MA. Chemoenzymatic and Template-Directed Synthesis of Bioactive
775 Macrocyclic Peptides. *Microbiol Mol Biol Rev* 2006;70:121–146. doi: 10.1128/mmbr.70.1.121-
776 146.2006

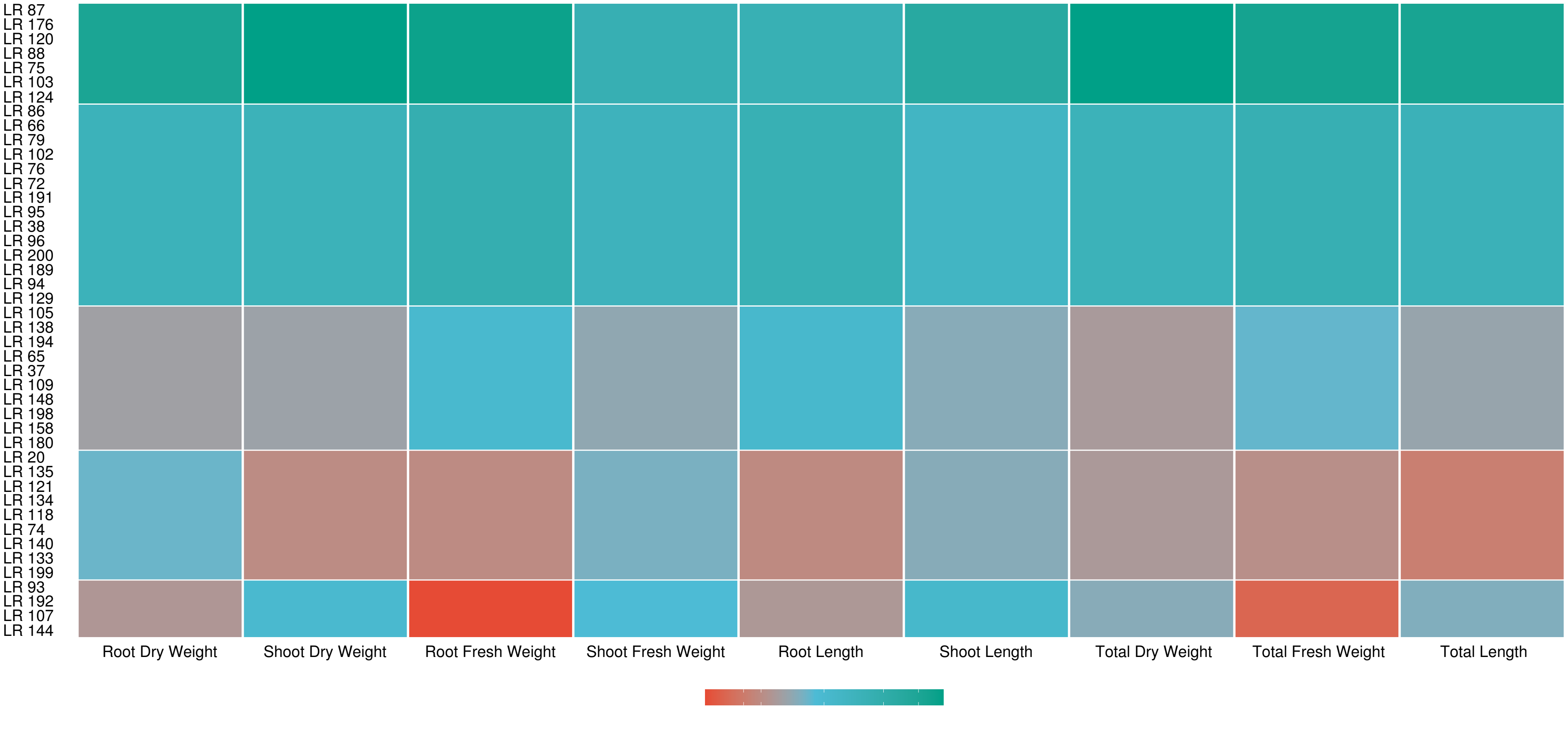
777 131. Khalid S, Baccile JA, Spraker JE, Tannous J, Imran M, et al. NRPS-Derived Isoquinolines
778 and Lipopeptides Mediate Antagonism between Plant Pathogenic Fungi and Bacteria. *ACS Chem*
779 *Biol* 2018;13:171–179. doi: 10.1021/acscchembio.7b00731

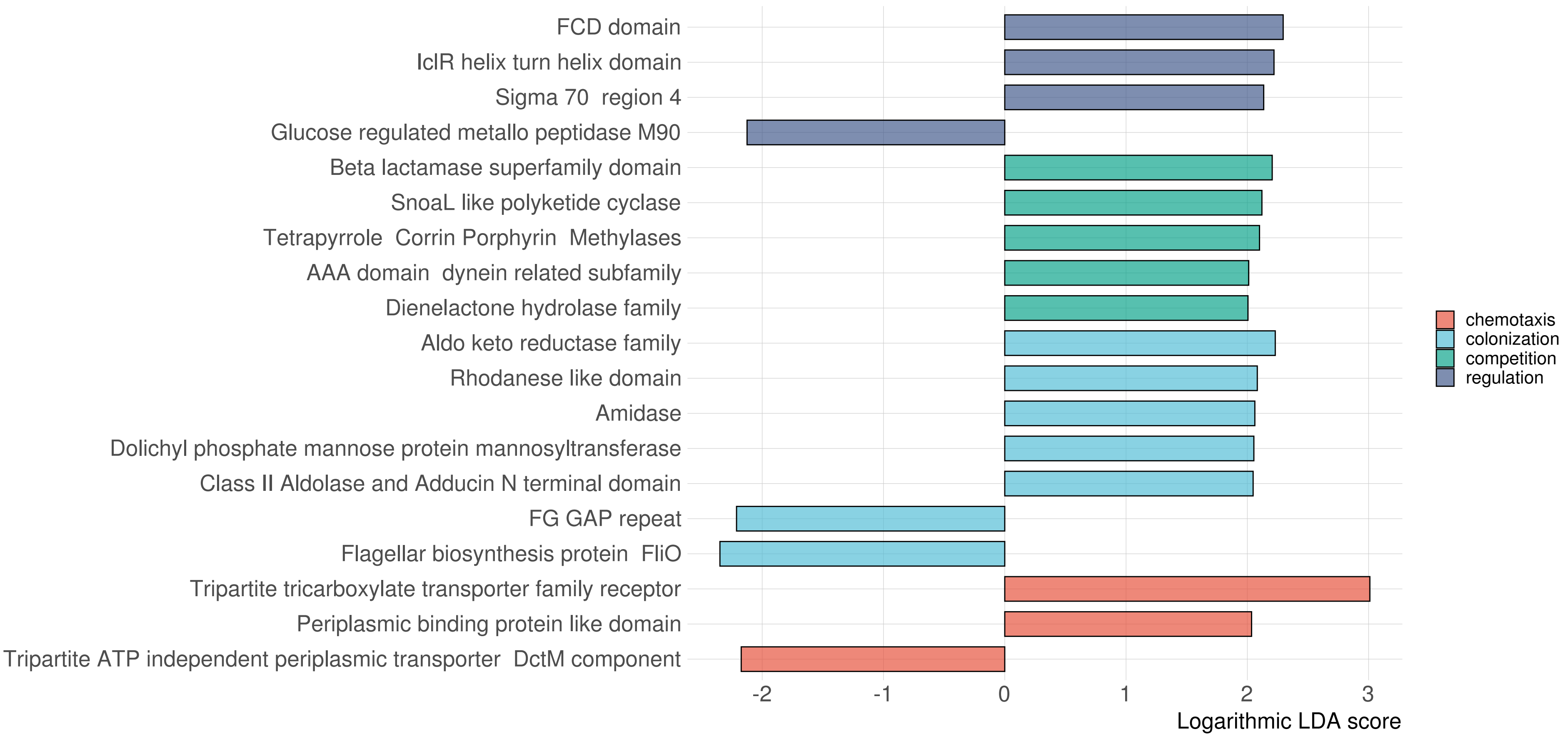
780 132. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. *Genome*
781 *Medicine* 2019;11:70. doi: 10.1186/s13073-019-0689-8

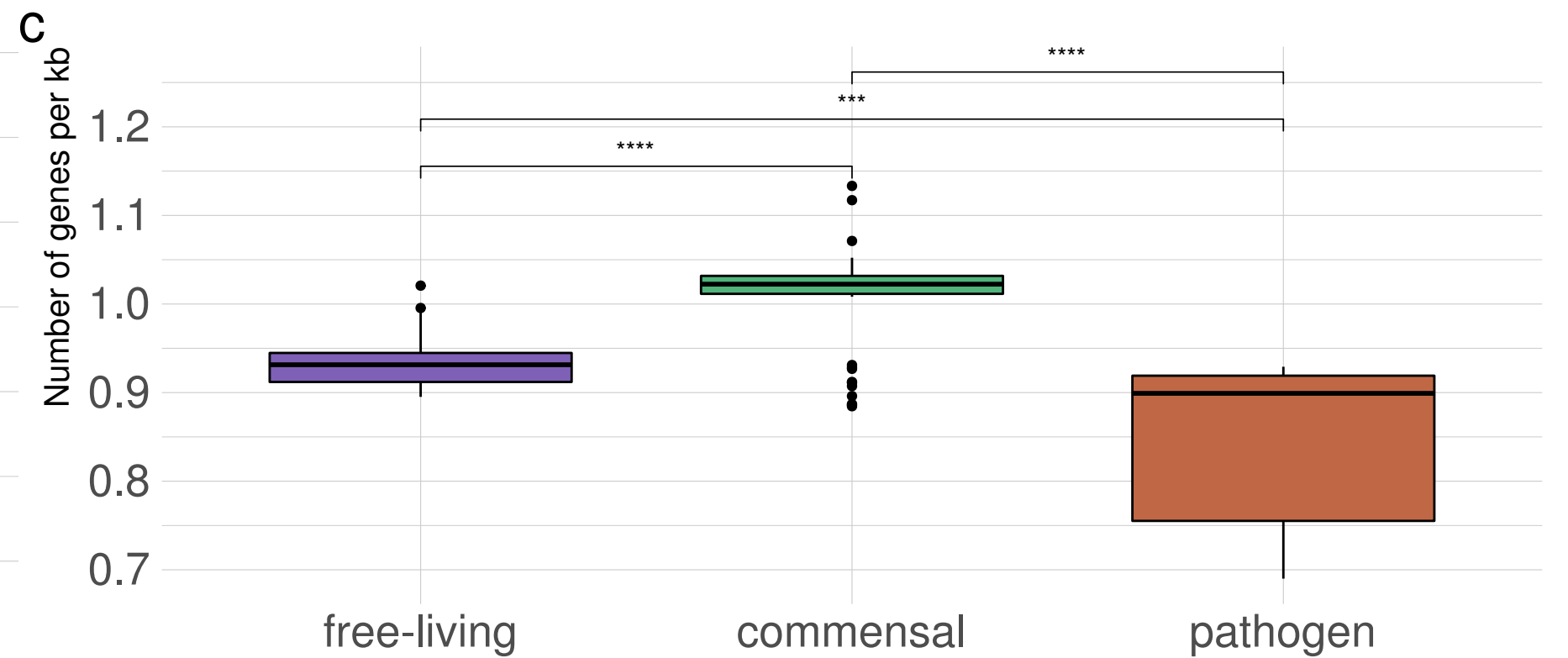
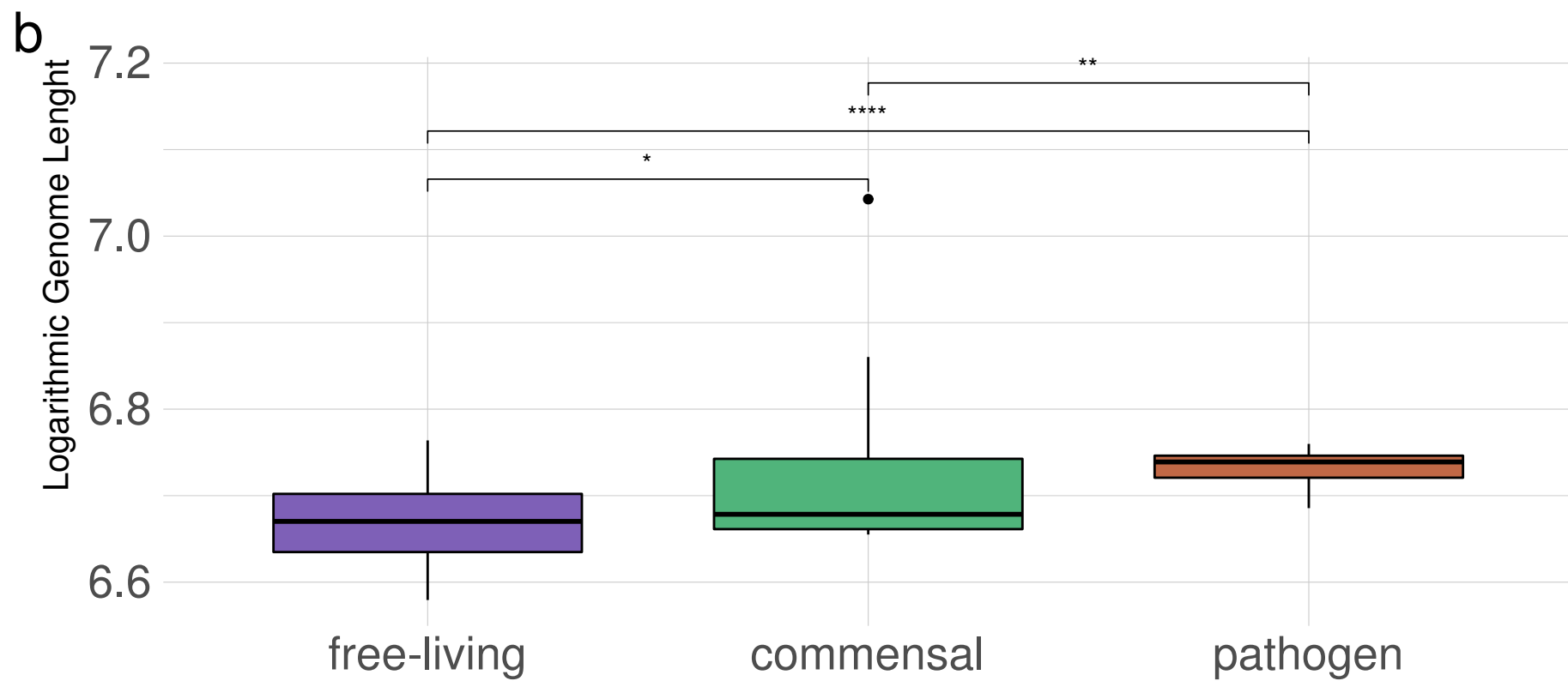
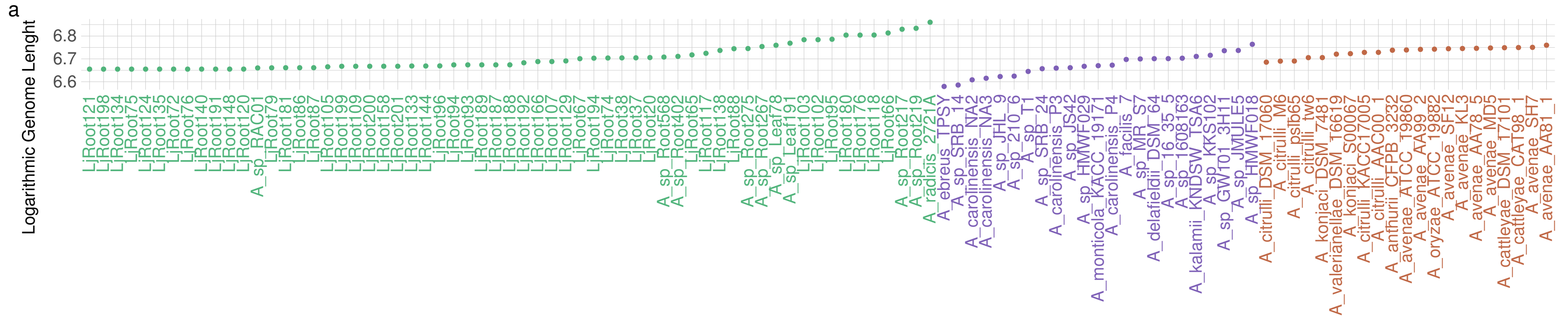
782 133. Iversen KH, Rasmussen LH, Al-Nakeeb K, Armenteros JJA, Jensen CS, et al. Similar
783 genomic patterns of clinical infective endocarditis and oral isolates of *Streptococcus sanguinis* and
784 *Streptococcus gordonii*. *Sci Rep* 2020;10:1–11. doi: 10.1038/s41598-020-59549-4

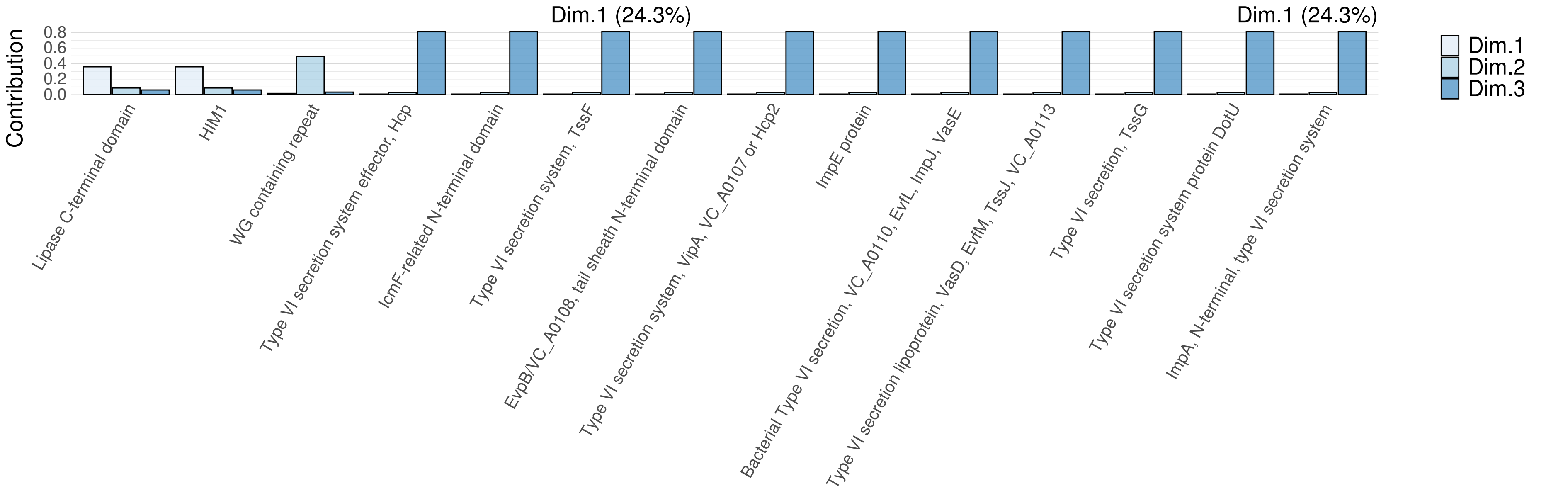
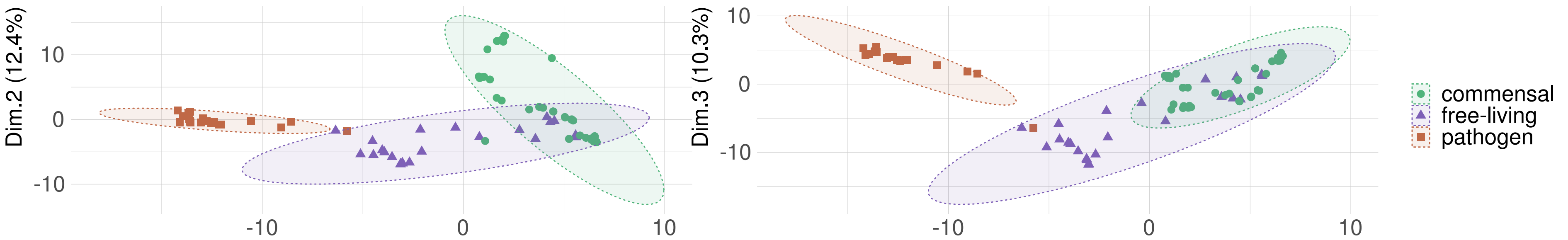
785 134. Qu K, Guo F, Liu X, Lin Y, Zou Q. Application of machine learning in microbiology. *Front*
786 *Microbiol* 2019;10:827. doi: 10.3389/fmicb.2019.00827

787 135. Tarca AL, Carey VJ, Chen X wen, Romero R, Drăghici S. Machine learning and its
788 applications to biology. *PLoS Comput Biol* 2007;3:e116. doi: 10.1371/journal.pcbi.0030116









b

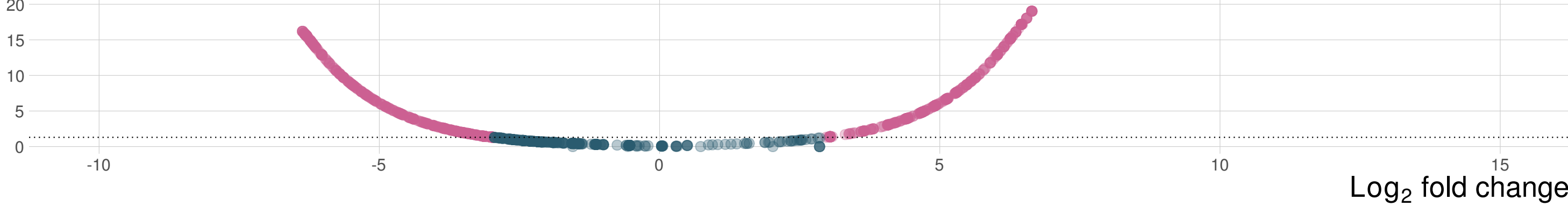
Enrichment Score



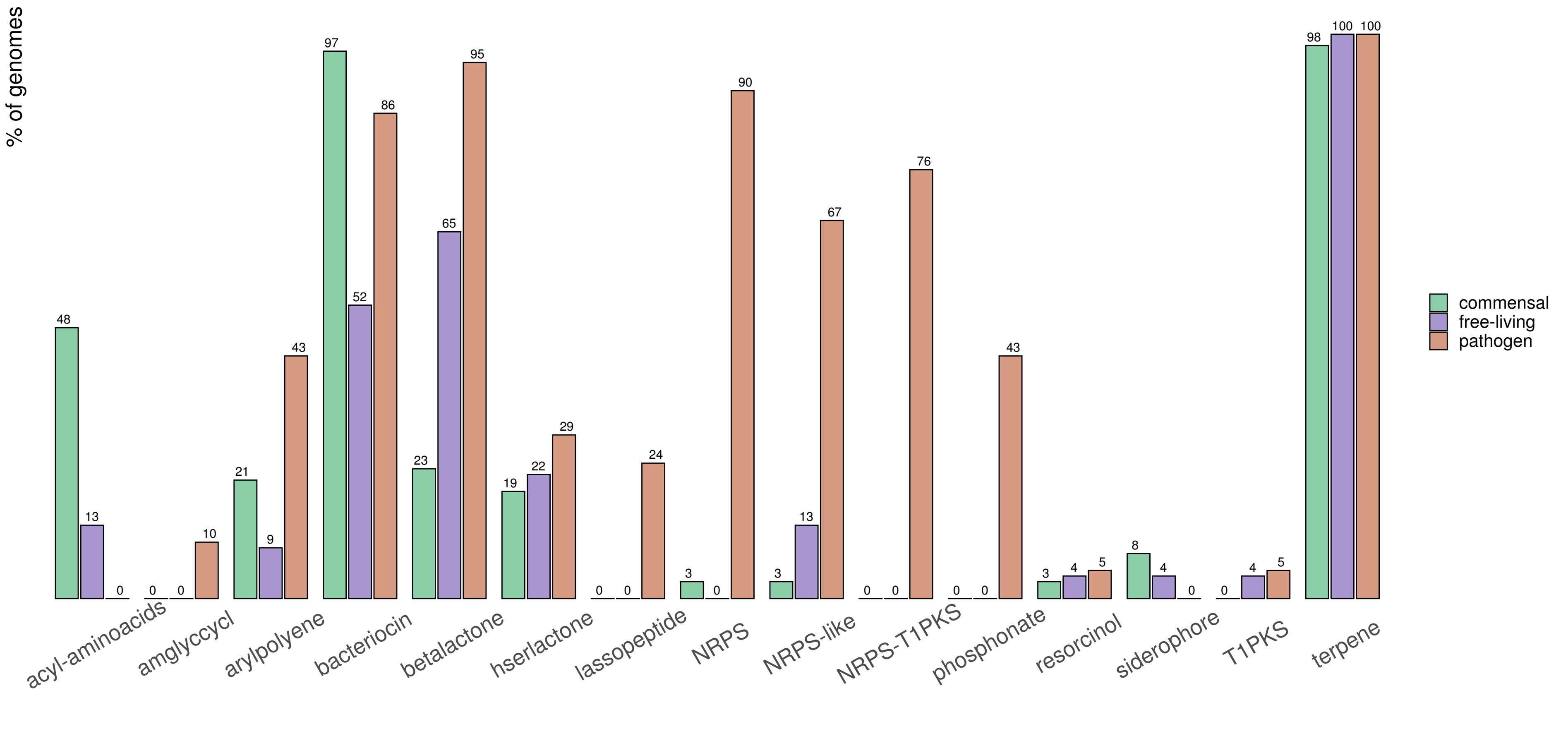
- effector
- hydrolytic enzyme
- motility
- secondary metabolism

a

$-\text{Log}_{10}$ q-value



% of genomes



FIGURES:

Fig.1: Heat-map showing the effects of single-strains on *Lotus japonicus* growth parameters as observed in the *L. japonicus* growth bioassays. Measures are shown as the median, for each observed group, of the percentage change from the control.

Fig.2: Bar-plot of the logarithmic scores from the Linear Discriminant analysis for the discriminant Pfams. Pfams are assigned to 4, color-coded, functional groups.

Fig.3: Anvi'o 6.2 circular representation of *Acidovorax* pan-genome and heat-map of average nucleotide identity across the genomes.

Fig.4: a) Dot-chart of individuals logarithmic genome length. b) box-plot of groups logarithmic genome length. c) box-plot of groups gene density expressed as number of genes per 1000 base-pairs. Significance was assessed by Student's t-test with Holm's correction ($p < 0.05$).

Fig.5: a) Ordination plot for the first and second component of the principal component analysis (PCA) b) Ordination for the first and third component of the PCA. c) Bar-plot of the Pfams with the highest contribution for first, second and third component of the PCA.

Fig.6: a) Volcano plot displaying Pfams enriched in commensal (left) and pathogenic strains (right). On the x axis is represented logarithmic 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 the curated list of enriched Pfams related to host-microbe or microbe-microbe interaction. Each dot is color coded based on the functional family of assignment.

Fig.7: Bar-plot of the percentage of genomes per group featuring each of the retrieved biosynthesis gene clusters.