nature genetics

SUPPLEMENTARY INFORMATION

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

In the format provided by the authors and unedited.

Genomic features of bacterial adaptation to plants

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

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

Supplementary Information

Table of Contents

Detailed bacterial isolation and genome sequencing process

Bacterial strains from Brassicaceae and Poplar were isolated using previously described protocols[1,2]. Poplar strains were cultured from root tissues collected from *Populus deltoides* and *Populus trichocarpa* trees. Strains designated CF and YR were isolated from native *Populus deltoides* growing in central Tennessee along the Caney Fork river and eastern North Carolina along the Yadkin river, respectively. The strains designated OV and OK were isolated from common garden grown *Populus trichocarpa* trees in Corvallis and Clatskanie, Oregon, respectively. Root samples were processed as described previously[3–6]. Briefly, rhizosphere strains were isolated by plating serial dilutions of root wash, while for endosphere strains, surface sterilized roots were pulverized with a sterile mortar and pestle in 10 mL of $MgSO_4$ (10 mM) solution followed by plating serial dilutions. For surface sterilization, roots were washed 5 times with sterile water, followed by 30s incubation in 95% ethanol, 3 min incubation in 5% NaOCl, then 6 washes with sterile water[3]. Strains were isolated on R2A agar media, and resulting colonies were picked and re-streaked a minimum of three times to ensure isolation. Isolated strains were identified by 16S rDNA PCR using primers 8F (AGAGTTTGATCCTGGCTCAG) and 1492R (GGTTACCTTGTTACGACTT) followed by Sanger sequencing and analysis.

For maize isolates, we selected soils associated with two different maize genotypes grown in two regions. Il14h, a sweet corn inbred line, and Mo17, a non-stiff stalk maize inbred line, were grown in two different fields (Lansing, NY and Urbana, IL). The rhizosphere soil samples from three replicates of each maize genotype grown at each field at week 12 after planting were collected as previously described[7]. A total of 12 rhizosphere soil samples were used to culture *Pseudomonas* isolates. From each rhizosphere soil sample, 0.1g soil was washed in 5 mL sterile phosphate buffered saline with 10% glycerol for 1 hour with gentle rocking at room temperature. 100 µL of the wash liquid was plated onto *Pseudomonas* Isolation Agar (BD Diagnostic Systems, Franklin Lakes, NJ) using a disposable inoculating loop. The plates were incubated at 30°C until colonies formed. To extract genomic DNA, single colonies were inoculated into 5 mL LB, and grown at 30°C overnight. The cultures were harvested by centrifugation at 5000 × g for 5 min, and the cells were lysed using the B1 and B2

solutions as described in the Qiagen Genomic DNA Handbook (Qiagen, Valencia, CA). The genomic DNA was precipitated with ethanol and sodium acetate, and pelleted after a centrifugation at 1811 × g for 30 min. PCR and Sanger sequencing of 16S rDNA were used to confirm the identity and purity of the genomic DNA preparations. The genomes of the first four *Pseudomonas* isolates cultured from each rhizosphere soil sample were sequenced. Thus, a total of 48 *Pseudomonas* isolate genomes were sequenced.

For isolation of single cells, *A. thaliana* accessions Col-0 and Cvi-0 were grown to maturity in 2.5 cm KORD pots (Canada) in Mason Farm or Clayton soil/sand mix (2 parts soil, 1 part sand)[8]. The pots containing the roots were turned upside down and the root mass was removed, carefully rolling and kneading the root mass to allow most of the soil to fall away. Dirty root masses were submerged and stirred in a separate 4L beaker of distilled water, allowing most soil to dislodge and sink. Roots were transferred to clean water and the process repeated, until the stirred water no longer appeared murky. All remaining soil and biological debris were carefully picked away from each root using sterile tweezers. For surface sterilization, the pool of visually clean roots was transferred to separate 250 mL glass bottles, each containing 200mL of a 1:10 dilution of household bleach in water containing 0.1% Triton X-100. Surface sterilization proceeded for 10 min with gentle agitation (inversion). The bleach solution was decanted and immediately replaced twice with autoclaved distilled water. The plant material was then treated for 2 minutes with 200mL of 2.5% sodium thiosulfate to fully neutralize the bleach; this was then washed twice more with autoclaved distilled water. Surfaced sterilized roots and leaves were then ground using a sterile mortar and pestle (grinding surfaces were sprayed with 95% ethanol and flamed several times) in a laminar flow hood. MES buffer (2.5 mM, ph 6.0) was added as needed to maintain a liquid consistency while grinding. Two parts of plant lysate were mixed with one part of autoclaved 80% glycerol for a 27% final glycerol concentration. This mixture was thoroughly mixed and pipetted in 1.5 mL aliquots into 2mL capacity cryovials and snap-frozen until further processing, resulting in 10-20 vials per condition.

To prepare cells for cell sorting, each glycerol stock was thawed on ice and diluted with 10mL sterile MES buffer (pH 6) to reduce viscosity. The solution was filtered

through a 100 micron cell strainer (Fisherbrand) into a 50mL tube, and subsequently aspirated with a syringe and passed through an 11 micron syringe filter (11µm Millipore nylon mesh in a Millipore Swinnex Filter Holder) to remove remaining plant particulates. The flow through was centrifuged for 5 min at 10,000 x g to re-concentrate the cells, and resuspended in 1mL of MES buffer plus 500uL 80% glycerol. This tube was vortexed, flash frozen, and shipped to the JGI for further processing. Individual cells were isolated using fluorescence-activated cell sorting (FACS) followed by DNA amplification using multiple displacement amplification (MDA), and 16S rDNA screening as described previously[9,10].

DNA from isolates and single cells was sequenced using next generation sequencing platforms, mostly using the Illumina HiSeq technology (Table S3). Libraries for Illumina sequencing were prepared using the following protocol: Plate-based DNA library preparation for Illumina sequencing was performed on the PerkinElmer Sciclone NGS robotic liquid handling system using a Kapa Biosystems library preparation kit. 200 ng of sample DNA was sheared to 300 bp using a Covaris LE220 focused-ultrasonicator. The sheared DNA fragments were size selected with solid phase reversible immobilization (SPRI) beads two times and then the selected fragments were end-repaired, A-tailed, and ligated with Illumina compatible sequencing adaptors from IDT containing a unique molecular index barcode for each sample library. The prepared library was then quantified using KAPA Biosystem's next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. The quantified library was then then multiplexed with other libraries, and the pool of libraries was then prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v3 or v4, and Illumina's cBot instrument to generate a clustered flowcell for sequencing. Sequencing of the flowcell was performed on the Illumina HiSeq2000/2500/1TB sequencer using a TruSeq SBS sequencing kit, v3 or v4, following a 2x150 indexed run recipe.

For the three genomes that were sequenced using MiSeq (Table S3), the libraries were prepared using the Nextera XT kit.

Some genomes were sequenced using 454 technology. Libraries were prepared using the following protocol: double-stranded genomic DNA samples were fragmented

via sonication to 400-800 base pairs. These fragments were end polished and ligated to a set of Y-shape adaptors. The 454 library fragments were then clonally amplified in bulk by capturing them through hybridization on microparticle beads and subjecting them to emulsion based PCR resulting in beads that were covered with millions of copies of a single DNA fragment (size range 400-800bp) where each bead contained a different clonally amplified library fragment. After amplification, the beads were recovered from the emulsions and loaded into the wells of a PicoTiterPlate device (PTP) such that wells contained single DNA beads. The PTP was then inserted into the 454 Genome Sequencer FLX-Titanium instrument for sequencing where sequencing reagents were sequentially flowed over the plate and the sequence of the DNA fragments was determined.

Sequenced genomic DNA was assembled using different assembly methods (Table S3). Genomes were annotated using the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4)[11] and were deposited at the Integrated Microbial Genomes (IMG) database[12], ENA[13] or Genbank[14] for public usage.

Analysis of the nine taxa prevalence in 16S and metagenome surveys

We used 16S rDNA surveys and metagenomes of the plant environment of *Arabidopsis*[8,15], barley[16], wheat, and cucumber[17]. The published information of the relative abundance and taxonomic assignments of operational taxonomic units (OTUs) was retrieved. Based on the taxonomic assignment the relative abundances of OTUs within a specific taxon were summed to yield the relative abundance of that taxon. If there were multiple replicates of an experiment we used the median value. Reads mapped outside of the bacteria kingdom were ignored in relative abundance calculations.

Assessment clustering quality using taxon-specific markers

In order to estimate the quality of the clusters output by UCLUST and Orthofinder, we ran the Phyla_Amphora[18] script MarkerScanner.pl using the default parameters over the 3837 genomes in our dataset. This resulted in the identification of sequences that are

homologues to the curated set of taxon markers contained in Phyla_Amphora. We used a custom script to summarize the marker scanning results in a table depicting the distribution of detected markers across the CDS ids and Genome ids in our dataset. Next, we compared the distribution of each of the taxon-specific markers identified by Phyla_Amphora across the clusters output by UCLUST and Orthofinder.

For each taxon-specific marker, we determined how many proteins in a given taxon (e.g Pseudomonas) were identified as a homologue to that marker. Then, we quantified the distribution of these homologues across the clusters output by UCLUST and Orthofinder. Ideally, all the homologues of a taxon-specific marker should be clustered in a single cluster of CDS (an orthogroup), in addition, that single cluster should contain only the CDS identified as homologues to the taxon marker. Using this logic, we estimated two metrics: the purity and fragmentation index. The purity index quantifies how many CDS are contained across all the clusters (UCLUST, Orthofinder) needed to cover the total number of CDS identified as homologues to a specific Phyla_Amphora taxon marker. The fragmentation index quantifies the number of clusters (UCLUST, Orthofinder) needed to cover the total number of CDS identified as homologues to a specific Phyla_Amphora taxon marker. The two metrics described above were calculated over each of the 9 taxa individually. The data and scripts utilized to compute these measurements across all taxa are available on: https://github.com/isaisg/gfobap/tree/master/phyla_amphora_benchmark
The plots generated from this analysis were generated using the ggplot[19] (v 2.2.1) package from R.

Construction of pan genome matrices, relational tables and HMM databases from the Orthofinder orthogroups

For each of the nine taxa, we used custom scripts to transform the orthogroup result output from Orthofinder into pan genome matrices depicting the distribution of orthogroups across the genomes of that given taxon. Additionally, we constructed tables exhibiting the distribution of orthogroups across genomes based on the CDS IDs. The pan genome matrices and tables described above can be downloaded from:
http://labs.bio.unc.edu/Dangl/Resources/gfobap_website/matrices_df_ogs.html

The scripts used to compute the matrices and tables can be found in:
https://github.com/isaisg/gfobap/tree/master/orthofinder_orthogroups_to_matrices_dataframes

Additionally, for each orthogroup in our dataset consisting of more than two CDS, we built multiple sequence alignments using MAFFT[20] (v7.305b). Subsequently, we used these alignments as inputs for Hidden Markov Model (HMMs) construction using the hmmbuild command from the HMMER suite[21] (v 3.1b2). We built nine HMM databases, one corresponding to each of the nine taxa analyzed in this study. Also, to complement these databases, we developed a scanning pipeline that utilizes the HMMER suite to search for the HMM orthogroups in our nine databases over any genome provided to the pipeline. The MAFFT alignments, HMM profiles and HMM databases can be downloaded from:
http://labs.bio.unc.edu/Dangl/Resources/gfobap_website/mafft_hmm.html

The scripts to compute the alignments and the HMM profiles plus the pipeline to scan novel genomes using the HMM databases can be downloaded from:
https://github.com/isaisg/gfobap/tree/master/mafft_hmm
https://github.com/isaisg/gfobap/tree/master/scanner_orthogroups_scripts

Assesment of PA/NPA prediction robustness using validation genome datasets

Seven validation genome groups were assembled representing the following genera: *Bacillus* (order Bacillales, n=222), *Burkholderia* (order Burkholderiales, n=121), *Chryseobacterium* (phylum Bacteroidetes, n=35), *Flavobacterium* (phylum Bacteroidetes, n=44), *Paenibacillus* (order Bacillales, n=60), *Sphingomonas* (Class Alphaproteobacteria, n=59), *Streptomyces* (Group Actinobacteria1, n=90). These datasets contained at least 10 genomes in PA and NPA groups and were relatively balanced. In addition, we compiled a new set of genomes from *Bacillus* (n=66) and *Pseudomonas* (n=24) that were labeled as PA or NPA based on their isolation sites. Each of the statistical approaches was run on the nine validation datasets to yield

significant PA and NPA Pfam domains. The significant domains were compared against the significant Pfam domains predicted from the original and much larger genome datasets to find the overlap between these two sets. The significant PA/NPA Pfam domains predicted based on the validation sets had on average 65-73%, 40%, and 20-24% overlap with the original results based on the entire dataset, for Hyperg, Scoary, and PhyloGLM, respectively. We deduce that the power and reproducibilty of Phyloglm and Scoary is limited when taxa are compared at low taxonomic ranks genus/family).

Growth and transformation of *Paraburkholderia Kururiensis* M130 affecting rice root colonization

Paraburkholderia kururiensis strain M130 was grown at 30^{o}C in King's B medium[22]. *E coli* strains were grown at 37^{o} C in Luria Bertani medium. When needed, antibiotics were used in the following concentrations: ampicillin, 100 µg/mL; kanamycin, 50 µg/mL; nitrofurantoin, 50 µg/mL; rifampicin 50 µg/mL.

Recombinant DNA techniques

Recombinant DNA techniques, including digestion using restriction enzymes (New England Biolabs UK), agarose gel electrophoresis, purification of DNA fragments, ligation with T4 ligase, and transformation of *E. coli* were performed as described [23]. Plasmids were purified using EuroClone columns (EuroClone S.p.A., Italy). Triparental matings to mobilize DNA from *E. coli* to *P. kururiensis* were carried out with the helper strain *E. coli* (pRK2013)[24]. PCR amplifications were performed using GoTaq Flexi DNA Polymerase (Promega, Madison, WI, USA).

Genes reproducibly enriched or depleted in phylogenetically diverse PA and RA genomes

Pfam domains and COG proteins ('terms') that were found as significantly PA and RA, or soil and NPA in multiple taxa according to the hypergeometric test were retrieved. The proportions of genes carrying the term (or multiple terms in cases of a term combination in a gene) in each genome were used in a *t*-test.

We searched the direct neighbors of *LacI*-family genes across all analyzed PA genomes. For each gene found, we retrieved its COG annotation which was translated to a COG category. Only informative COG categories were used. COGs belonging to "Function unknown" and "General function prediction only" were filtered out due to limited functional information.

In order to find LacI-family TF DNA binding sites, we scanned the intergenic regions of over 25 bp length between LacI-family genes and directly adjacent (in upstream, downstream or antisense head-to-head and tail-to-tail orientations) carbohydrate-related genes for abundant kmers of different lengths using wordcount (Emboss package[25]). The most abundant motifs found in multiple taxa were compared against their distribution in random intergenic sequences using the Fisher exact test.

Annotating proteins with PREPARADOs as being candidates for secretion

We annotated PREPARADO-containing proteins as secreted by Sec if they had a predicted signal peptide[26] and lacked a transmembrane domain according to IMG annotation. A protein was marked as being secreted by T3SS if it had a score > 0.999 according to EffectiveT3 as implemented by the effectivedb server[27]. A domain was predicted to be associated with secretion by Sec or T3SS if over 50% of the proteins carrying the domain were predicted to be secreted by these secretion systems. The proportions of proteins carrying the different domains and being secreted are mentioned in Table S21.

Detailed construction of Δ5-Hyde1 strain

Acidovorax Citrulli (*A. citrulli*) strain AAC00-1 and its derived mutants were grown on nutrient agar (NA) medium (Thermo Fisher Scientific Inc, Waltham, MA) supplemented with rifampicin (100 µg/ml). To delete a cluster of five *Hyde1* genes (Aave_3191, Aave_3192, Aave_3193, Aave_3194, Aave_3195), we performed a marker-exchange mutagenesis as previously described[28]. Briefly, DNA fragments from regions flanking the *Hyde1* gene cluster were amplified using the following primers: Aave3187SwaFor and Aave3187SwaRev (upstream region, 1.2Kb), Aave3196PmeFor and Aave3196PmeRev (downstream region, 1.485Kb). A kanamycin (Km) resistance gene

(nptII) was amplified from pDK4[29] with primers km_for and km_rev. The derived NptII gene is flanking with the FLP recognition target sites. The two flanking fragments of the *Hyde1* cluster were then fused to the nptII gene by overlap PCR[30]. The derived cassettes were cloned into the PCR8/GW-Topo vector (Invitrogen), and cloned into the suicide vector pLVC18L-Des[31] using LR clonase (Invitrogen, Carlsbad, CA). The derived construct was then mobilized into *A. citrulli* strain AAC00-1 by tri-parental mating as previously described[32]. Double crossover mutants were selected using marker-exchange mutagenesis as previously reported[31]. *A. citrulli* strain that contained impaired genes were selected on NA medium supplemented with rifampicin (100 µg/ml) and kanamycin (50 µg/ml). The kanamycin resistant mutant strain was further transformed with a modified plasmid vector pBBR1FLP2 that carries the FLP recombinase gene[33]. The mutant strain lost the kanamycin resistance gene and the modified pBBR1FLP2 plasmid was further selected as previously described[33]. The marker-free mutant was designated as ∆1-Hyde1, and its genotype was confirmed by PCR amplification with primers "Aave3187 check for" and "Aave3196 check rev". The PCR product was confirmed by sequencing.

The marker-exchange mutagenesis procedure was repeated to further delete four *Hyde1* loci: Aave_0989, Aave_4706, Aave_4335, and Aave_1108. Primers used to amplify the up- and downstream flanking sequences and check the deletions are listed in Table S25**.** The final mutant with deletion of 9 out of 11 *Hyde1* genes was designated as ∆5-Hyde1, which has been used for competition assay.

A similar procedure was also used to generate the ∆T6SS mutant. The primers used for amplify the upstream and downstream DNA sequences around the *vasD* gene homologue (Aave_1470) are Aave1469 Swa For and Aave1469 Swa Rev (upstream); Aave1471 PmeFor and Aave1471 PmeFor (downstream). The ∆T6SS mutant was checked with primers "Aave1470 check for" and "Aave1470 check rev".

Supplementary Methods references

1. Lebeis, S. L. *et al.* Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa. *Science (80-.).* **349,** (2015).

2. Doty, S. L. *et al.* Diazotrophic endophytes of native black cottonwood and willow. *Symbiosis* **47,** 23–33 (2009).

3. Gottel, N. R. *et al.* Distinct microbial communities within the endosphere and rhizosphere of *Populus deltoides* roots across contrasting soil types. *Appl. Environ. Microbiol.* **77,** 5934–5944 (2011).

4. Weston, D. J. *et al. Pseudomonas fluorescens* induces strain-dependent and strain-independent host plant responses in defense networks, primary metabolism, photosynthesis, and fitness. *Mol. Plant-Microbe Interact.* **25,** 765–778 (2012).

5. Shakya, M. *et al.* A multifactor analysis of fungal and bacterial community structure in the root microbiome of mature *Populus deltoides* trees. *PLoS One* **8,** e76382 (2013).

6. Klingeman, D. M. *et al.* Draft genome sequences of four Streptomyces isolates from the *Populus trichocarpa* root endosphere and rhizosphere. *Genome Announc.* **3,** (2015).

7. Peiffer, J. A. *et al.* Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 6548–53 (2013).

8. Lundberg, D. S. *et al.* Defining the core *Arabidopsis thaliana* root microbiome. *Nature* **488,** 86–90 (2012).

9. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499,** 431–437 (2013).

10. Rinke, C. *et al.* Obtaining genomes from uncultivated environmental microorganisms using FACS–based single-cell genomics. *Nat. Protoc.* **9,** 1038–1048 (2014).

11. Huntemann, M. *et al.* The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v . 4). *Stand. Genomic Sci.* 1–6 (2015). doi:10.1186/s40793-015-0077-y

12. Integrated Microbial Genomes. Available at: https://img.jgi.doe.gov/cgi-bin/mer/main.cgi.

13. Toribio, A. L. *et al.* European Nucleotide Archive in 2016. *Nucleic Acids Res.* **45,** D32–D36 (2017).

14. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **45,** D37–D42 (2017).

15. Bulgarelli, D. *et al.* Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. *Nature* **488,** 91–5 (2012).

16. Bulgarelli, D. *et al.* Structure and function of the bacterial root microbiota in wild and domesticated barley. *Cell Host Microbe* **17,** 392–403 (2015).

17. Ofek-Lalzar, M. *et al.* Niche and host-associated functional signatures of the root surface microbiome. *Nat. Commun.* **5,** 4950 (2014).

18. Wang, Z. & Wu, M. A Phylum-Level Bacterial Phylogenetic Marker Database. *Mol. Biol. Evol.* **30,** 1258–1262 (2013).

19. Wickham, H. *Ggplot2 : elegant graphics for data analysis*. (Springer, 2009).

20. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30,** 3059–66 (2002).

21. Finn, R. D. *et al.* HMMER web server: 2015 update. *Nucleic Acids Res.* **43,** W30–

W38 (2015).

22. KING, E. O., WARD, M. K. & RANEY, D. E. Two simple media for the demonstration of pyocyanin and fluorescin. *J. Lab. Clin. Med.* **44,** 301–7 (1954).

23. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular cloning : a laboratory manual*. (Cold Spring Harbor Laboratory, 1989).

24. Figurski, D. H. & Helinski, D. R. Replication of an origin-containing derivative of plasmid RK2 dependent on a plasmid function provided in trans. *Proc. Natl. Acad. Sci. U. S. A.* **76,** 1648–52 (1979).

25. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–7 (2000).

26. Deng, P. *et al.* Comparative genome-wide analysis reveals that *Burkholderia contaminans* MS14 possesses multiple antimicrobial biosynthesis genes but not major genetic loci required for pathogenesis. *Microbiologyopen* **5,** 353–369 (2016).

27. Eichinger, V. *et al.* EffectiveDB--updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. *Nucleic Acids Res.* **44,** D669-74 (2016).

28. Traore, S. M. Characterization of Type Three Effector Genes of A. citrulli, the Causal Agent of Bacterial Fruit Blotch of Cucurbits. (Virginia Polytechnic Institute and State University, 2014).

29. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc. Natl. Acad. Sci.* **97,** 6640–6645 (2000).

30. Innis, M. A. *PCR protocols : a guide to methods and applications*. (Academic Press, 1990).

31. Zhao, B., Dahlbeck, D., Krasileva, K. V., Fong, R. W. & Staskawicz, B. J. Computational and Biochemical Analysis of the Xanthomonas Effector AvrBs2 and Its Role in the Modulation of Xanthomonas Type Three Effector Delivery. *PLoS Pathog.* **7,** e1002408 (2011).

32. Ditta, G., Stanfield, S., Corbin, D. & Helinski, D. R. Broad host range DNA cloning system for gram-negative bacteria: construction of a gene bank of Rhizobium meliloti. *Proc. Natl. Acad. Sci. U. S. A.* **77,** 7347–51 (1980).

33. Jittawuttipoka, T. *et al.* Mini-Tn 7 vectors as genetic tools for gene cloning at a single copy number in an industrially important and phytopathogenic bacteria, Xanthomonas spp. *FEMS Microbiol. Lett.* **298,** 111–117 (2009).

34. Zhang, Z. *et al.* Disruption of PAMP-Induced MAP Kinase Cascade by a Pseudomonas syringae Effector Activates Plant Immunity Mediated by the NB-LRR Protein SUMM2. *Cell Host Microbe* **11,** 253–263 (2012).

35. Bai, Y. *et al.* Functional overlap of the Arabidopsis leaf and root microbiota. *Nature* **528,** 364–369 (2015).

36. Ho, B. T., Dong, T. G. & Mekalanos, J. J. A view to a kill: the bacterial type VI secretion system. *Cell Host Microbe* **15,** 9–21 (2014).

37. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25,** 1043–55 (2015).

Supplementary Figure Legends

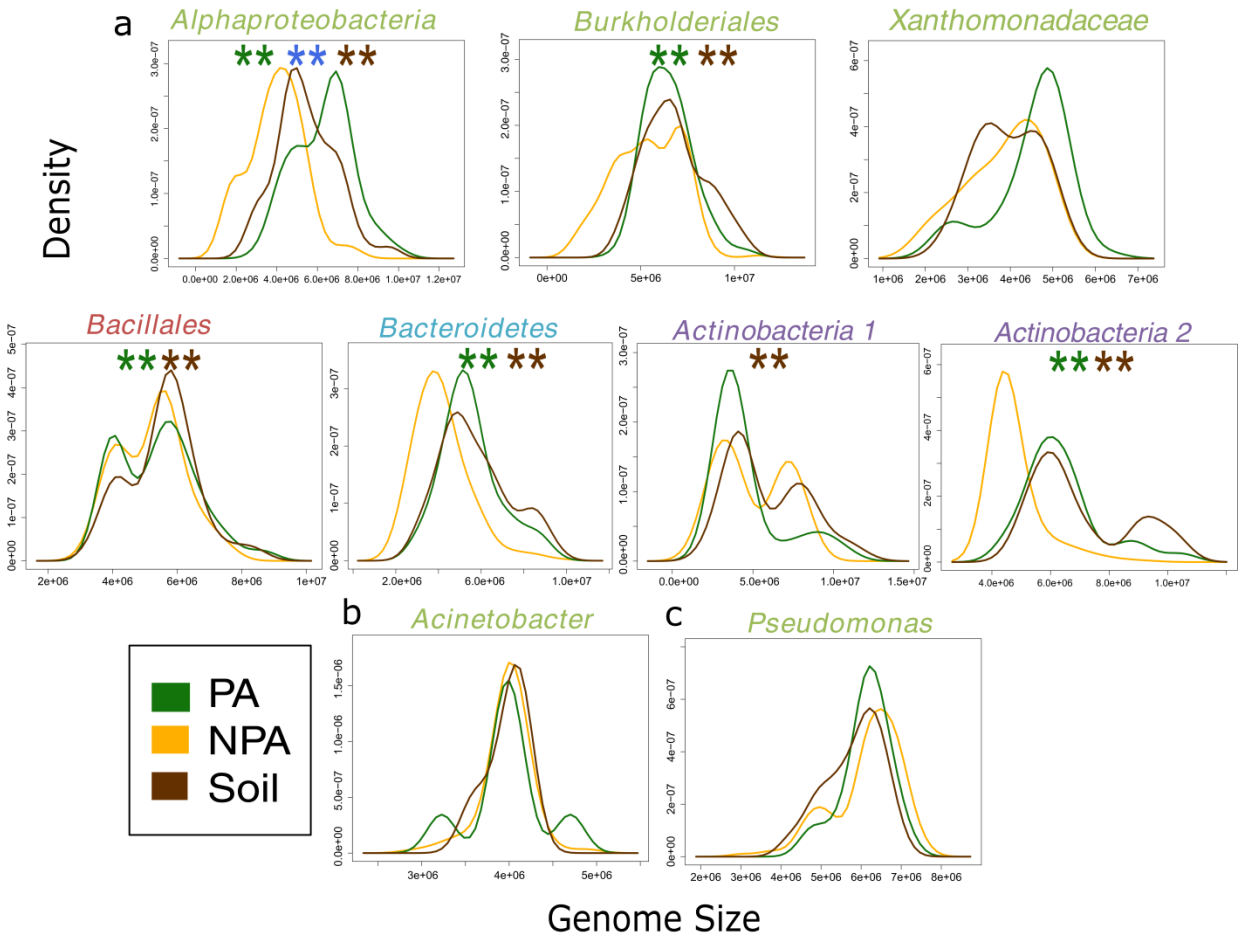
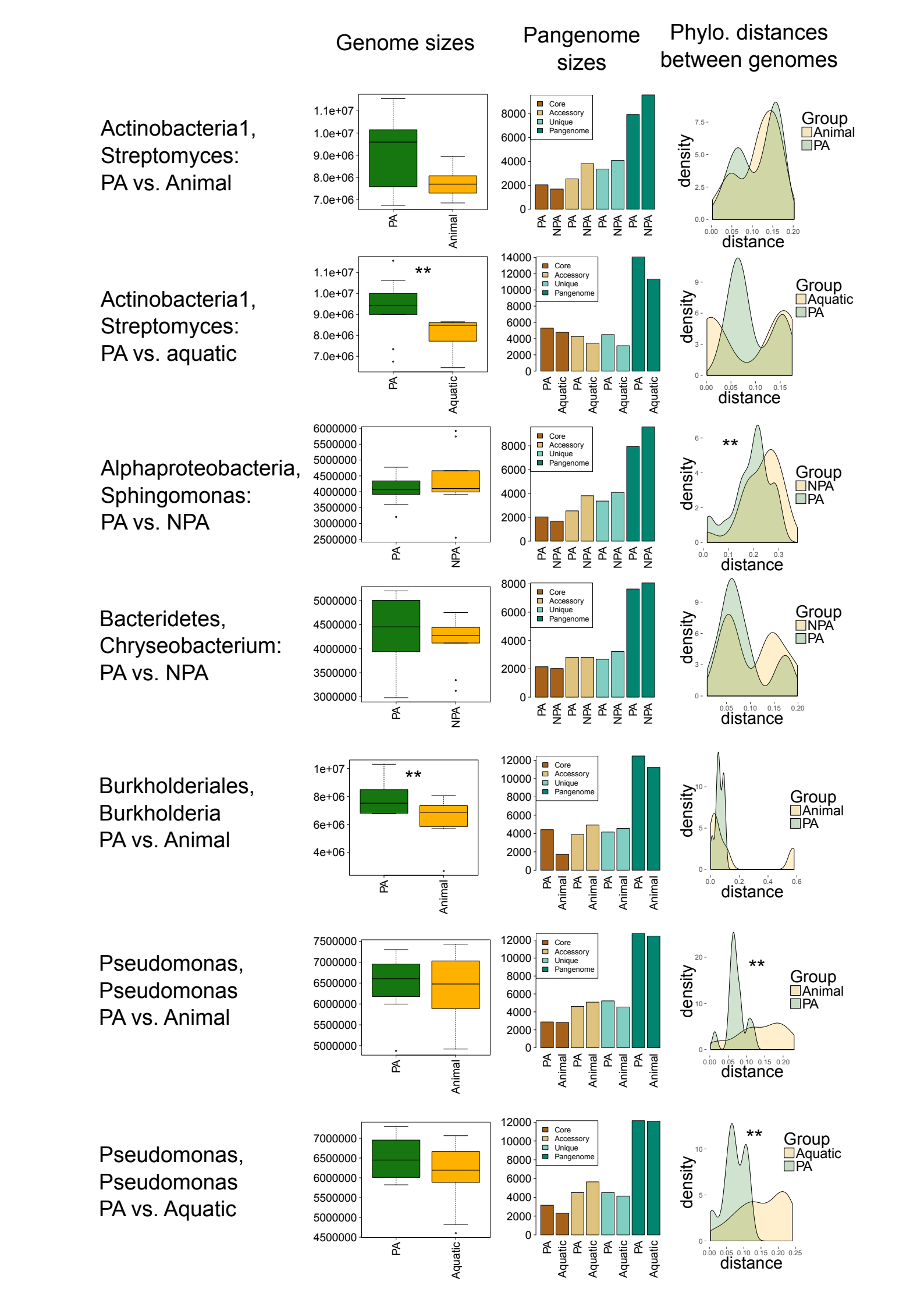


Figure S1. Density plots of genome size as a function of genome label. X and Y axes are genome size and the density in all panels, respectively. **a**. Taxa in which PA genomes and/or soil genomes are significantly larger than NPA genomes. **b**. A taxon in which there are no significant differences among any of the groups tested. **c**. A taxon in which PA and NPA genomes are both significantly larger than soil genomes and there is no significant difference in size between PA and NPA genomes. In all kernel density plots, the color of the main title represents a distinct phylum; green – *Proteobacteria*, blue – *Bacteroidetes*, red – *Firmicutes*, purple – *Actinobacteria*. Significant genome size differences are based on *t*-test ($P < 0.05$ is considered statistically significant). Double asterisks denote PhyloGLM results. Green – PA genomes are larger than NPA genomes, brown – soil genomes are larger than NPA genomes, blue – RA (root-associated) genomes are larger than soil genomes. Full results are presented in Supplementary Table 5.

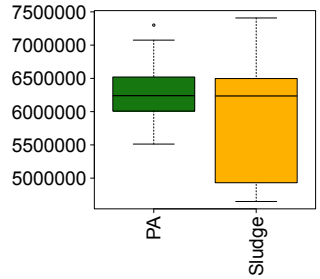

Genome sizes
Pangenome sizes
Phylo. distances between genomes
Actinobacteria1, Streptomyces: PA vs. Animal
Actinobacteria1, Streptomyces: PA vs. aquatic
Alphaproteobacteria, Sphingomonas: PA vs. NPA
Bacteridetes, Chryseobacterium: PA vs. NPA
Burkholderiales, Burkholderia PA vs. Animal
Pseudomonas, Pseudomonas PA vs. Animal
Pseudomonas, Pseudomonas PA vs. Aquatic
Core
Accessory
Unique
Pangenome
density
distance
Group
Animal
Aquatic
NPA
PA
**

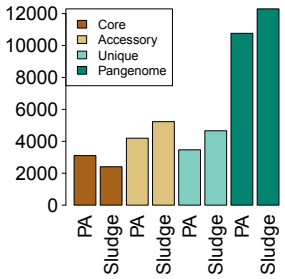
Genome sizes

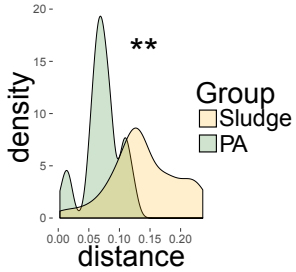
Pangenome sizes

Phylo. distances between genomes

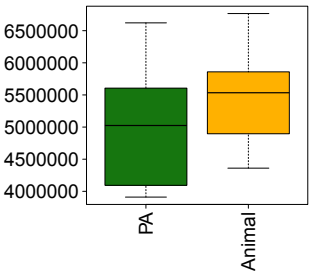
Pseudomonas, Pseudomonas PA vs. Sludge

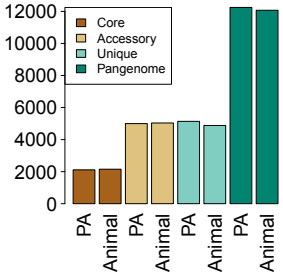


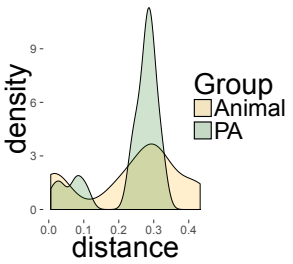




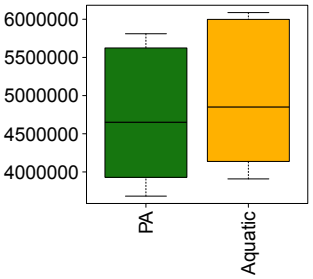
Bacillales, Bacillus: PA vs. Animal

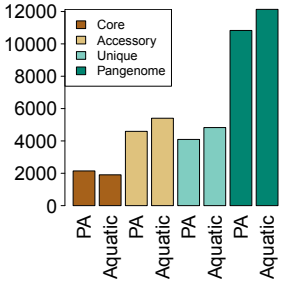


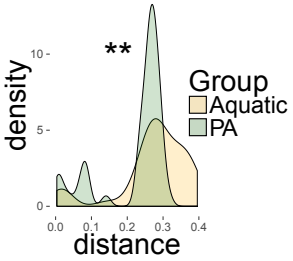




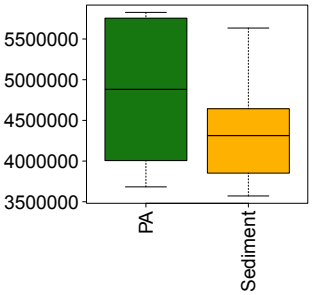
Bacillales, Bacillus: PA vs. Aquatic

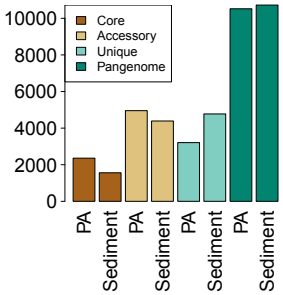


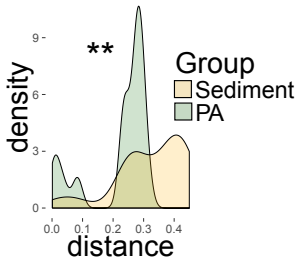




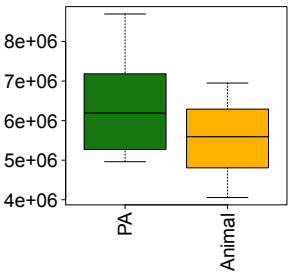
Bacillales, Bacillus: PA vs. Sediment

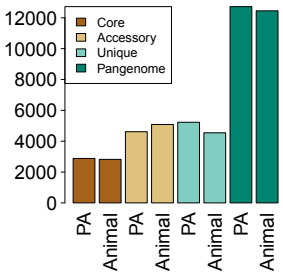


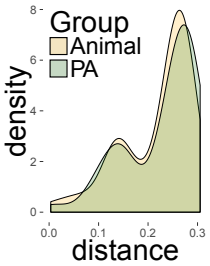




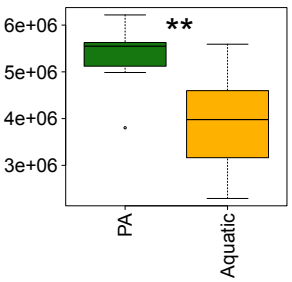
Bacillales, Paenibacillus: PA vs. Animal

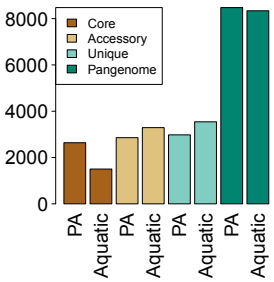






Bacteroidetes, Flavobacterium: PA vs. Aquatic





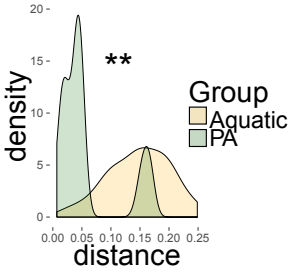


Figure S2. PA and NPA bacteria have similar pangenome size. PA genomes tend to be more closely related (lower phylogenetic distance between genomes) and therefore have large core genome and lower number of unique genes in comparison to NPA genomes. Twenty random genomes were selected from genera that have mixed PA and NPA classifications. For each NPA group, we used only organisms isolated from the same type of environment (e.g. Aquatic environment). The genome selection process aimed to minimize differences in phylogenetic distance distribution between the PA and NPA groups (Methods). Left panels: differences in genome size between PA and NPA groups. Y axis is genome size. Central panels: number of genes within each group of genes (core, accessory genes, unique genes, pangenome). Right panel: phylogenetic distances between all pairs in each group. ** significant difference (*t*-test $P < 0.05$).

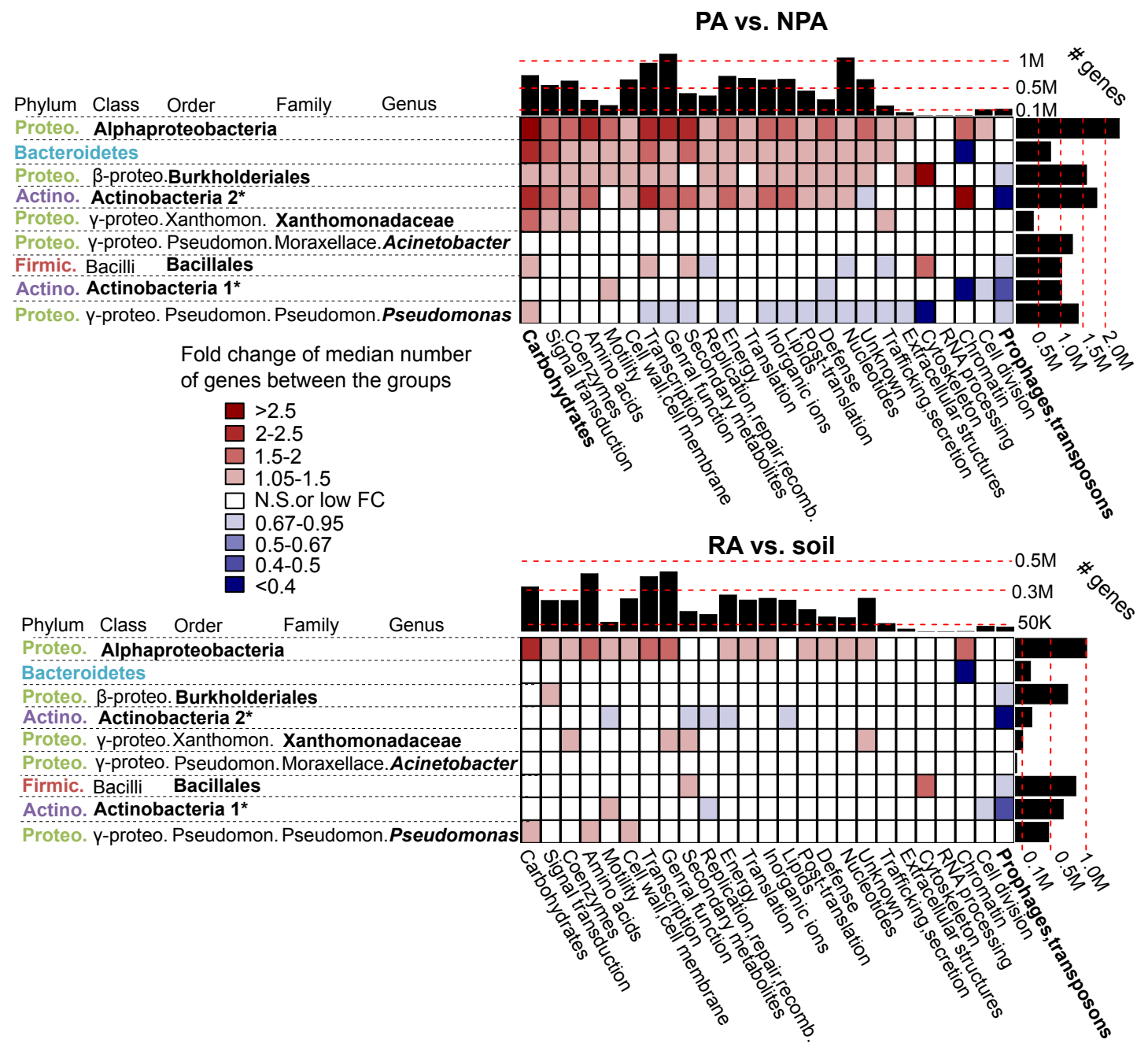

PA vs. NPA
Phylum Class Order Family Genus
Proteo. Alphaproteobacteria
Bacteroidetes
Proteo. β-proteo. Burkholderiales
Actino. Actinobacteria 2*
Proteo. γ-proteo. Xanthomon. Xanthomonadaceae
Proteo. γ-proteo. Pseudomon. Moraxellace. Acinetobacter
Firmic. Bacilli Bacillales
Actino. Actinobacteria 1*
Proteo. γ-proteo. Pseudomon. Pseudomon. Pseudomonas
1M
0.5M
0.1M
#genes
Carbohydrates
Signal transduction
Coenzymes
Amino acids
Motility
Cell wall,cell membrane
Transcription
Genral function
Secondary metabolites
Replication,repair,recomb.
Energy
Translation
Inorganic ions
Lipids
Post-translation
Defense
Nucleotides
Unknown
Trafficking,secretion
Extracellular structures
Cytoskeleton
RNA processing
Chromatin
Cell division
Prophages,transposons
0.5M
1.0M
1.5M
2.0M
Fold change of median number of genes between the groups
>2.5
2-2.5
1.5-2
1.05-1.5
N.S.or low FC
0.67-0.95
0.5-0.67
0.4-0.5
<0.4
RA vs. soil
Phylum Class Order Family Genus
Proteo. Alphaproteobacteria
Bacteroidetes
Proteo. β-proteo. Burkholderiales
Actino. Actinobacteria 2*
Proteo. γ-proteo. Xanthomon. Xanthomonadaceae
Proteo. γ-proteo. Pseudomon. Moraxellace. Acinetobacter
Firmic. Bacilli Bacillales
Actino. Actinobacteria 1*
Proteo. γ-proteo. Pseudomon. Pseudomon. Pseudomonas
0.5M
0.3M
50K
#genes
Carbohydrates
Signal transduction
Coenzymes
Amino acids
Motility
Cell wall,cell membrane
Transcription
Genral function
Secondary metabolites
Replication,repair,recomb.
Energy
Translation
Inorganic ions
Lipids
Post-translation
Defense
Nucleotides
Unknown
Trafficking,secretion
Extracellular structures
Cytoskeleton
RNA processing
Chromatin
Cell division
Prophages,transposons
0.1M
0.5M
1.0M

Figure S3.

Fold change differences in gene categories between PA/NPA and RA/soil genomes of the same taxon. Top panel: PA vs. NPA genomes. Bottom panel: RA vs. soil genomes. For both panels, the heat map indicates the level of enrichment or depletion. Hot colored cells indicate significantly more genes (q value < 0.05, FDR corrected two-sided *t*-test) in PA and RA genomes in the upper and lower panels, respectively. Histograms on the upper and right margins represent the total number of genes compared in each column and row, respectively. PA – plant-associated, NPA – non-plant associated, RA – root associated, soil –soil-associated. * not a formal class name. Carbohydrates – Carbohydrate metabolism and transport gene category. N.S – non-significant; q value >= 0.05 (FDR corrected two-sided *t*-test). FC – fold change. Full COG category names from the x axis appear in Table S6. Note that cells with high estimate absolute values (dark colors) are based on categories of few genes and are therefore more likely less accurate.

PA and NPA genomes of taxon X

Protein/domain pooling

Protein/domain clustering + calculation of phylogentic diversity

OrthoFinder (proteins)

COG (proteins)

Pfam (protein domains)

.... Other protein annotations TIGRFAM, KO

.....

Calling significant PA and NPA clusters

PhyloGLM copy number

Hyperg copy number

Other approaces (hypergbin, phyloglmbin, scoary)

PhyloGLM copy number

Hyperg copy number

Three other approaces

PhyloGLM copy number

Hyperg copy number

Three other approaces

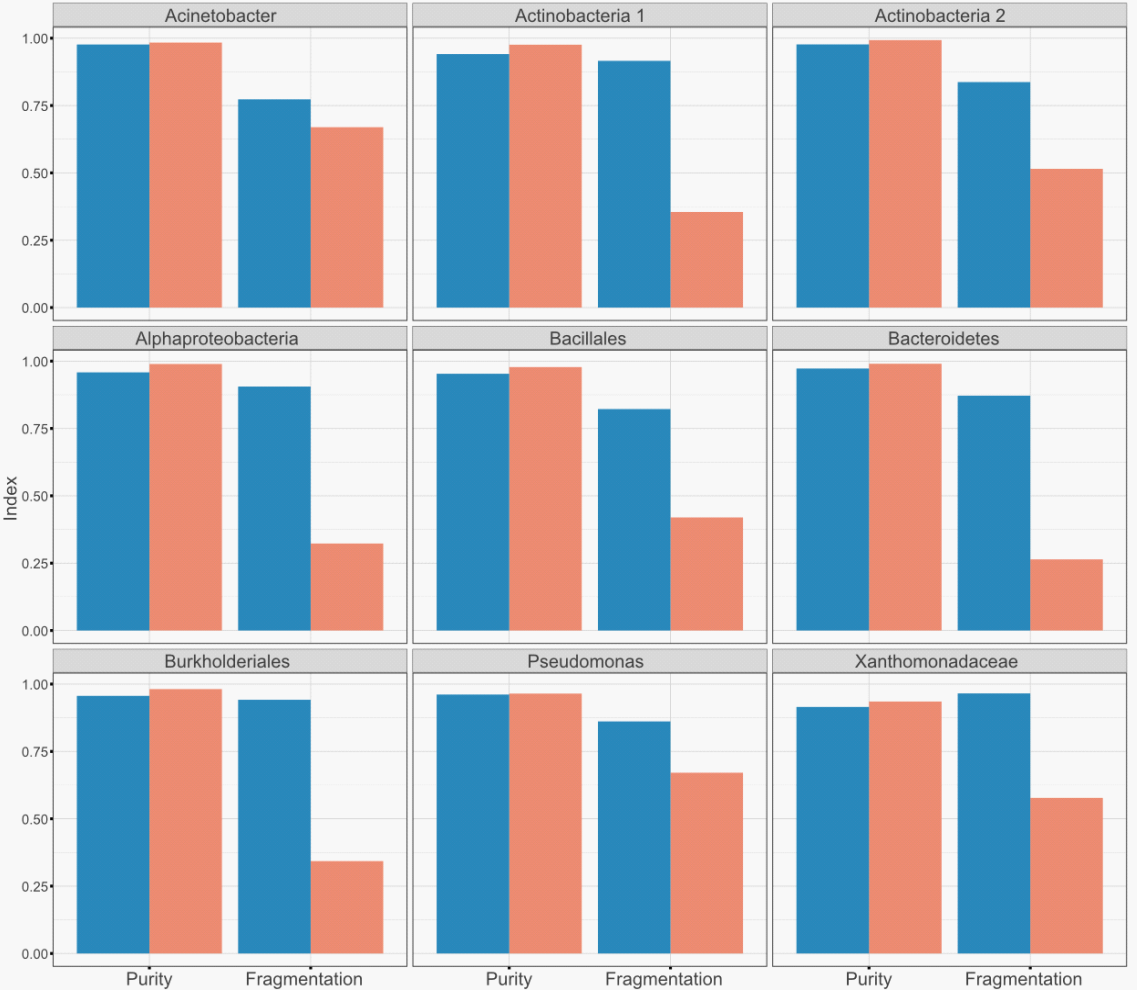
Calling PA and NPA gene operons:

Plant domains found within PA genes:

Plant genome

PA bacterial genome

Figure S4. Overview of the algorithm used to call PA and NPA genes (proteins) and gene operons. High quality PA and NPA genomes were collected. All protein and protein domains were retrieved from genomes. Different protein/domain clustering approaches were used based on existing functional annotation (COG, Pfam, TIGRfam, KEGG orthology) or based on running OrthoFinder over all protein coding genes (for simplicity TIGRfam and KEGG orthology were not mentioned in the figure). Note that clusters may contain a combination of orthologous and paralogous genes. Significant PA/NPA clusters (enriched with PA/NPA proteins/domains) were called based on five tests: PhyloGLM and the Hypergeometric test, both gene copy number and gene presence/absence versions (phyloglmcn, phyloglmbin, hypergcn, hypergbin), and Scoary. Genes from PA and NPA genomes in PA and NPA clusters, respectively, are marked with a triangle. Genes from the significant protein clusters (OrthoFinder, COG) were separately used to predict PA/NPA gene operons comprised of nearly exclusively adjacent PA/NPA genes sharing the same orientation. PA Pfam domains were used to search the overlap between those and plant-like protein domains (PREPARADOs).



Taxon	Number of Markers
Acinetobacter	570
Actinobacteria 1	436
Actinobacteria 2	436
Alphaproteobacteria	400
Bacillales	336
Bacteroidetes	430
Burkholderiales	606
Pseudomonas	582
Xanthomonadaceae	578

Figure S5. Orthofinder exhibits lower clustering fragmentation than UCLUST.

We detected taxon specific markers in the 3837 genomes utilized in this study using Phyla_Amphora[18]. Ideally, all the coding sequences (CDS) that were identified as a homologue of a particular Phyla_Amphora marker should be clustered in a single orthogroup. Using this logic, we derived two metrics from the Phyla_Amphora markers detected to quantify the quality of the orthogroups output by UCLUST and Orthofinder. The purity index quantifies how many CDS are contained in all the orthogroups needed to cover the total number of CDS identified by Phyla_Amphora as homologues of a specific marker. The fragmentation index quantifies the total number of orthogroups needed to recover the total number of homologues identified by Phyla_Amphora of a specific marker. For ease of visualization both metrics are transformed to span values between 0 to 1, values closer to 1 denote higher purity and integrity of the orthogroups analyzed. The data and scripts utilized to compute these measurements across all taxa are available on: https://github.com/isaisg/gfobap/tree/master/phyla_amphora_benchmark.

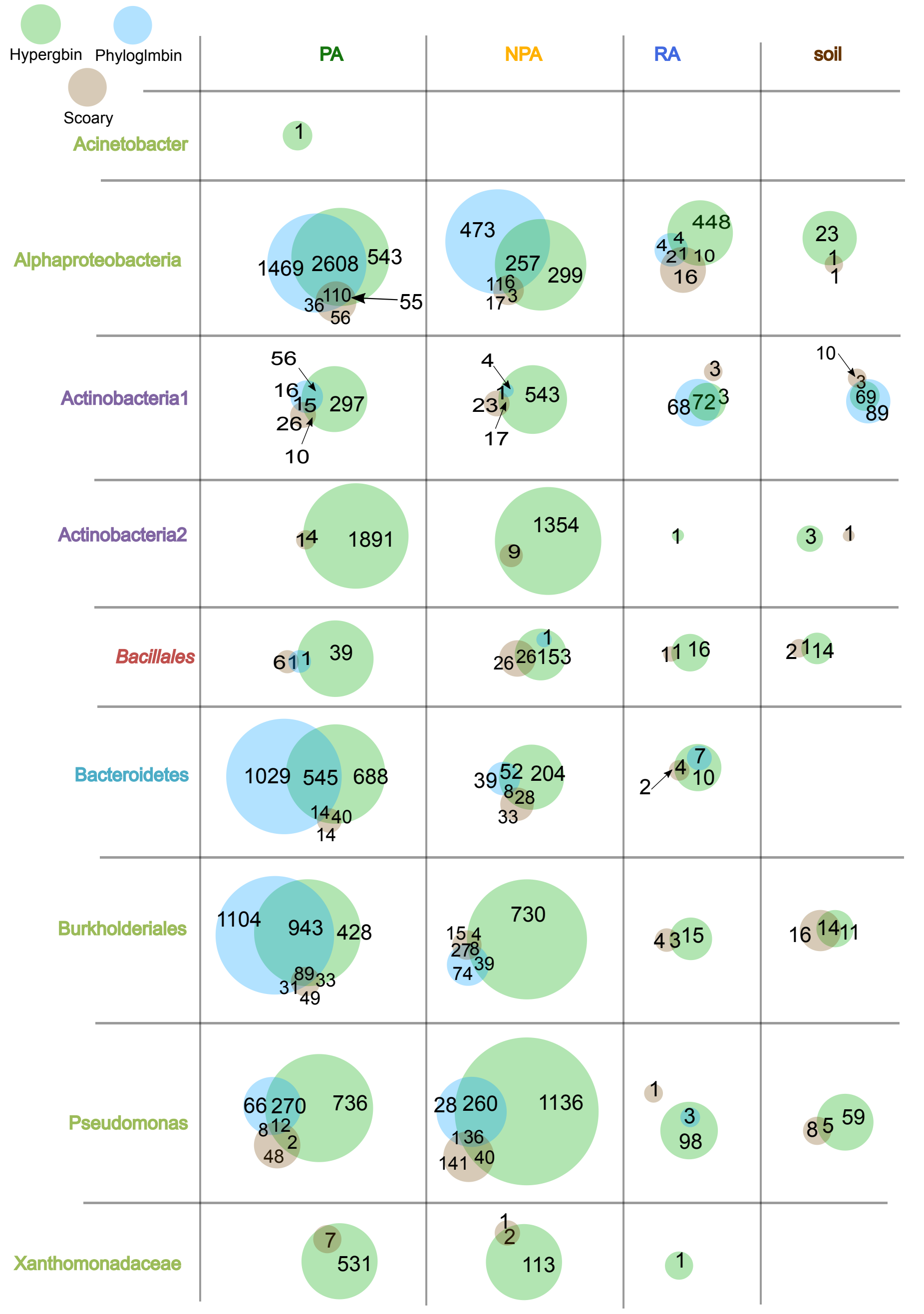

Hypergbin
PhylogImbin
Scoary
PA
NPA
RA
soil
Acinetobacter
1
Alphaproteobacteria
1469
2608
543
110
36
56
55
473
257
299
116
17
3
448
4
4
21
10
16
23
1
1
Actinobacteria1
56
16
15
297
26
10
4
1
543
23
17
3
3
68
72
3
10
3
69
89
Actinobacteria2
14
1891
1354
9
1
3
1
Bacillales
6
1
1
39
1
26
26
153
11
16
2
1
14
Bacteroidetes
1029
545
688
14
40
14
39
52
204
8
28
33
2
4
7
10
Burkholderiales
1104
943
428
89
33
31
49
730
15
4
27
8
74
39
4
3
15
16
14
11
Pseudomonas
66
270
736
8
12
2
48
28
260
1136
1
36
141
40
1
3
98
8
5
59
Xanthomonadaceae
7
531
1
2
113
1

Figure S6. Number of significant orthogroups predicted by the presence/absence (binary) versions of the Hypergeometric test (hypergbin) and PhyloGLM (phyloglmbin), and by Scoary. The numbers represent gene clusters found either in each group separately (where circles do not overlap) or in the overlap between the groups. Hypergbin is likely the most promiscuous and sensitive approach as it predict enriched genes in high numbers and does not require a phylogenetic signal (monophyletic genes can be significant in hypergbin). It may lead to many false positive predictions. Phyloglmbin is more stringent than hypergbin but it may be less sensitive than hypergbin as it cannot predict any significant gene in certain taxa that lack sufficiently strong phylogenetic signal (e.g. Actinobacteria2, Xanthomonadaceae). Scoary is probably the most stringent approach that combines a naive statistical test, a phylogenetic test, and label permutations. Therefore it frequently yields the lowest number of significant predictions.

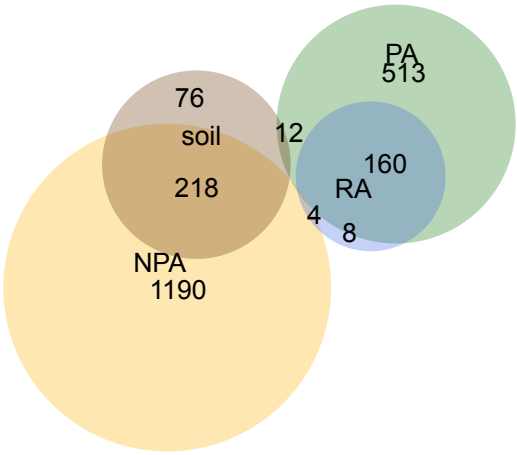
hypergcn
phyloglmcn
Scoary

	PA	NPA	RA	soil
Acinetobacter	7	6	6	1
Alphaproteobacteria	2742, 922, 325, 39, 44, 86, 88	71, 508, 1570, 14, 11, 12	69, 49, 85, 5, 1, 4, 19	253, 2
Actinobacteria1	122, 356, 2, 29, 11	2, 2, 24, 17, 303	2, 1, 61, 54, 27	39, 5, 8, 70
Actinobacteria2	14, 1672	1099, 9	9	49, 1
Bacillales	6, 1, 1, 136	1, 23, 29, 153	11, 73	3, 31
Bacteroidetes	908, 398, 286, 29, 16, 21, 2	21, 34, 385, 8, 20, 41	51, 43	12
Burkholderiales	1176, 522, 210, 72, 35, 13, 82	13, 1006, 10, 31, 35, 55	4, 3, 56	6, 24, 46
Pseudomonas	121, 183, 723, 5, 13, 4, 48	43, 197, 1044, 1, 17, 140, 60	2, 27	5, 8, 77
Xanthomonadaceae	7, 179	1, 2, 188	6	4

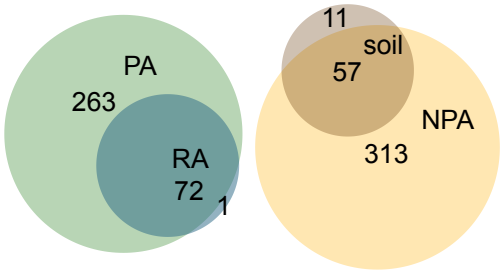
Figure S7. Number of significant orthogroups predicted by the copy number versions of the Hypergeometric test (hypergcn) and PhyloGLM (phyloglmcn), and by Scoary. The numbers represent gene clusters found either in each group separately (where circles do not overlap) or in the overlap between the groups. Hypergcn is likely the most promiscuous yet sensitive approach as it predict enriched genes in high numbers and does not require a phylogenetic signal (monophyletic genes can be significant in hypergbin). It may lead to many false positive predictions. Phyloglmcn is more stringent than hypergcn but it may be less sensitive than hypergcn as it cannot predict any significant gene in certain taxa that lack sufficiently strong phylogenetic signal (e.g. Actinobacteria2, Xanthomonadaceae). Scoary is probably the most stringent approach that combines a naive statistical test, a phylogenetic test, and label permutations. Therefore it frequently yields the lowest number of significant predictions.

Acinetobacter

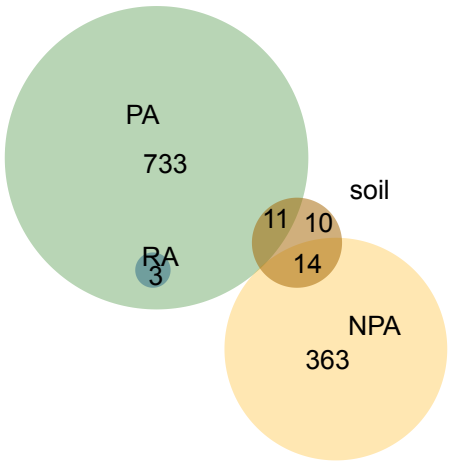
Alphaproteobacteria



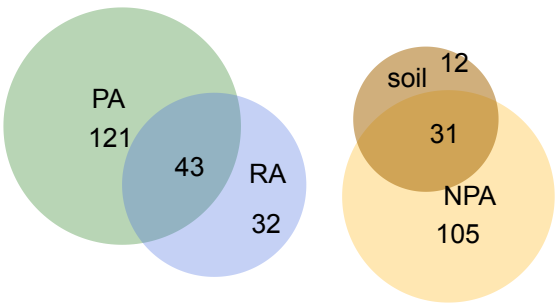
Actinobacteria1



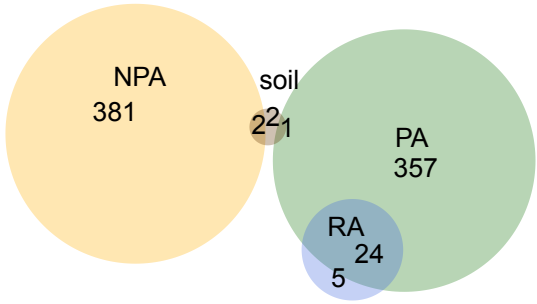
Actinobacteria2



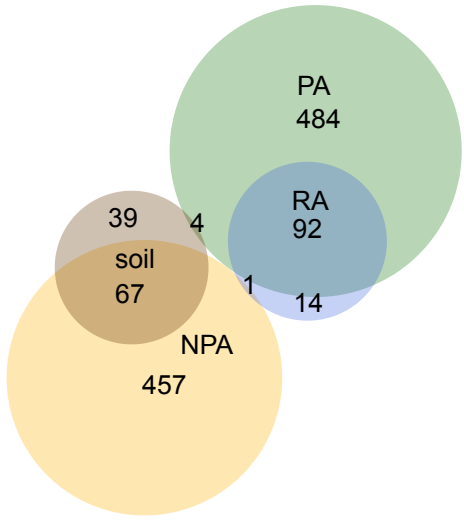
Bacillales



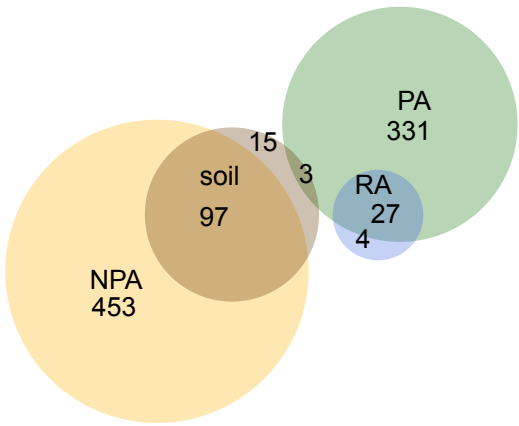
Bacteroidetes



Burkholderiales



Pseudomonas



Xanthomonadaceae

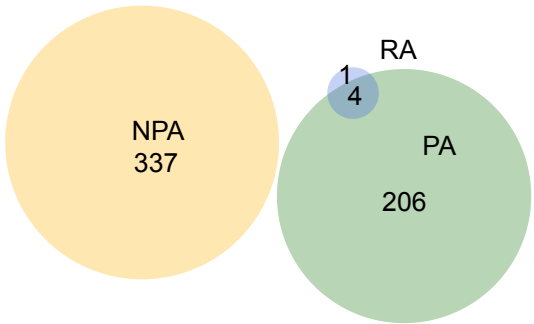
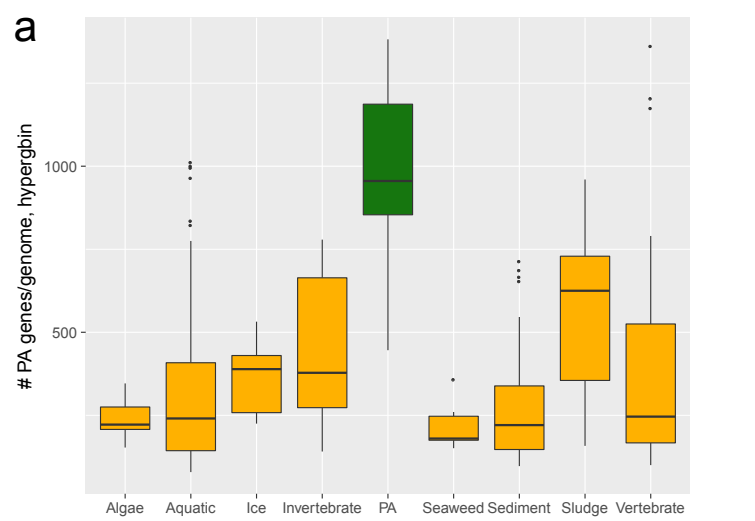
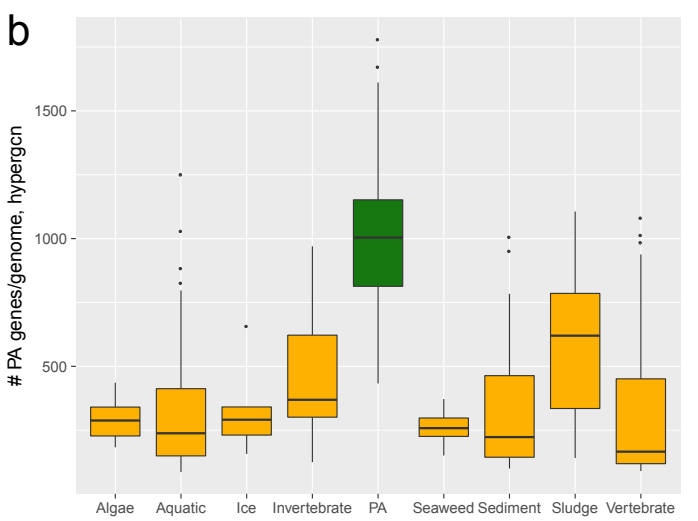
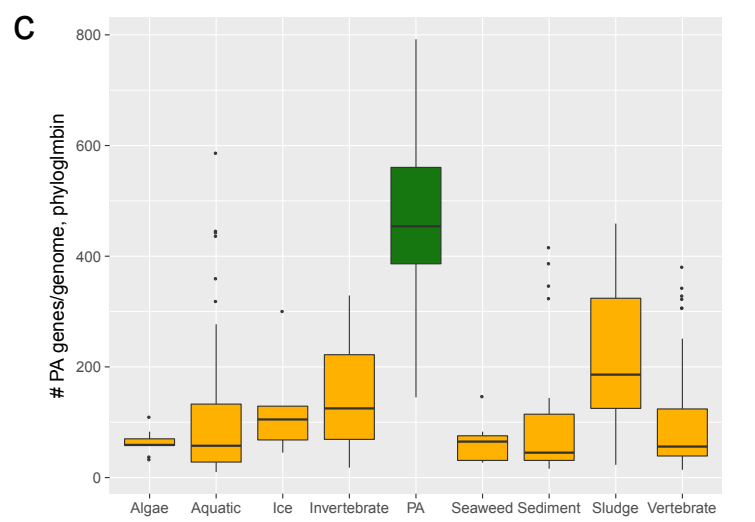
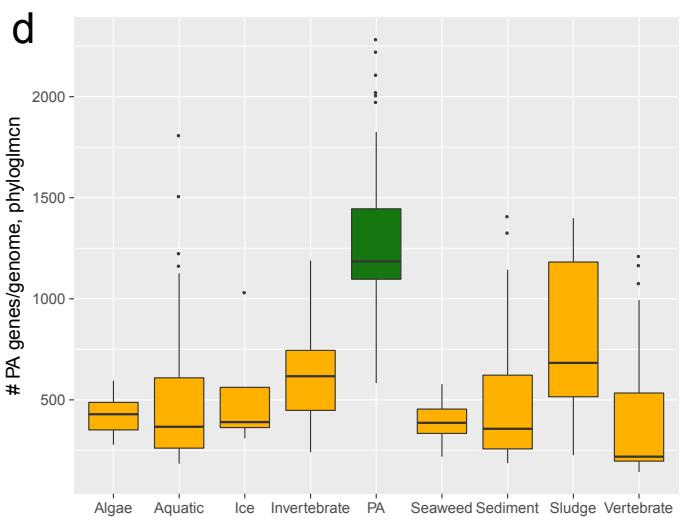


Figure S8. Euler diagrams of the significant PA (green), NPA (yellow), RA (blue), and soil (brown) COGs of all nine analyzed taxa. Statistical significance was estimated by hypergcn.


a
PA genes/genome, hypergbin
1000
500
Algae
Aquatic
Ice
Invertebrate
PA
Seaweed
Sediment
Sludge
Vertebrate


b
PA genes/genome, hypergcn
1500
1000
500
Algae
Aquatic
Ice
Invertebrate
PA
Seaweed
Sediment
Sludge
Vertebrate


c
PA genes/genome, phyloglmbin
800
600
400
200
0
Algae
Aquatic
Ice
Invertebrate
PA
Seaweed
Sediment
Sludge
Vertebrate


d
PA genes/genome, phyloglmcn
2000
1500
1000
500
Algae
Aquatic
Ice
Invertebrate
PA
Seaweed
Sediment
Sludge
Vertebrate

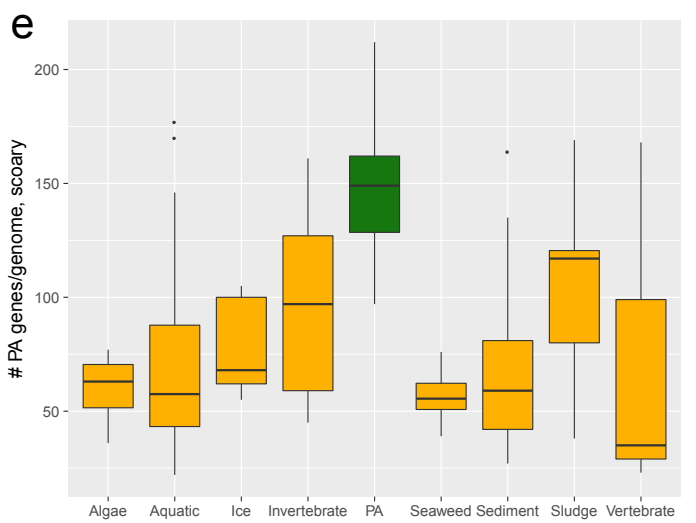

e
PA genes/genome, scoary
200
150
100
50
Algae
Aquatic
Ice
Invertebrate
PA
Seaweed
Sediment
Sludge
Vertebrate

Figure S9. Copy number of PA genes in PA and NPA genomes of Bacteroidetes. PA Genes were predicted by a. hypergbin, **b.** hypergcn, **c.** phyloglmbin**, d.** phyloglmcn**, e.** scoary**.** PA genes are more abundant in PA genomes than in NPA genomes from each of the different environments (*t*-test, $p < 0.05$).

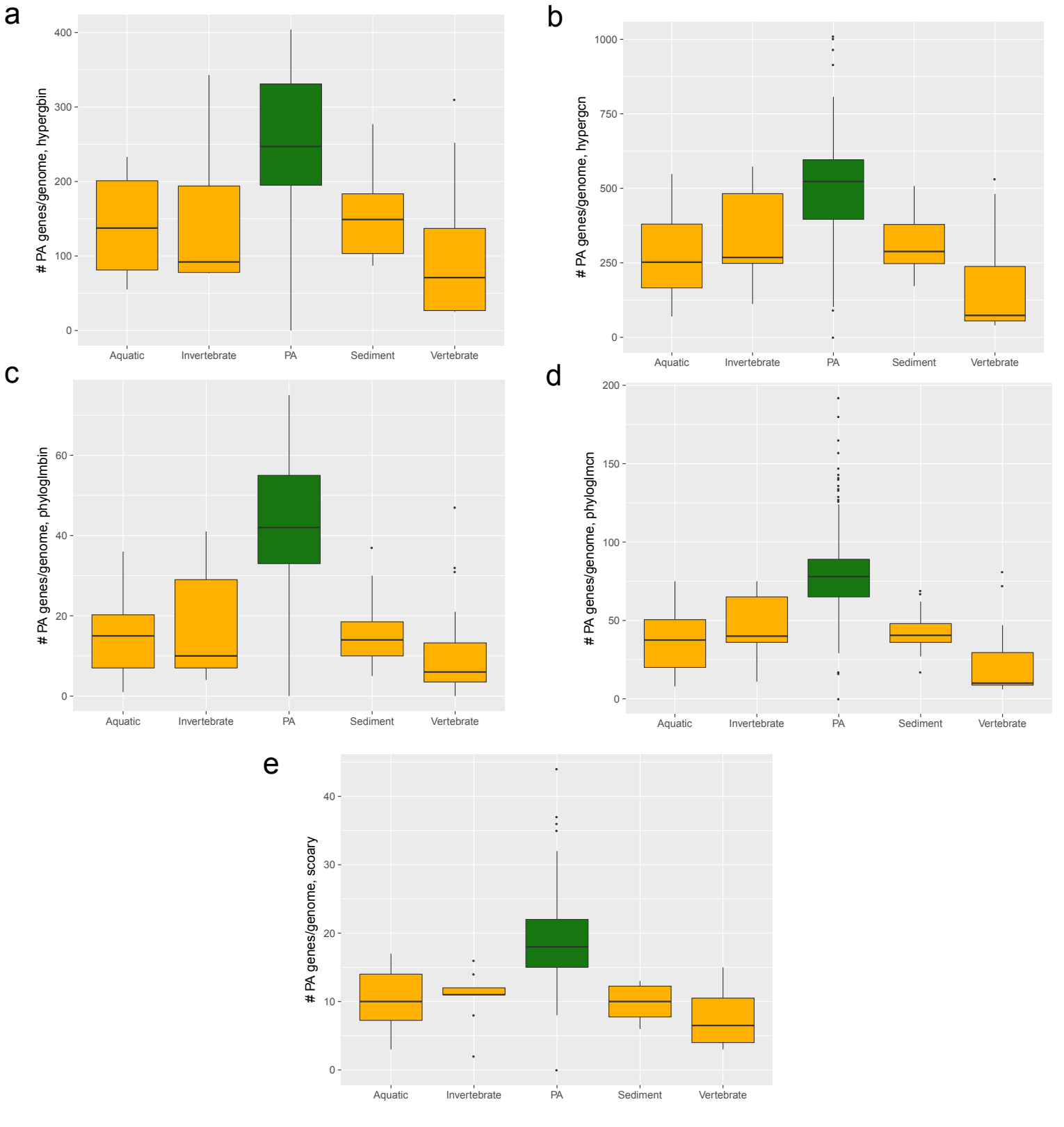
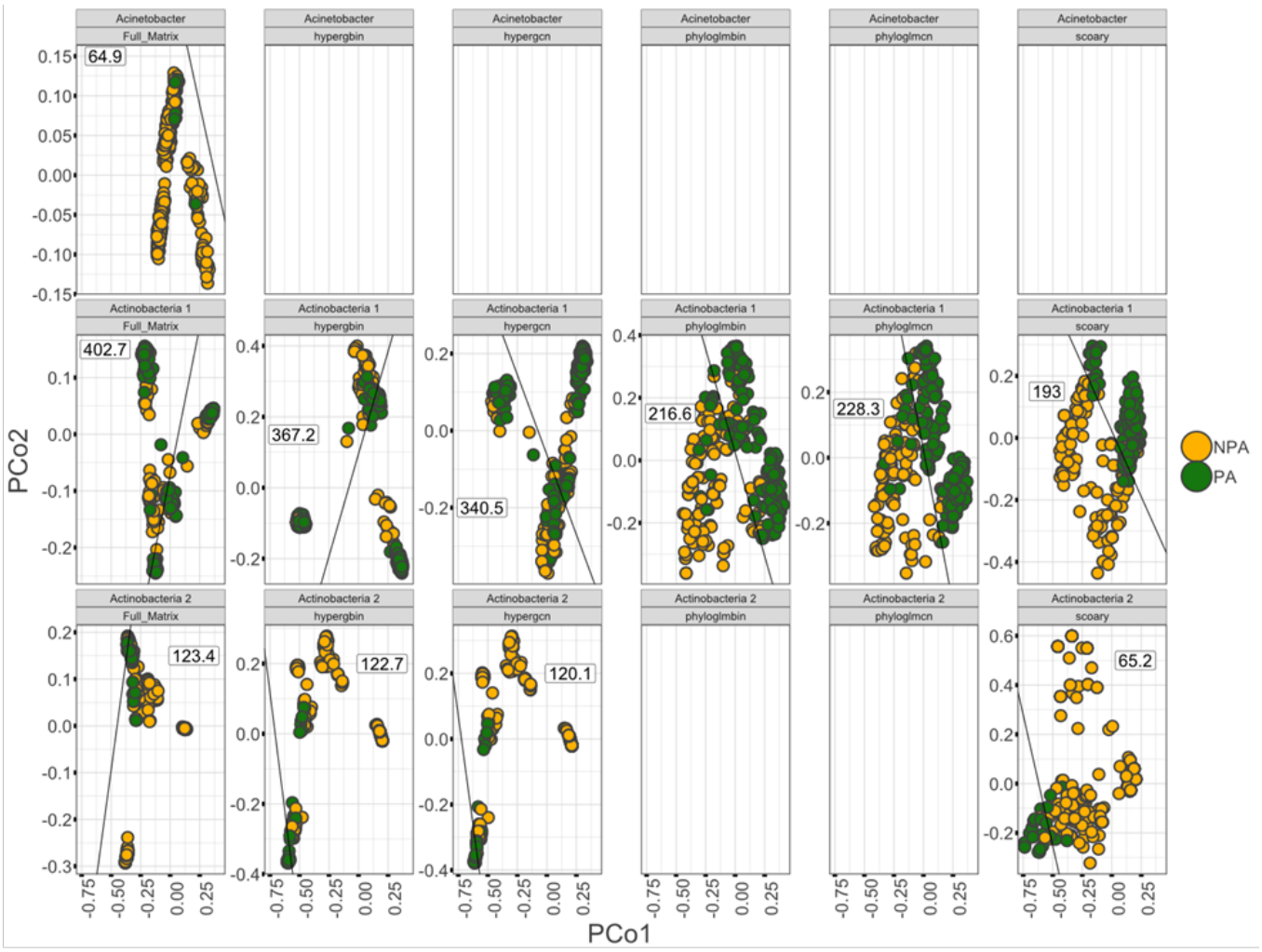
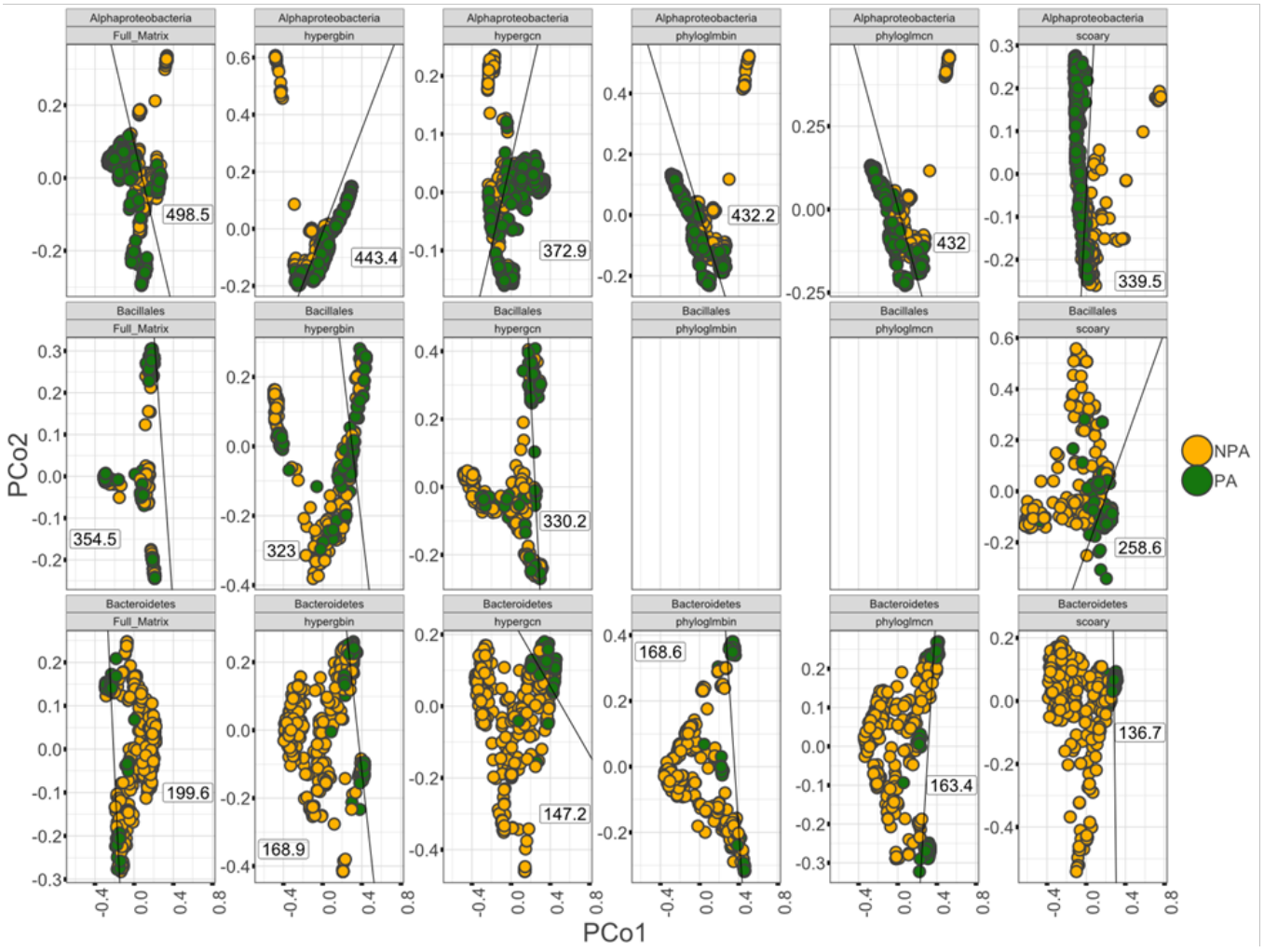

a
PA genes/genome, hypergbin
400
300
200
100
0
Aquatic
Invertebrate
PA
Sediment
Vertebrate
b
PA genes/genome, hypergcn
1000
750
500
250
0
Aquatic
Invertebrate
PA
Sediment
Vertebrate
c
PA genes/genome, phyloglmbin
60
40
20
0
Aquatic
Invertebrate
PA
Sediment
Vertebrate
d
PA genes/genome, phyloglmcn
200
150
100
50
0
Aquatic
Invertebrate
PA
Sediment
Vertebrate
e
PA genes/genome, scoary
40
30
20
10
0
Aquatic
Invertebrate
PA
Sediment
Vertebrate

Figure S10. Copy number of PA genes in PA and NPA genomes of Actinobacteria1. PA Genes were predicted by a. hypergbin, **b.** hypergcn, **c.** phyloglmbin**, d.** phyloglmcn**, e.** scoary**.** PA genes are more abundant in PA genomes than in NPA genomes from each of the different environments (*t*-test, $p < 0.05$).

a


Acinetobacter
Full_Matrix
64.9
Acinetobacter
hypergbin
Acinetobacter
hypergcn
Acinetobacter
phyloglmbin
Acinetobacter
phyloglmcn
Acinetobacter
scoary
Actinobacteria 1
Full_Matrix
402.7
Actinobacteria 1
hypergbin
367.2
Actinobacteria 1
hypergcn
340.5
Actinobacteria 1
phyloglmbin
216.6
Actinobacteria 1
phyloglmcn
228.3
Actinobacteria 1
scoary
193
Actinobacteria 2
Full_Matrix
123.4
Actinobacteria 2
hypergbin
122.7
Actinobacteria 2
hypergcn
120.1
Actinobacteria 2
phyloglmbin
Actinobacteria 2
phyloglmcn
Actinobacteria 2
scoary
65.2
PCo2
PCo1
NPA
PA


Alphaproteobacteria
Full_Matrix
498.5
Alphaproteobacteria
hypergbin
443.4
Alphaproteobacteria
hypergcn
372.9
Alphaproteobacteria
phyloglmbin
432.2
Alphaproteobacteria
phyloglmcn
432
Alphaproteobacteria
scoary
339.5
Bacillales
Full_Matrix
354.5
Bacillales
hypergbin
323
Bacillales
hypergcn
330.2
Bacillales
phyloglmbin
Bacillales
phyloglmcn
Bacillales
scoary
258.6
Bacteroidetes
Full_Matrix
199.6
Bacteroidetes
hypergbin
168.9
Bacteroidetes
hypergcn
147.2
Bacteroidetes
phyloglmbin
168.6
Bacteroidetes
phyloglmcn
163.4
Bacteroidetes
scoary
136.7
PCo2
PCo1
NPA
PA

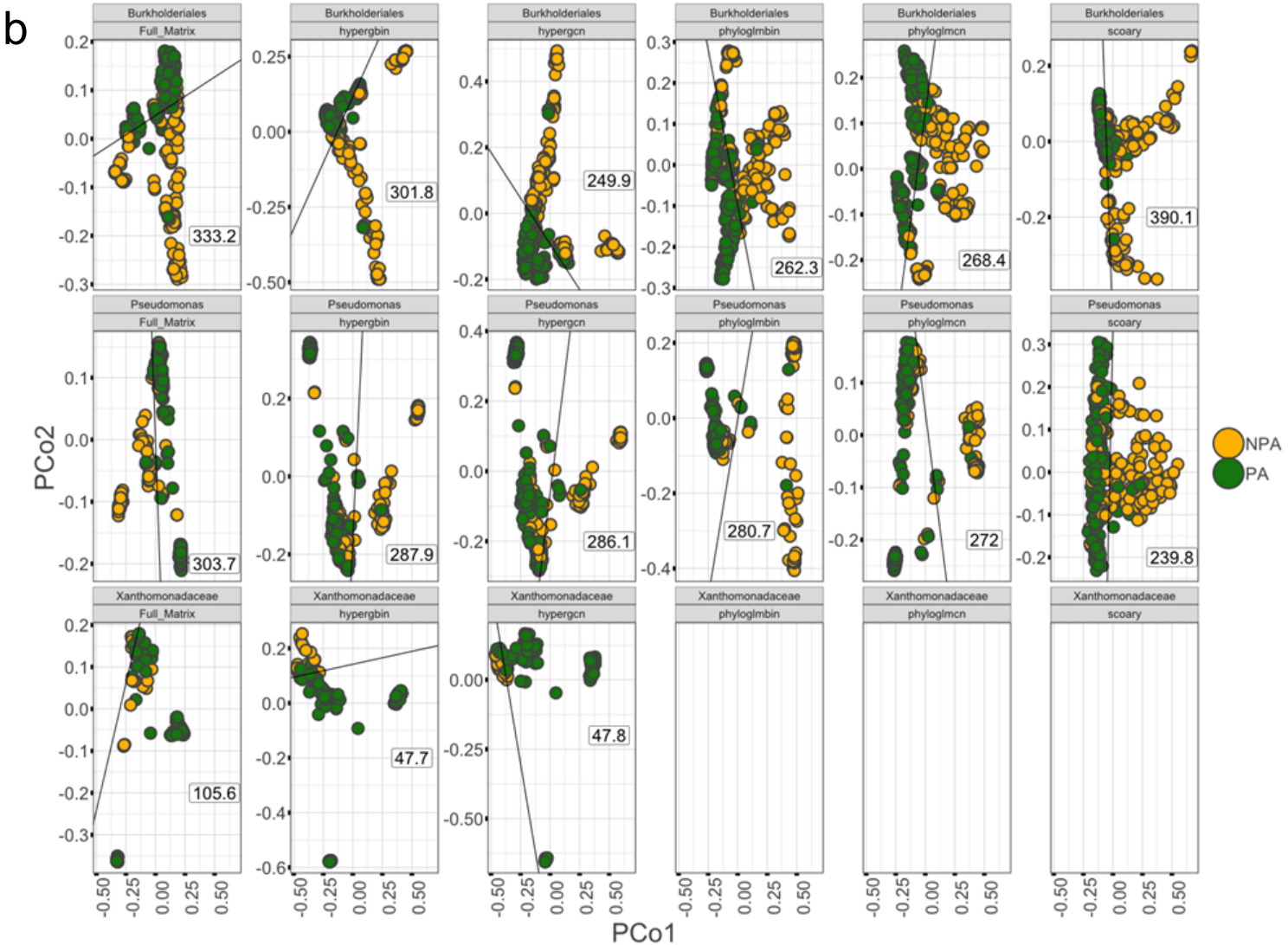

b
Burkholderiales
Full_Matrix
333.2
Burkholderiales
hypergbin
301.8
Burkholderiales
hypergcn
249.9
Burkholderiales
phyloglmbin
262.3
Burkholderiales
phyloglmcn
268.4
Burkholderiales
scoary
390.1
Pseudomonas
Full_Matrix
303.7
Pseudomonas
hypergbin
287.9
Pseudomonas
hypergcn
286.1
Pseudomonas
phyloglmbin
280.7
Pseudomonas
phyloglmcn
272
Pseudomonas
scoary
239.8
Xanthomonadaceae
Full_Matrix
105.6
Xanthomonadaceae
hypergbin
47.7
Xanthomonadaceae
hypergcn
47.8
Xanthomonadaceae
phyloglmbin
Xanthomonadaceae
phyloglmcn
Xanthomonadaceae
scoary
PCo2
PCo1
NPA
PA

Figure S11. PCoA analysis. a and **b** show different taxa. We visualized the overall contribution of statistically significantly enriched/depleted orthogroups to the differentiation of PA and NPA genomes based on Principal Coordinates Analysis (PCoA). We used the Canberra distance to perform PCoA over a pan genome matrix with the total number of orthogroups across a taxon (Full Matrix, including non-significant orthogroups) and pan genome matrices containing only the corresponding enriched/depleted orthogroups as called by the different algorithms utilized (hypergbin, hypercn, phyloglmbin, phyloglmcn and scoary). The black lines in the plots correspond to the fit of a logistic regression that modeled the binary label of each genome (PA, NPA) given the two first axes of each plot. The blank plots occur when there were not statistically significantly enriched/depleted orthogroups reported by a particular method or because the number of enriched/depleted orthogroups were not sufficient to recover the total number of genomes across a taxon. Inside each scatterplot, we include the Akaike Information Criterion (AIC) value output from the logistic regression fit. Smaller AIC values represent a better quality of the model to the data.

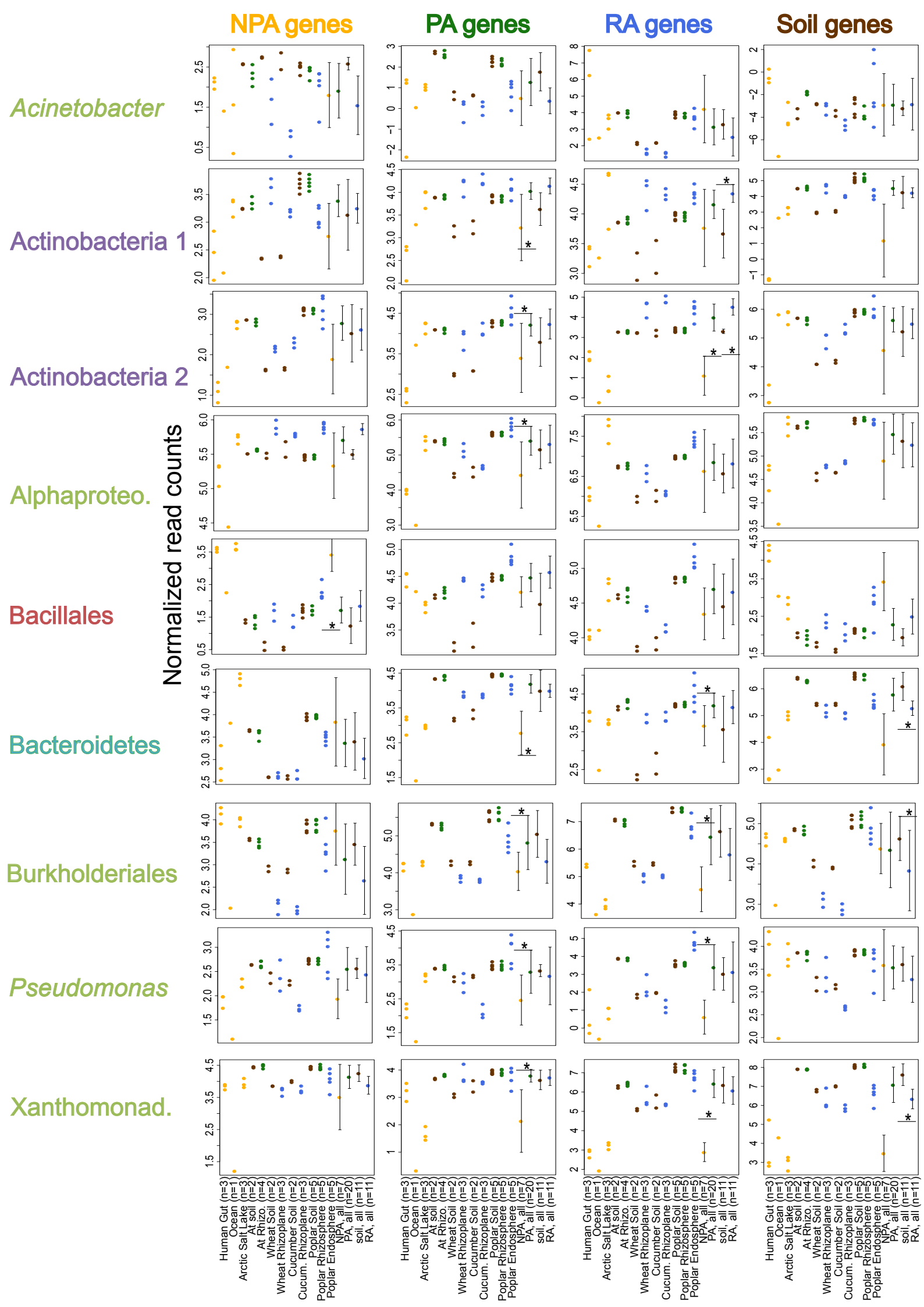

NPA genes
PA genes
RA genes
Soil genes
Acinetobacter
Actinobacteria 1
Actinobacteria 2
Alphaproteo.
Bacillales
Bacteroidetes
Burkholderiales
Pseudomonas
Xanthomonad.
Normalized read counts
Human Gut (n=3)
Ocean (n=1)
Arctic Salt Lake (n=3)
At soil (n=2)
At Rhizo. (n=4)
Wheat Soil (n=2)
Wheat Rhizoplane (n=3)
Cucumber Soil (n=2)
Cucum. Rhizoplane (n=3)
Poplar Soil (n=5)
Poplar Rhizosphere (n=5)
Poplar Endosphere (n=7)
NPA, all (n=20)
PA, all (n=7)
soil, all (n=11)
RA, all (n=11)

Figure S12. Reads from 38 shotgun metagenome samples were mapped to significant PA, NPA, RA, and soil genes predicted by the **hypergeometric test, gene copy number version (hypergcn).** Normalized reads are defined as log2((A * 10000)/(B * C)), where A is defined as hits of metagenomes reads against predicted gene set (e.g. PA genes of taxon X), B is defined as hits against phylum-specific phylogenetic marker genes (taking into account the taxon fraction within the sample), and C is defined as number of predicted gene clusters (e.g. PA gene clusters of taxon X). Samples in which either B or C equals 0 (namely either taxon is absent from the metagenome or there is an unavailable reference) and therefore could not be normalized were omitted. The *Arabidopsis*, cucumber, wheat, and poplar metagenomes were paired with the other samples from the same plant. These paired samples were taken from the same soil (cucumber and wheat samples were taken from the same pot) and were sequenced together. RA samples are also PA samples but not the other way around, because some PA samples were taken from rhizosphere which is different from our operational definition of RA: rhizoplane and endophytic compartment. The last four columns represent the distribution across all relevant samples. An asterisk above/below a boxplot pair represents a significant difference in the expected direction ($P < 0.05$, two sided *t*-test). RA genes are required to be more abundant in PA over NPA metagenomes and also in RA over soil metagenomes.

NPA genes
PA genes
RA genes
Soil genes
Acinetobacter
Actinobacteria 1
Actinobacteria 2
Alphaproteo.
Bacillales
Bacteroidetes
Burkholderiales
Pseudomonas
Xanthomonad.
Normalized Read counts
Human Gut (n=3)
Ocean (n=1)
Arctic Salt Lake (n=3)
At soil (n=2)
At Rhizo. (n=4)
Wheat Soil (n=2)
Wheat Rhizoplane (n=3)
Cucumber Soil (n=2)
Cucum. Rhizoplane (n=3)
Poplar Soil (n=5)
Poplar Rhizosphere (n=5)
Poplar Endosphere (n=5)
NPA, all (n=7)
PA, all (n=20)
soil, all (n=11)
RA, all (n=11)

Figure S13. Reads from 38 shotgun metagenome samples were mapped to significant PA, NPA, RA, and soil genes predicted by the **hypergeometric test, gene presence/absence version (hypergbin).** Normalized reads are defined as log2((A * 10000)/(B * C)), where A is defined as hits of metagenomes reads against predicted gene set (e.g. PA genes of taxon X), B is defined as hits against phylum-specific phylogenetic marker genes (taking into account the taxon fraction within the sample), and C is defined as number of predicted gene clusters (e.g. PA gene clusters of taxon X). Samples in which either B or C equals 0 (namely either taxon is absent from the metagenome or there is an unavailable reference) and therefore could not be normalized were omitted. The *Arabidopsis*, cucumber, wheat, and poplar metagenomes were paired with the other samples from the same plant. These paired samples were taken from the same soil (cucumber and wheat samples were taken from the same pot) and were sequenced together. RA samples are also PA samples but not the other way around, as some PA samples were taken from rhizosphere which is different from our operational definition of RA: rhizoplane and endophytic compartment. The last four columns represent the distribution across all relevant samples. An asterisk above/below a boxplot pair represents a significant difference in the expected direction ($P < 0.05$, two sided *t*-test). RA genes are required to be more abundant in PA over NPA metagenomes and also in RA over soil metagenomes.

NPA genes
PA genes
RA genes
Soil genes
Acinetobacter
Actinobacteria 1
Actinobacteria 2
Alphaproteo.
Bacillales
Bacteroidetes
Burkholderiales
Pseudomonas
Xanthomonad.
Normalized Read scoary
Human Gut (n=3)
Ocean (n=1)
Arctic Salt Lake (n=3)
At soil (n=2)
At Rhizo. (n=4)
Wheat Soil (n=2)
Wheat Rhizoplane (n=3)
Cucumber Soil (n=2)
Cucum. Rhizoplane (n=3)
Poplar Soil (n=5)
Poplar Rhizosphere (n=5)
Poplar Endosphere (n=5)
NPA, all (n=7)
PA, all (n=20)
soil, all (n=11)
RA, all (n=11)

Figure S14. Reads from 38 shotgun metagenome samples were mapped to significant PA, NPA, RA, and soil genes predicted by **Scoary.** Normalized reads are defined as log2((A * 10000)/(B * C)), where A is defined as hits of metagenomes reads against predicted gene set (e.g. PA genes of taxon X), B is defined as hits against phylum-specific phylogenetic marker genes (taking into account the taxon fraction within the sample), and C is defined as number of predicted gene clusters (e.g. PA gene clusters of taxon X). Samples in which either B or C equals 0 (namely either taxon is absent from the metagenome or there is an unavailable reference) and therefore could not be normalized were omitted. The *Arabidopsis*, cucumber, wheat, and poplar metagenomes were paired with the other samples from the same plant. These paired samples were taken from the same soil (cucumber and wheat samples were taken from the same pot) and were sequenced together. RA samples are also PA samples but not the other way around, as some PA samples were taken from rhizosphere which is different from our operational definition of RA: rhizoplane and endophytic compartment. The last four columns represent the distribution across all relevant samples. An asterisk above/below a boxplot pair represents a significant difference in the expected direction ($P < 0.05$, two sided *t*-test). RA genes are required to be more abundant in PA over NPA metagenomes and also in RA over soil metagenomes.

NPA genes
PA genes
RA genes
Soil genes
Acinetobacter
Actinobacteria 1
Actinobacteria 2
Alphaproteo.
Bacillales
Bacteroidetes
Burkholderiales
Pseudomonas
Xanthomonad.
Normalized Read counts
Human Gut (n=3)
Ocean (n=1)
Arctic Salt Lake (n=3)
At soil (n=2)
At Rhizo. (n=4)
Wheat Soil (n=2)
Wheat Rhizoplane (n=3)
Cucumber Soil (n=2)
Cucum. Rhizoplane (n=3)
Poplar Soil (n=5)
Poplar Rhizosphere (n=5)
Poplar Endosphere (n=5)
NPA, all (n=7)
PA, all (n=20)
soil, all (n=11)
RA, all (n=11)

Figure S15. Reads from 38 shotgun metagenome samples were mapped to significant PA, NPA, RA, and soil genes predicted by **PhyloGLM, gene copy number version (phyloglmcn).** Normalized reads are defined as log2((A * 10000)/(B * C)), where A is defined as hits of metagenomes reads against predicted gene set (e.g. PA genes of taxon X), B is defined as hits against phylum-specific phylogenetic marker genes (taking into account the taxon fraction within the sample), and C is defined as number of predicted gene clusters (e.g. PA gene clusters of taxon X). Samples in which either B or C equals 0 (namely either taxon is absent from the metagenome or there is an unavailable reference) and therefore could not be normalized were omitted. The *Arabidopsis*, cucumber, wheat, and poplar metagenomes were paired with the other samples from the same plant. These paired samples were taken from the same soil (cucumber and wheat samples were taken from the same pot) and were sequenced together. RA samples are also PA samples but not the other way around, as some PA samples were taken from rhizosphere which is different from our operational definition of RA: rhizoplane and endophytic compartment. The last four columns represent the distribution across all relevant samples. An asterisk above/below a boxplot pair represents a significant difference in the expected direction ($P < 0.05$, two sided *t*-test). RA genes are required to be more abundant in PA over NPA metagenomes and also in RA over soil metagenomes.

NPA genes
PA genes
RA genes
Soil genes
Acinetobacter
Actinobacteria 1
Actinobacteria 2
Alphaproteo.
Bacillales
Bacteroidetes
Burkholderiales
Pseudomonas
Xanthomonad.
Normalized Read counts
Human Gut (n=3)
Ocean (n=1)
Arctic Salt Lake (n=3)
At soil (n=2)
At Rhizo. (n=4)
Wheat Soil (n=2)
Wheat Rhizoplane (n=3)
Cucumber Soil (n=2)
Cucum. Rhizoplane (n=3)
Poplar Soil (n=5)
Poplar Rhizosphere (n=5)
Poplar Endosphere (n=5)
NPA, all (n=7)
PA, all (n=20)
soil, all (n=11)
RA, all (n=11)

Figure S16. Reads from 38 shotgun metagenome samples were mapped to significant PA, NPA, RA, and soil genes predicted by **PhyloGLM, gene presence/absence version (phylogImbin).** Normalized reads are defined as log2((A * 10000)/(B * C)), where A is defined as hits of metagenomes reads against predicted gene set (e.g. PA genes of taxon X), B is defined as hits against phylum-specific phylogenetic marker genes (taking into account the taxon fraction within the sample), and C is defined as number of predicted gene clusters (e.g. PA gene clusters of taxon X). Samples in which either B or C equals 0 (namely either taxon is absent from the metagenome or there is an unavailable reference) and therefore could not be normalized were omitted. The *Arabidopsis*, cucumber, wheat, and poplar metagenomes were paired with the other samples from the same plant. These paired samples were taken from the same soil (cucumber and wheat samples were taken from the same pot) and were sequenced together. RA samples are also PA samples but not the other way around, as some PA samples were taken from rhizosphere which is different from our operational definition of RA: rhizoplane and endophytic compartment. The last four columns represent the distribution across all relevant samples. An asterisk above/below a boxplot pair represents a significant difference in the expected direction ($P < 0.05$, two sided *t*-test). RA genes are required to be more abundant in PA over NPA metagenomes and also in RA over soil metagenomes.

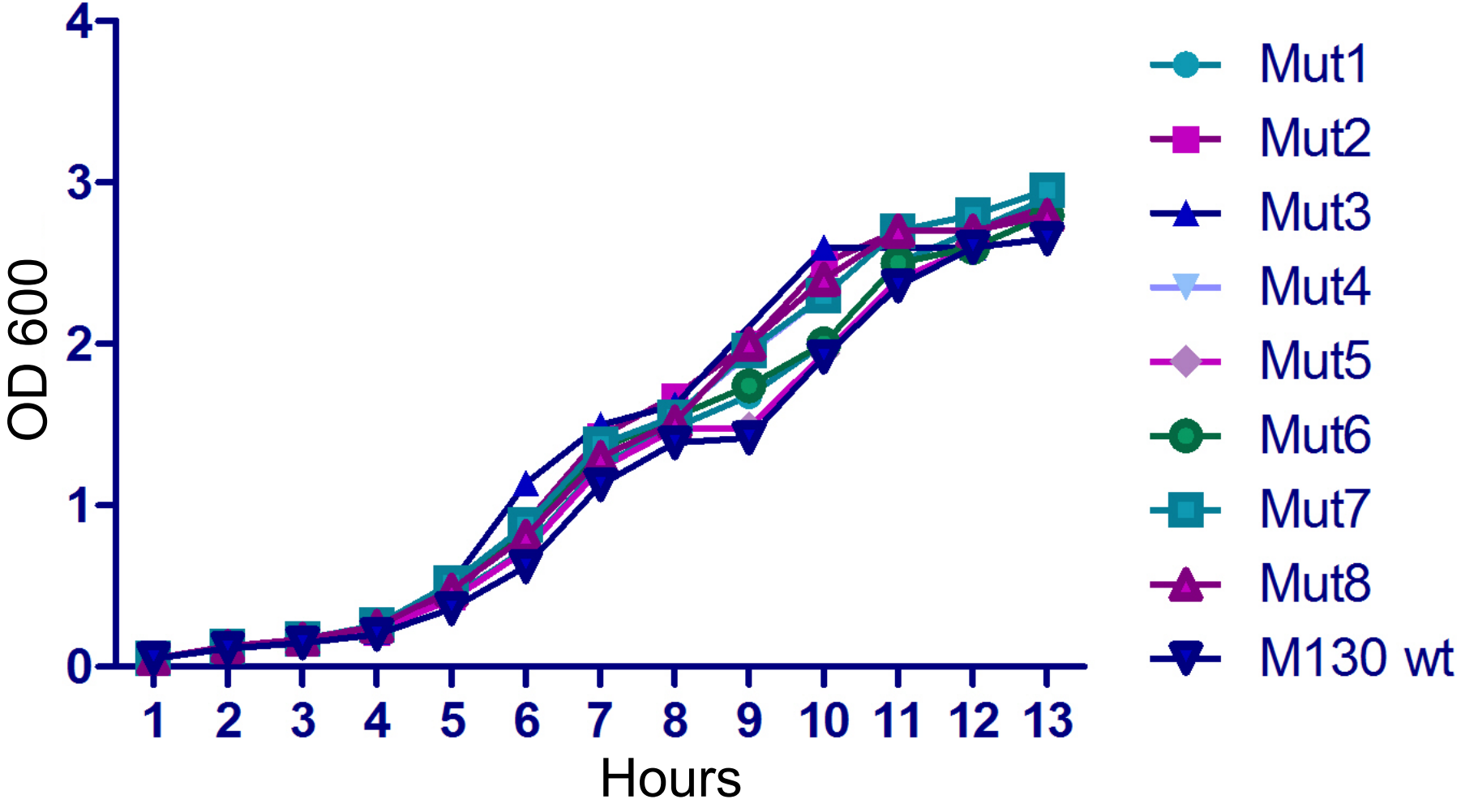

OD 600
0
1
2
3
4
1 2 3 4 5 6 7 8 9 10 11 12 13
Hours
Mut1
Mut2
Mut3
Mut4
Mut5
Mut6
Mut7
Mut8
M130 wt

Figure S17. All eight deletion mutant gene strains of putative PA genes in *Paraburkholderia kururiensis* M130 do not affect bacterial growth rates. Mutants are described in Table S17.

Sense, upstream

Sense, downstream

Antisense, upstream and downstream

Actinobacteria 1

Actinobacteria 2

Alphaproteobacteria

Bacillales

Bacteroidetes

Burkholderiales

Pseudomonas

Xanthomonadaceae

per category

neighbor genes of LacI TF / # genes

Cytoskeleton
Mobilome: prophages, transposons
Extracellular structures
Intracellular trafficking, secretion, and vesicular transport
Defense mechanisms
Signal transduction mechanisms
Function unknown
General function prediction only
Secondary metabolites biosynthesis, transport and catabolism
Inorganic ion transport and metabolism
Posttranslational modification, protein turnover, chaperones
Cell motility
Cell wall/membrane/envelope biogenesis
Replication, recombination and repair
Translation, ribosomal structure and biogenesis
Transcription
Lipid transport and metabolism
Coenzyme transport and metabolism
Carbohydrate transport and metabolism
Nucleotide transport and metabolism
Amino acid transport and metabolism
Cell cycle control, cell division, chromosome partitioning
Energy production and conversion
Chromatin structure and dynamics
RNA processing and modification

Figure S18. Direct neighbors of *LacI*-family transcription factor genes. Genes next to *LacI*-family genes were retrieved in all possible orientations relative to the *LacI*-family gene: upstream sense, downstream sense, or antisense (head-to-head and tail-to-tail). The COG annotation of each gene was retrieved and was translated into COG category.

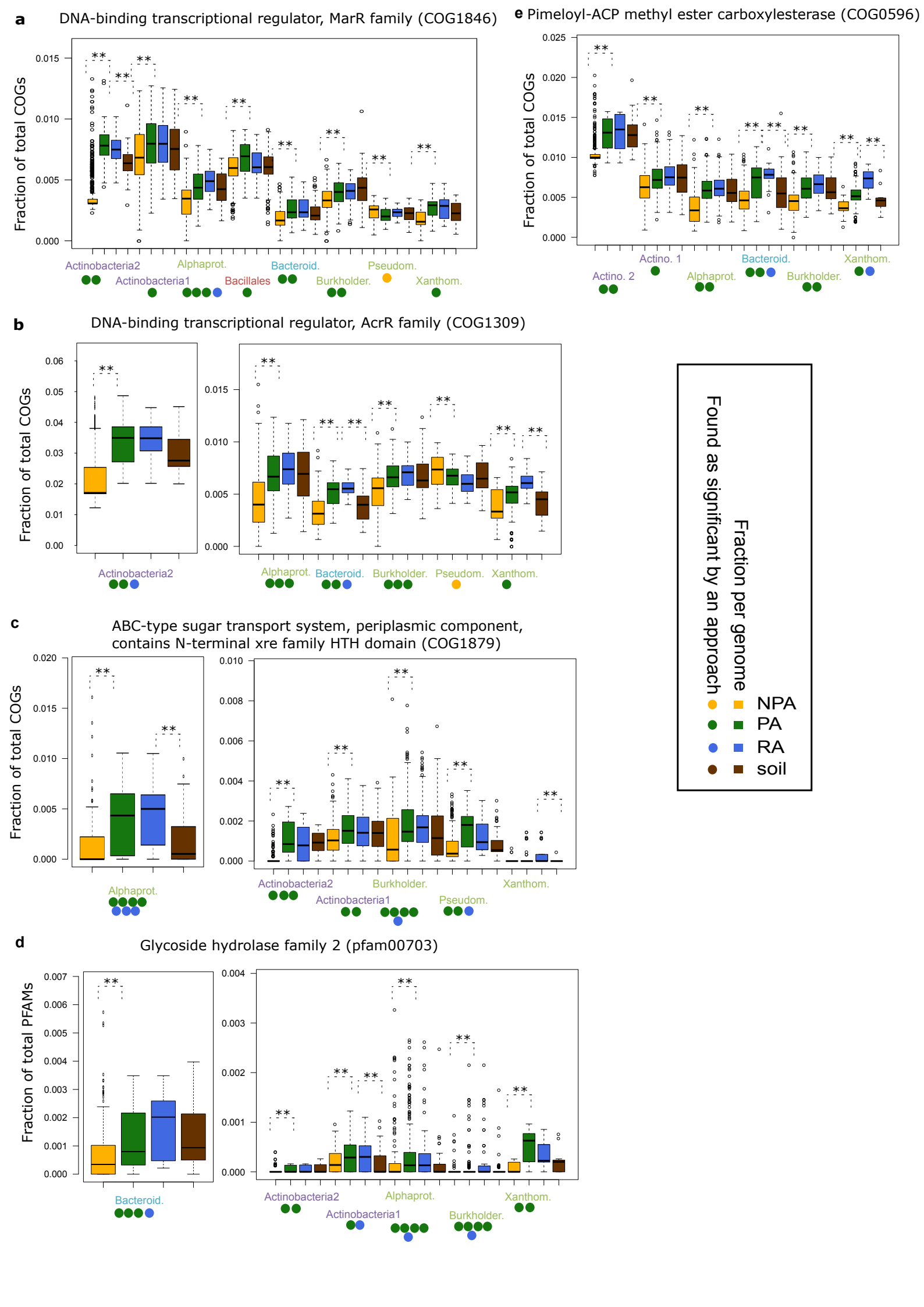

a
DNA-binding transcriptional regulator, MarR family (COG1846)
Fraction of total COGs
Actinobacteria2
Actinobacteria1
Alphaprot.
Bacillales
Bacteroid.
Burkholder.
Pseudom.
Xanthom.
e Pimeloyl-ACP methyl ester carboxylesterase (COG0596)
Fraction of total COGs
Actino. 2
Actino. 1
Alphaprot.
Bacteroid.
Burkholder.
Xanthom.
b
DNA-binding transcriptional regulator, AcrR family (COG1309)
Fraction of total COGs
Actinobacteria2
Alphaprot.
Bacteroid.
Burkholder.
Pseudom.
Xanthom.
Found as significant by an approach
Fraction per genome
NPA
PA
RA
soil
c
ABC-type sugar transport system, periplasmic component,
contains N-terminal xre family HTH domain (COG1879)
Fraction of total COGs
Alphaprot.
Actinobacteria2
Actinobacteria1
Burkholder.
Pseudom.
Xanthom.
d
Glycoside hydrolase family 2 (pfam00703)
Fraction of total PFAMs
Bacteroid.
Actinobacteria2
Actinobacteria1
Alphaprot.
Burkholder.
Xanthom.

Figure S19. Various protein and protein domains enriched in PA and RA bacteria from multiple taxa and occasionally identified by multiple approaches. Double asterisks indicate a significant difference between the compared groups ($P < 0.05$, *t*-test). Filled circles below each axis denotes the number of approaches in which the protein/domain was found to be significant (maximum is five).

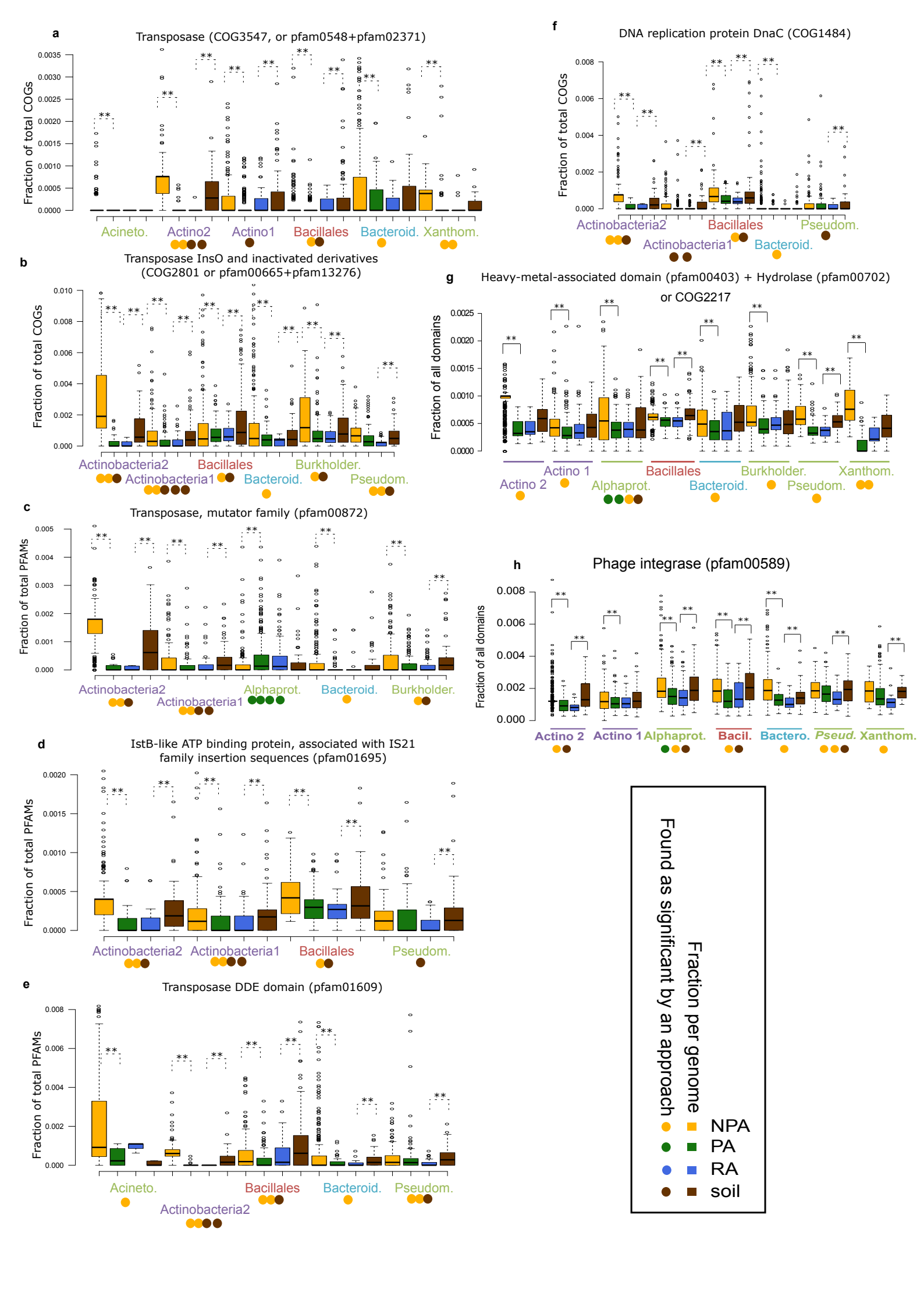

a
Transposase (COG3547, or pfam0548+pfam02371)
Fraction of total COGs
Acineto.
Actino2
Actino1
Bacillales
Bacteroid.
Xanthom.
b
Transposase InsO and inactivated derivatives
(COG2801 or pfam00665+pfam13276)
Fraction of total COGs
Actinobacteria2
Actinobacteria1
Bacillales
Bacteroid.
Burkholder.
Pseudom.
c
Transposase, mutator family (pfam00872)
Fraction of total PFAMs
Actinobacteria2
Actinobacteria1
Alphaprot.
Bacteroid.
Burkholder.
d
IstB-like ATP binding protein, associated with IS21
family insertion sequences (pfam01695)
Fraction of total PFAMs
Actinobacteria2
Actinobacteria1
Bacillales
Pseudom.
e
Transposase DDE domain (pfam01609)
Fraction of total PFAMs
Acineto.
Actinobacteria2
Bacillales
Bacteroid.
Pseudom.
f
DNA replication protein DnaC (COG1484)
Fraction of total COGs
Actinobacteria2
Actinobacteria1
Bacillales
Bacteroid.
Pseudom.
g
Heavy-metal-associated domain (pfam00403) + Hydrolase (pfam00702)
or COG2217
Fraction of all domains
Actino 2
Actino 1
Alphaprot.
Bacillales
Bacteroid.
Burkholder.
Pseudom.
Xanthom.
h
Phage integrase (pfam00589)
Fraction of all domains
Actino 2
Actino 1
Alphaprot.
Bacil.
Bactero.
Pseud.
Xanthom.
Found as significant by an approach
Fraction per genome
NPA
PA
RA
soil

Figure S20. Various protein and protein domains enriched in NPA and soil bacteria from multiple taxa and occasionally identified by multiple approaches. Double asterisks indicate a significant difference between the compared groups ($P < 0.05$, *t*-test). Filled circles below each axis denote the number of approaches in which the protein/domain was found to be significant (maximum is five).

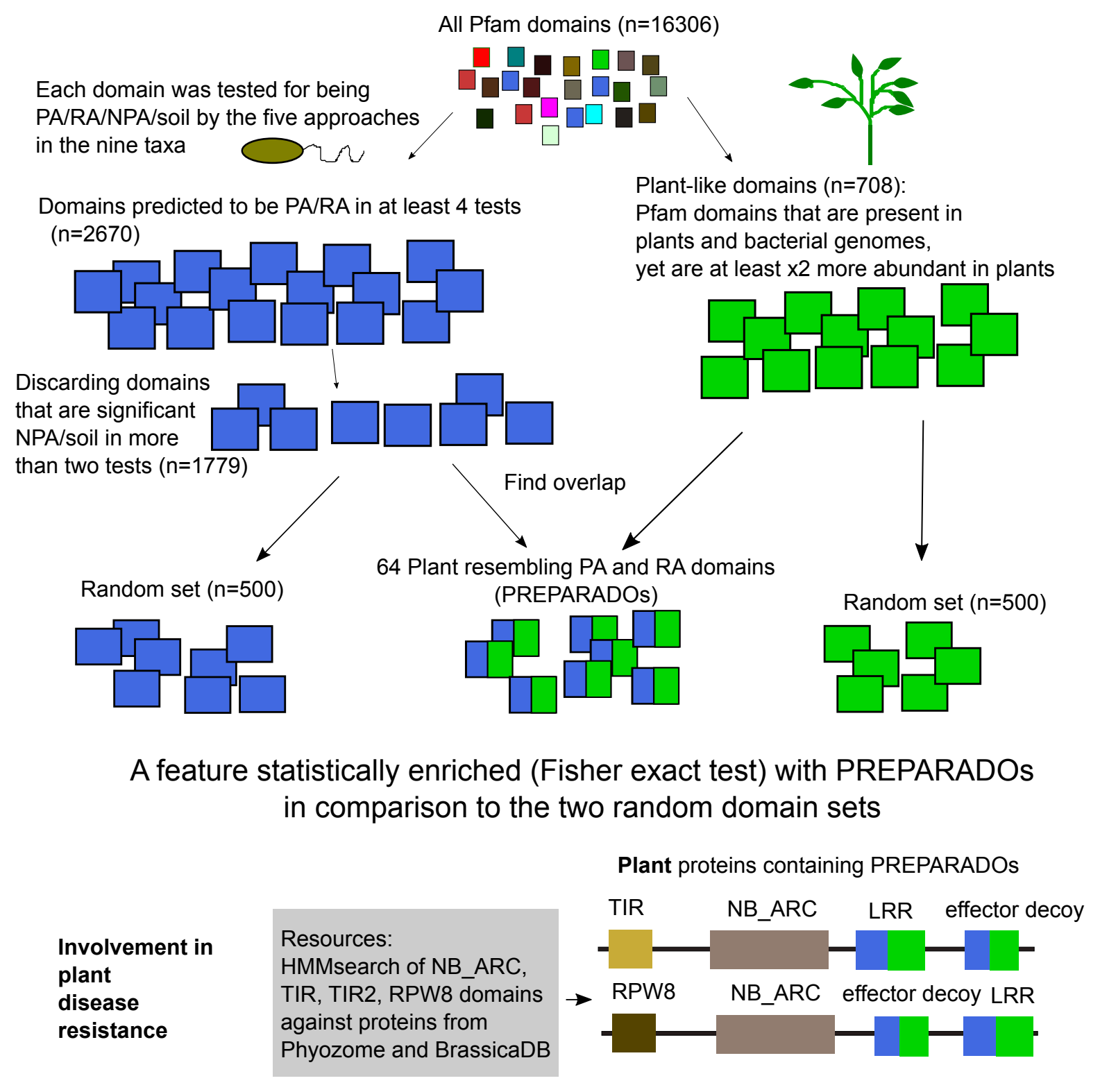

All Pfam domains (n=16306)
Each domain was tested for being PA/RA/NPA/soil by the five approaches in the nine taxa
Plant-like domains (n=708): Pfam domains that are present in plants and bacterial genomes, yet are at least x2 more abundant in plants
Domains predicted to be PA/RA in at least 4 tests (n=2670)
Discarding domains that are significant NPA/soil in more than two tests (n=1779)
Find overlap
Random set (n=500)
64 Plant resembling PA and RA domains (PREPARADOs)
Random set (n=500)
A feature statistically enriched (Fisher exact test) with PREPARADOs in comparison to the two random domain sets
Involvement in plant disease resistance
Resources: HMMsearch of NB_ARC, TIR, TIR2, RPW8 domains against proteins from Phyozome and BrassicaDB
Plant proteins containing PREPARADOs
TIR
NB_ARC
LRR
effector decoy
RPW8
NB_ARC
effector decoy
LRR

Figure S21. The algorithm used to predict PREPARADO and their co-enrichment with domains common to plant disease resistance proteins of the NLR class. LRR is illustrated as a PREPARADO as LRR6 and LRR8 are also PREPARADOs.

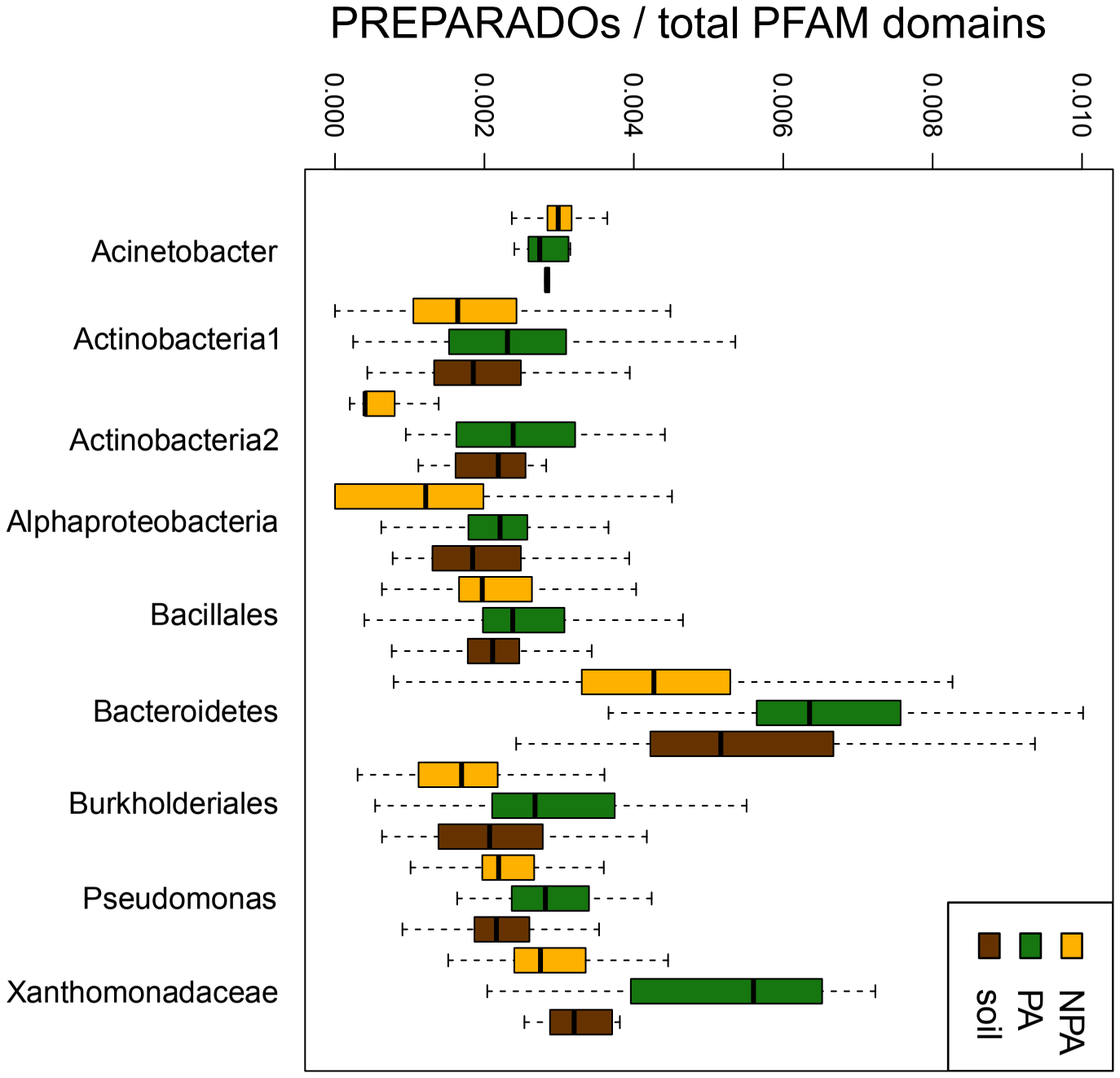
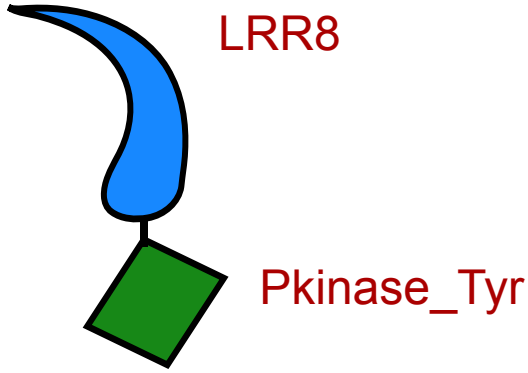
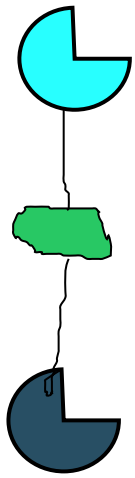

PREPARADOs / total PFAM domains
0.000
0.002
0.004
0.006
0.008
0.010
Acinetobacter
Actinobacteria1
Actinobacteria2
Alphaproteobacteria
Bacillales
Bacteroidetes
Burkholderiales
Pseudomonas
Xanthomonadaceae
NPA
PA
soil

Figure S22. PREPARADO abundance as a fraction of total Pfam domains across the nine taxa.

a.



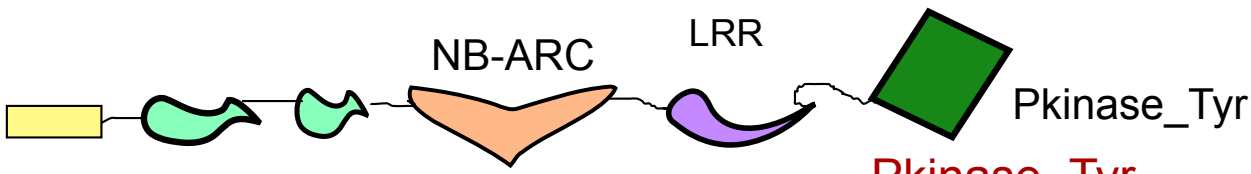
PSPTOimg_00032580
(IMG gene 2508869181)
Pseudomonas syringae pv. tomato DC3000



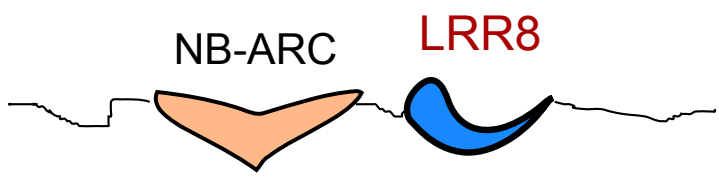
PREPARADOs

Aave_0386
(IMG gene ID 639821592)
Acidovorax avenae citrulli AAC00-1

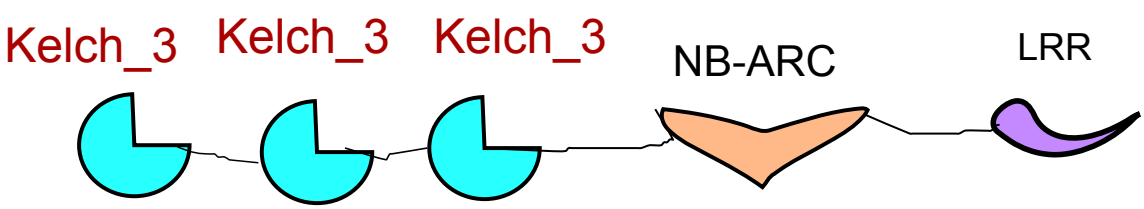
b.



WRKY19
Arabidopsis thaliana



SUMM2
Arabidopsis thaliana



XP_018514354
Probable disease resistance RPP8-like protein 2
Brassica rapa

Figure S23. An illustration of PREPARADOs contained in putative effector binding or disease resistance proteins in plants. a. Examples of microbial proteins, each with two PREPARADOs (LRR8, Pkinase_Tyr, Kelch_3, Kelch_4). **b.** Integration of PREPARADOs into NB-ARC domains in different plant proteins. NB-ARC is present in many disease resistance (R) proteins. SUMM2 was suggested to act as an R gene[34].

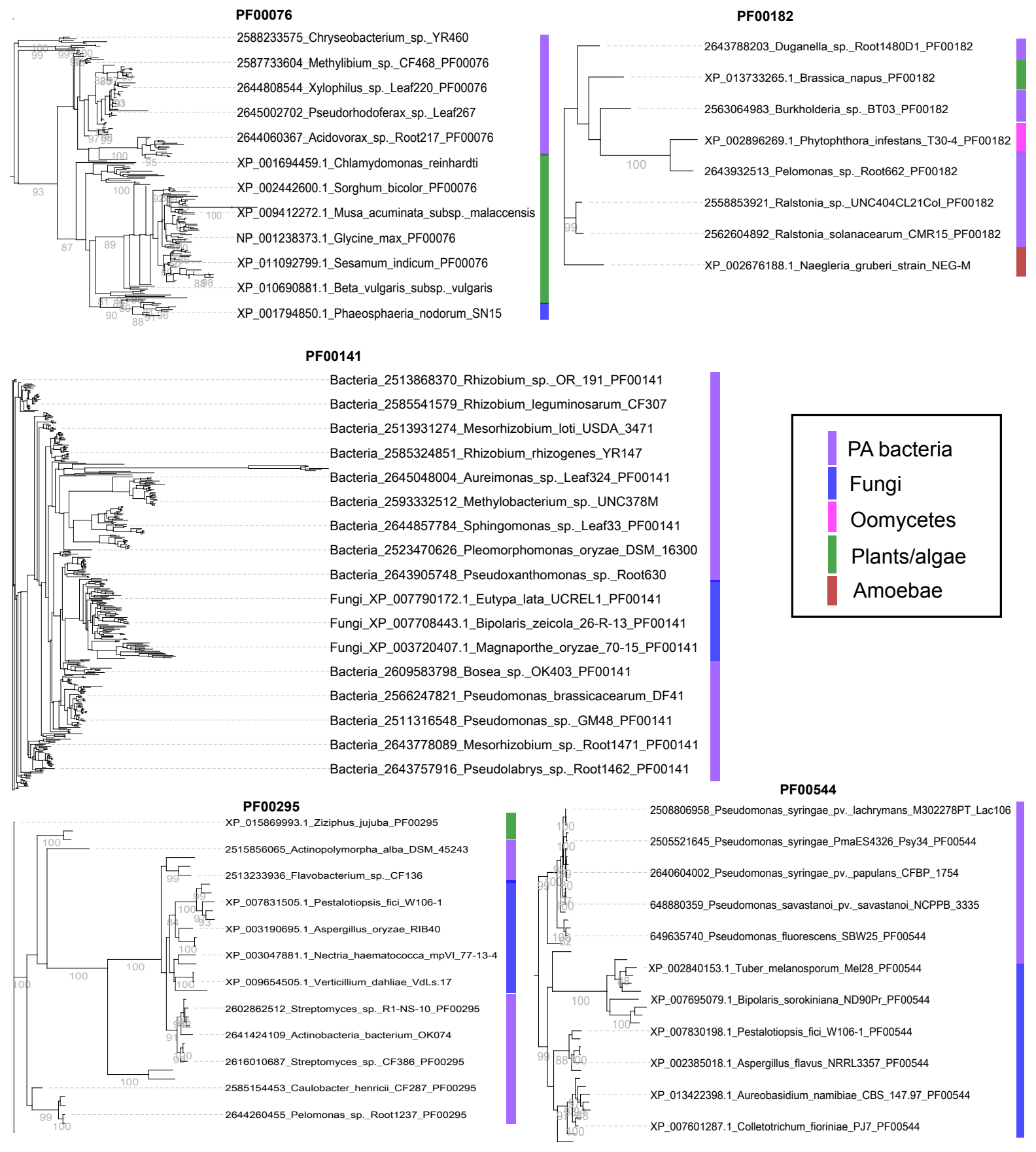

PF00076
2588233575_Chryseobacterium_sp._YR460
2587733604_Methylibium_sp._CF468_PF00076
2644808544_Xylophilus_sp._Leaf220_PF00076
2645002702_Pseudorhodoferax_sp._Leaf267
2644060367_Acidovorax_sp._Root217_PF00076
XP_001694459.1_Chlamydomonas_reinhardti
XP_002442600.1_Sorghum_bicolor_PF00076
XP_009412272.1_Musa_acuminata_subsp._malaccensis
NP_001238373.1_Glycine_max_PF00076
XP_011092799.1_Sesamum_indicum_PF00076
XP_010690881.1_Beta_vulgaris_subsp._vulgaris
XP_001794850.1_Phaeosphaeria_nodorum_SN15
PF00182
2643788203_Duganella_sp._Root1480D1_PF00182
XP_013733265.1_Brassica_napus_PF00182
2563064983_Burkholderia_sp._BT03_PF00182
XP_002896269.1_Phytophthora_infestans_T30-4_PF00182
2643932513_Pelomonas_sp._Root662_PF00182
2558853921_Ralstonia_sp._UNC404CL21Col_PF00182
2562604892_Ralstonia_solanacearum_CMR15_PF00182
XP_002676188.1_Naegleria_gruberi_strain_NEG-M
PF00141
Bacteria_2513868370_Rhizobium_sp._OR_191_PF00141
Bacteria_2585541579_Rhizobium_leguminosarum_CF307
Bacteria_2513931274_Mesorhizobium_loti_USDA_3471
Bacteria_2585324851_Rhizobium_rhizogenes_YR147
Bacteria_2645048004_Aureimonas_sp._Leaf324_PF00141
Bacteria_2593332512_Methylobacterium_sp._UNC378M
Bacteria_2644857784_Sphingomonas_sp._Leaf33_PF00141
Bacteria_2523470626_Pleomorphomonas_oryzae_DSM_16300
Bacteria_2643905748_Pseudoxanthomonas_sp._Root630
Fungi_XP_007790172.1_Eutypa_lata_UCREL1_PF00141
Fungi_XP_007708443.1_Bipolaris_zeicola_26-R-13_PF00141
Fungi_XP_003720407.1_Magnaporthe_oryzae_70-15_PF00141
Bacteria_2609583798_Bosea_sp._OK403_PF00141
Bacteria_2566247821_Pseudomonas_brassicacearum_DF41
Bacteria_2511316548_Pseudomonas_sp._GM48_PF00141
Bacteria_2643778089_Mesorhizobium_sp._Root1471_PF00141
Bacteria_2643757916_Pseudolabrys_sp._Root1462_PF00141
PA bacteria
Fungi
Oomycetes
Plants/algae
Amoebae
PF00295
XP_015869993.1_Ziziphus_jujuba_PF00295
2515856065_Actinopolymorpha_alba_DSM_45243
2513233936_Flavobacterium_sp._CF136
XP_007831505.1_Pestalotiopsis_fici_W106-1
XP_003190695.1_Aspergillus_oryzae_RIB40
XP_003047881.1_Nectria_haematococca_mpVI_77-13-4
XP_009654505.1_Verticillium_dahliae_VdLs.17
2602862512_Streptomyces_sp._R1-NS-10_PF00295
2641424109_Actinobacteria_bacterium_OK074
2616010687_Streptomyces_sp._CF386_PF00295
2585154453_Caulobacter_henricii_CF287_PF00295
2644260455_Pelomonas_sp._Root1237_PF00295
PF00544
2508806958_Pseudomonas_syringae_pv._lachrymans_M302278PT_Lac106
2505521645_Pseudomonas_syringae_PmaES4326_Psy34_PF00544
2640604002_Pseudomonas_syringae_pv._papulans_CFBP_1754
648880359_Pseudomonas_savastanoi_pv._savastanoi_NCPPB_3335
649635740_Pseudomonas_fluorescens_SBW25_PF00544
XP_002840153.1_Tuber_melanosporum_Mel28_PF00544
XP_007695079.1_Bipolaris_sorokiniana_ND90Pr_PF00544
XP_007830198.1_Pestalotiopsis_fici_W106-1_PF00544
XP_002385018.1_Aspergillus_flavus_NRRL3357_PF00544
XP_013422398.1_Aureobasidium_namibiae_CBS_147.97_PF00544
XP_007601287.1_Colletotrichum_fioriniae_PJ7_PF00544

Figure S24. Maximum-likelihood phylogenetic trees of a few PREPARADO-containing proteins demonstrating high similarity between those found in PA bacteria, fungi, oomycetes, and plants. Only a small fraction of the proteins in the tree are presented due to size limitation. In each label the long integer represents an IMG gene ID. Accession starting with XP_ are Refseq proteins.

a

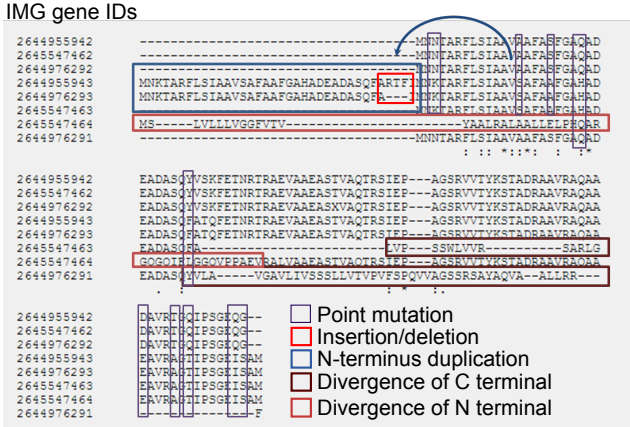
Alignment fraction

	Leaf191	Leaf76	Leaf84	Leaf78
Leaf191		0.99	0.97	0.87
Leaf76	1		0.97	0.87
Leaf84	0.96	0.96		0.86
Leaf78	0.87	0.87	0.86	

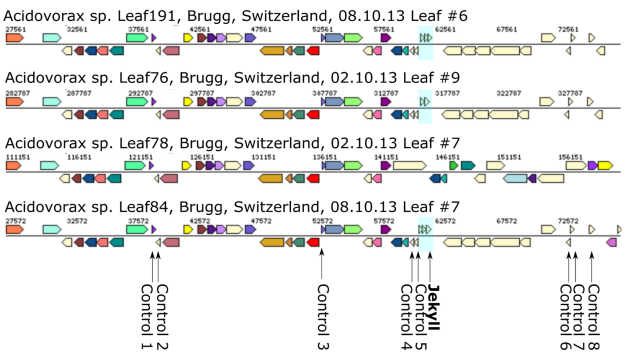
Average Nucleotide Identity

	Leaf191	Leaf76	Leaf84	Leaf78
Leaf191		100	99.9	94.76
Leaf76	100		99.9	94.77
Leaf84	99.9	99.9		94.77
Leaf78	94.76	94.77	94.77	

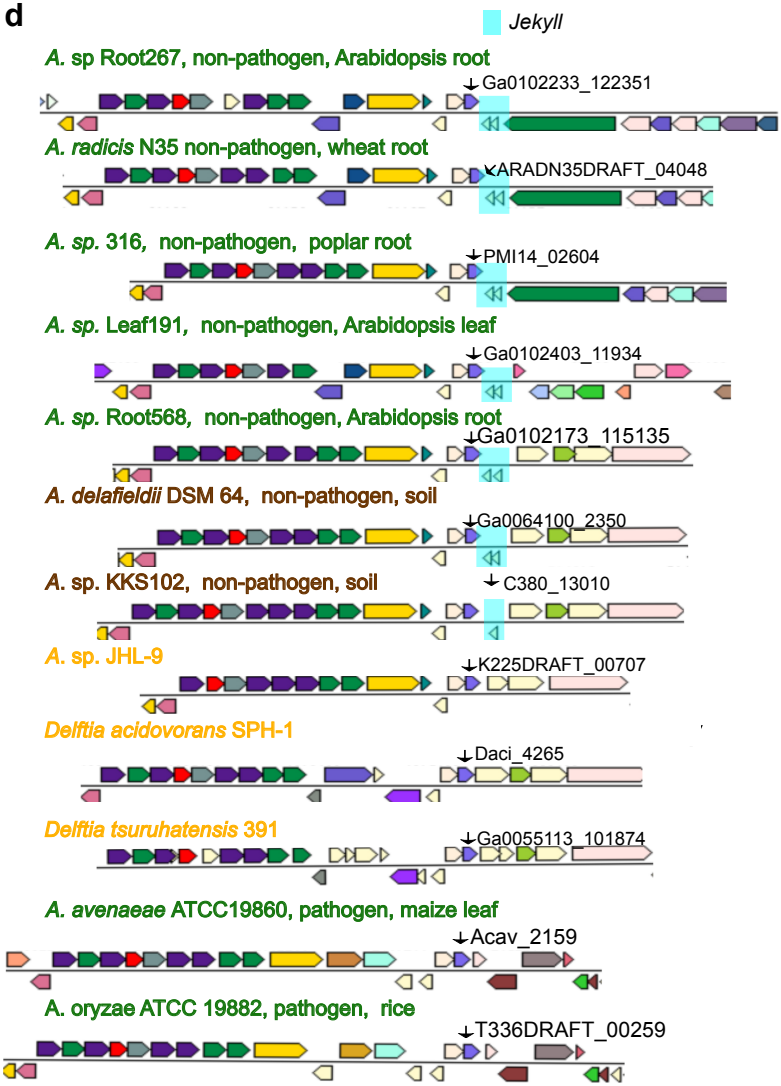
b



c



d



e

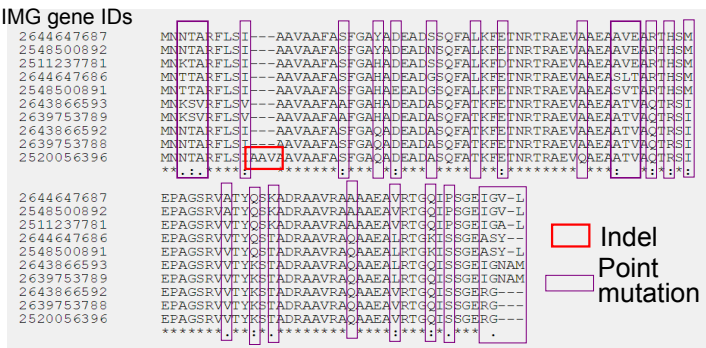


Figure S25. *Jekyll* gene variability and neighbors. a. Comparison of the genome similarity between four *Acidovorax* isolates from naturally grown *Arabidopsis* leaves in Switzerland[35]. **b.** Multiple sequence alignments using online MAFFT of the proteins shown in Figure 6b. **c.** Comparison of evolutionary changes between *Jekyll* genes versus short control genes in the same genomic neighborhood is presented in Table S24. **d.** A *Jekyll* locus within a conserved genomic region showing specific presence in non-pathogenic PA *Acidovorax* isolated from different plants (upper green labels) and soil-associated (brown labels) *Acidovorax*. Below are NPA *Acidovorax* and *Delftia* genomes (orange labels) and pathogenic PA *Acidovoarx* (two last green labels) **e.** Multiple sequence alignment using MAFFT of proteins marked in **d**.

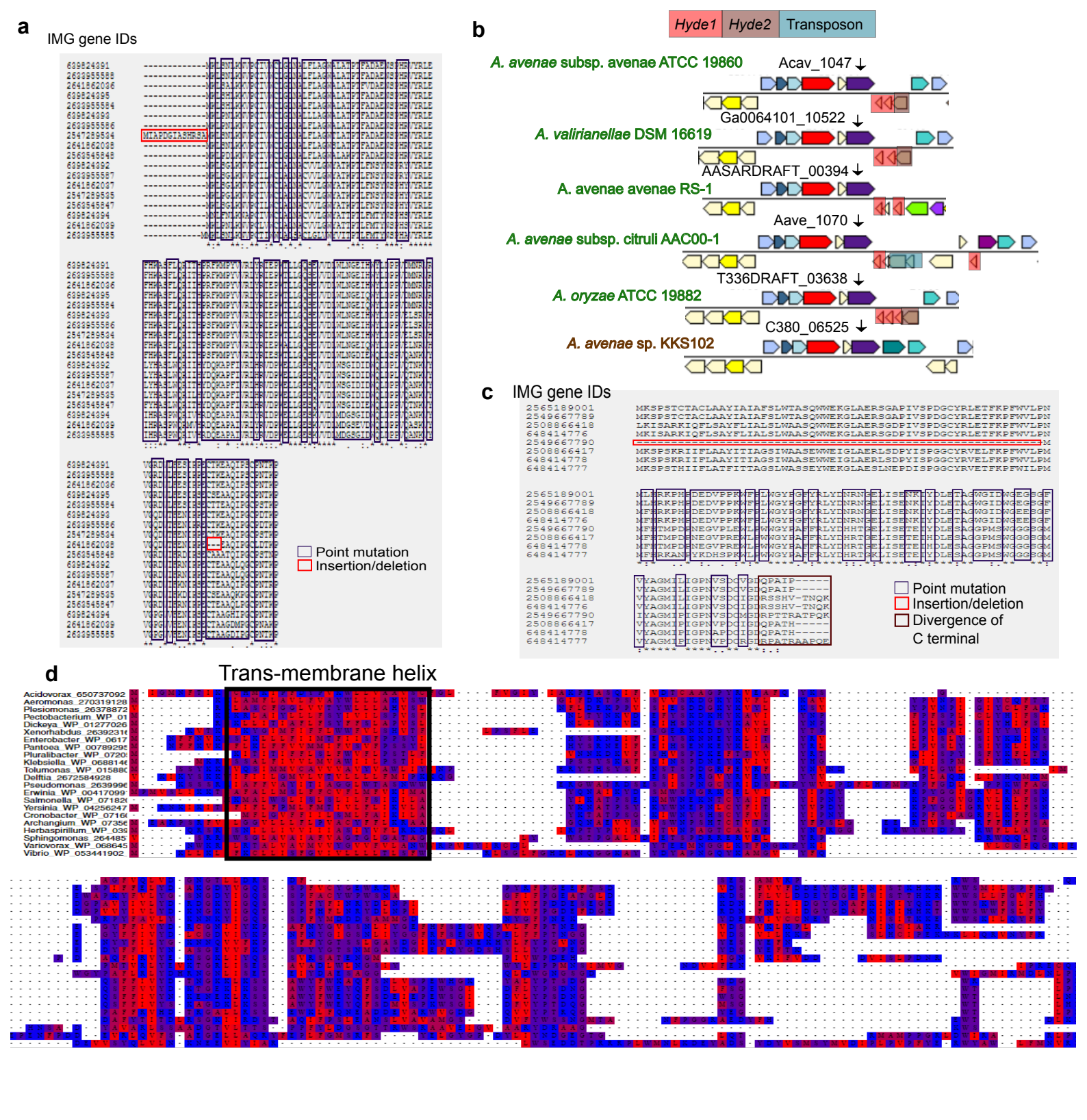

a
IMG gene IDs
Point mutation
Insertion/deletion
b
Hyde1
Hyde2
Transposon
A. avenae subsp. avenae ATCC 19860
Acav_1047
Ga0064101_10522
A. valirianellae DSM 16619
AASARDRAFT_00394
A. avenae avenae RS-1
Aave_1070
A. avenae subsp. citruli AAC00-1
T336DRAFT_03638
A. oryzae ATCC 19882
C380_06525
A. avenae sp. KKS102
c
IMG gene IDs
Point mutation
Insertion/deletion
Divergence of
C terminal
d
Trans-membrane helix

Figure S26. *Hyde* genes variability and protein motifs. a. Multiple sequence alignment by MAFFT of the Hyde1 proteins presented in Figure 6c. **b.** A variable Hyde locus. Note the absence of the locus from the last soil-associated isolate despite the conservation of the genomic environment. **c.** Multiple sequence alignment by MAFFT of the Hyde1 proteins presented in **b**. **d.** Similarity between Hyde and other Hyde1-like proteins of different Proteobacteria. From top to bottom, proteins in the following genera: *Acidovorax*, *Aeromonas*, *Plesiomonas*, *Dickeya*, *Xenorhabdus*, *Enterobacter*, *Pantoea*, *Pluralibacter*, *Klebsiella*, *Tolumonas*, *Delftia*, *Pseudomonas*, *Erwinia*, *Salmonella*, *Yersinia*, *Cronobacter*, *Archangium*, *Herbaspirillium*, *Sphingomonas*, *Variovorax*, *Vibrio*. Red color denotes hydrophobic amino acids. A transmembrane helix is predicted in the N terminus. The rest of the protein shows no conservation**.**

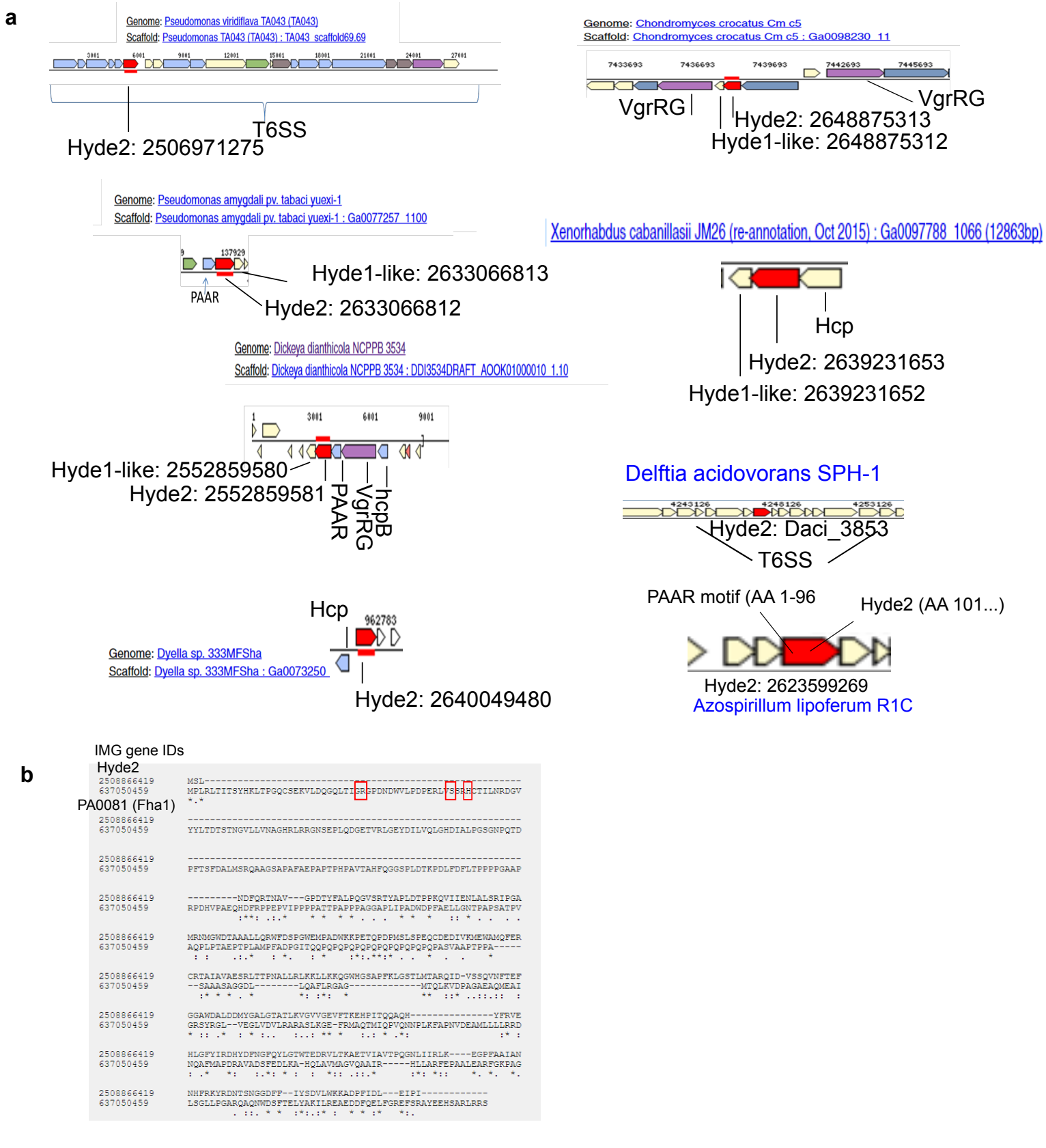

a
Genome: Pseudomonas viridiflava TA043 (TA043)
Scaffold: Pseudomonas TA043 (TA043) : TA043_scaffold69.69
T6SS
Hyde2: 2506971275
Genome: Chondromyces crocatus Cm c5
Scaffold: Chondromyces crocatus Cm c5 : Ga0098230_11
VgrRG
VgrRG
Hyde2: 2648875313
Hyde1-like: 2648875312
Genome: Pseudomonas amygdali pv. tabaci yuexi-1
Scaffold: Pseudomonas amygdali pv. tabaci yuexi-1 : Ga0077257_1100
PAAR
Hyde1-like: 2633066813
Hyde2: 2633066812
Xenorhabdus cabanillasii JM26 (re-annotation, Oct 2015) : Ga0097788_1066 (12863bp)
Hcp
Hyde2: 2639231653
Hyde1-like: 2639231652
Genome: Dickeya dianthicola NCPPB 3534
Scaffold: Dickeya dianthicola NCPPB 3534 : DDI3534DRAFT_AOOK01000010_1.10
Hyde1-like: 2552859580
Hyde2: 2552859581
PAAR
VgrRG
hcpB
Delftia acidovorans SPH-1
Hyde2: Daci_3853
T6SS
PAAR motif (AA 1-96
Hyde2 (AA 101...)
Hyde2: 2623599269
Azospirillum lipoferum R1C
Hcp
Genome: Dyella sp. 333MFSha
Scaffold: Dyella sp. 333MFSha : Ga0073250
Hyde2: 2640049480
b
IMG gene IDs
Hyde2
PA0081 (Fha1)

Figure S27. Association between *Hyde* loci and T6SS. a. Genomic proximity between different Hyde2 proteins (marked in red, number represent IMG gene number) and different T6SS components and a fusion event between Hyde2 and PAAR domain in *Azospirilium*. AA – amino acids. **b.** Similarity between Hyde2 protein of *Pseudomonas syringae* pv. tomato DC3000 (DC3000 gold standard) and FHA1 protein - a core scaffolding protein of the *P. aeruginosa* H-T6SS that is required for protein secretion by T6SS[36]. The amino acids marked in red are phosphopeptide binding motif. Hyde2 is shorter than the FHA protein and lacks the FHA domain (pfam00498).

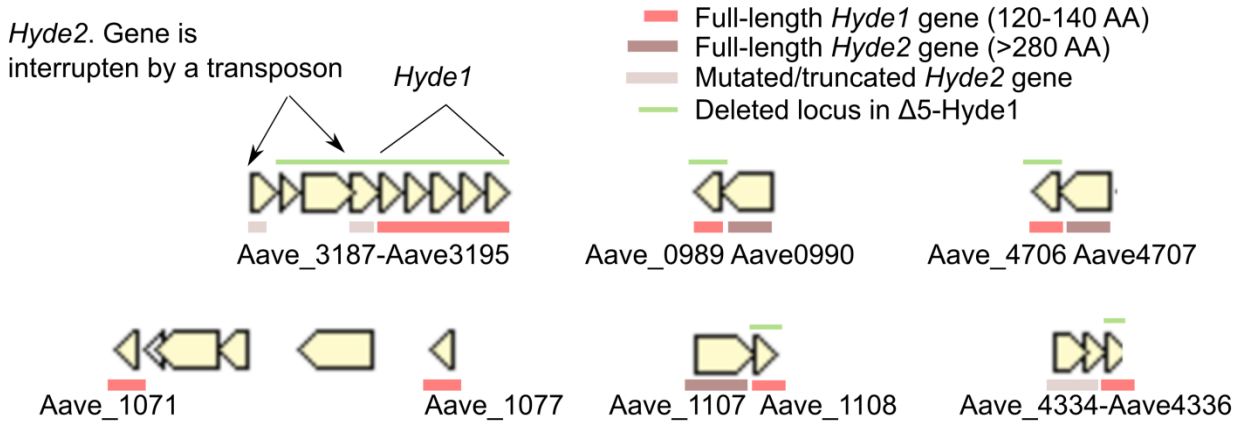
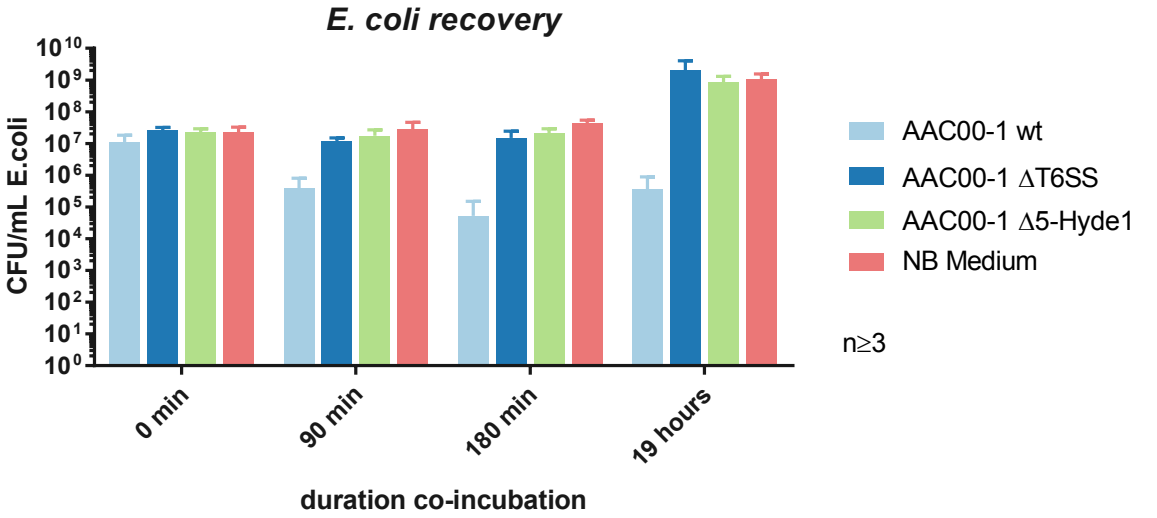
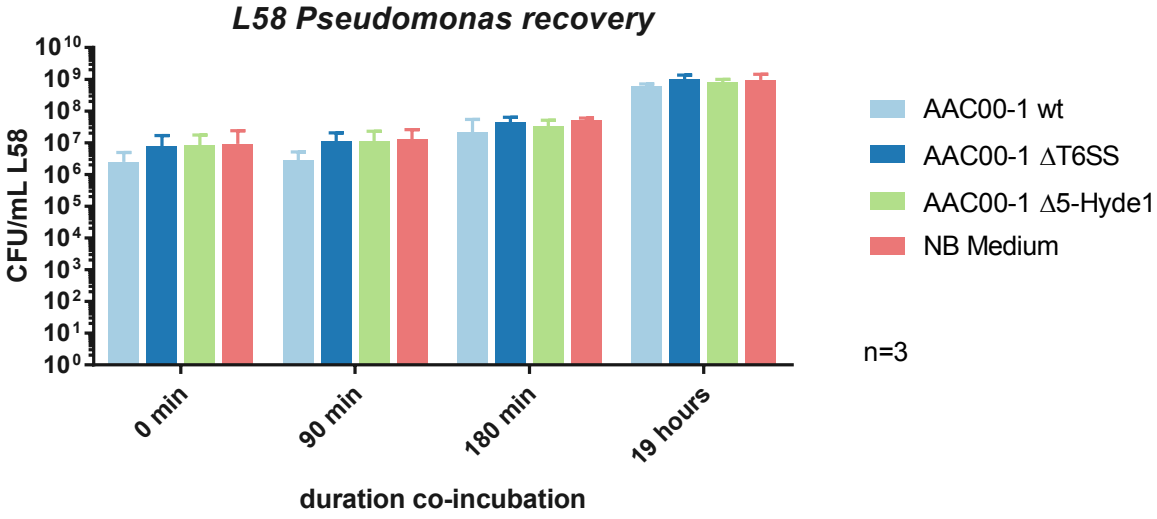


Figure S28. *Hyde* loci in *Acidovorax citruli* AAC00-1.

a


E. coli recovery
CFU/mL E.coli
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
10^{4}
10^{3}
10^{2}
10^{1}
10^{0}
0 min
90 min
180 min
19 hours
duration co-incubation
AAC00-1 wt
AAC00-1 ΔT6SS
AAC00-1 Δ5-Hyde1
NB Medium
n≥3

b


L58 Pseudomonas recovery
CFU/mL L58
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
10^{4}
10^{3}
10^{2}
10^{1}
10^{0}
0 min
90 min
180 min
19 hours
duration co-incubation
AAC00-1 wt
AAC00-1 ΔT6SS
AAC00-1 Δ5-Hyde1
NB Medium
n=3

c

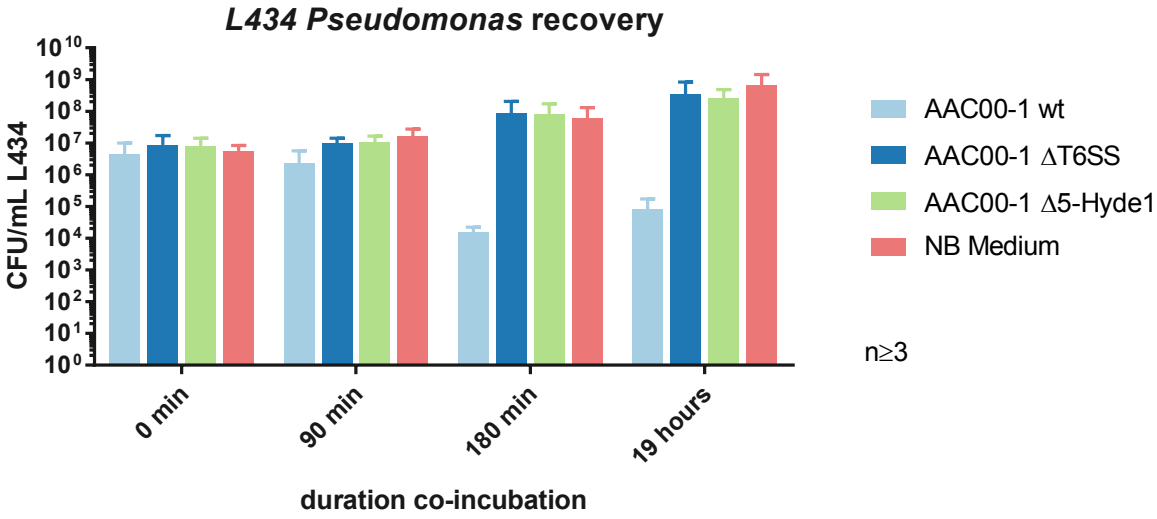

L434 Pseudomonas recovery
CFU/mL L434
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
10^{4}
10^{3}
10^{2}
10^{1}
10^{0}
0 min
90 min
180 min
19 hours
duration co-incubation
AAC00-1 wt
AAC00-1 ΔT6SS
AAC00-1 Δ5-Hyde1
NB Medium
n≥3

Figure S29. Recovered prey cells after co-incubation with *Acidovorax* aggresor strains. Chloramphenicol-resistant prey cells *E. coli* BW25113 (**a**), *L58 Pseudomonas* (**b**), and *L434 Pseudomonas* (**c**) were mixed at equal ratio with different *Acidovorax* strains or NB medium. After co-incubation of the indicated times on NB agar plates at 28C, mixed populations were resuspended in NB medium and spotted on Chloramphenicol-containing NB agar. Means with SD are shown (n≥3).