Community proteogenomics reveals insights into the physiology of phyllosphere bacteria

Nathanaël Delmotte[a,1], Claudia Knief[a,1], Samuel Chaffron[b], Gerd Innerebner[a], Bernd Roschitzki[c], Ralph Schlapbach[c], Christian von Mering[b], and Julia A. Vorholt[a,2]

[a]Institute of Microbiology, Eidgenössische Technische Hochschule Zurich, Wolfgang-Pauli-Strasse 10, 8093 Zurich, Switzerland; [b]Institute of Molecular Biology and Swiss Institute of Bioinformatics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland; and [c]Functional Genomics Center Zurich, University of Zurich/Eidgenössische Technische Hochschule Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

Edited by Steven E. Lindow, University of California, Berkeley, CA, and approved July 16, 2009 (received for review May 12, 2009)

Aerial plant surfaces represent the largest biological interface on Earth and provide essential services as sites of carbon dioxide fixation, molecular oxygen release, and primary biomass production. Rather than existing as axenic organisms, plants are colonized by microorganisms that affect both their health and growth. To gain insight into the physiology of phyllosphere bacteria under in situ conditions, we performed a culture-independent analysis of the microbiota associated with leaves of soybean, clover, and *Arabidopsis thaliana* plants using a metaproteogenomic approach. We found a high consistency of the communities on the 3 different plant species, both with respect to the predominant community members (including the alphaproteobacterial genera *Sphingomonas* and *Methylobacterium*) and with respect to their proteomes. Observed known proteins of *Methylobacterium* were to a large extent related to the ability of these bacteria to use methanol as a source of carbon and energy. A remarkably high expression of various TonB-dependent receptors was observed for *Sphingomonas*. Because these outer membrane proteins are involved in transport processes of various carbohydrates, a particularly large substrate utilization pattern for Sphingomonads can be assumed to occur in the phyllosphere. These adaptations at the genus level can be expected to contribute to the success and coexistence of these 2 taxa on plant leaves. We anticipate that our results will form the basis for the identification of unique traits of phyllosphere bacteria, and for uncovering previously unrecorded mechanisms of bacteria-plant and bacteria-bacteria relationships.

metaproteomics | methylotrophy | plant phyllosphere | *Pseudomonas* | *Sphingomonas*

For terrestrial plants, the phyllosphere represents the interface between the above-ground parts of plants and the air. Conservative estimates indicate that the roughly 1 billion square kilometers of worldwide leaf surfaces host more than 10^{26} bacteria, which are the most abundant colonizers of this habitat (1, 2). The overall microbiota in this ecosystem is thus sufficiently large to have an impact on the global carbon and nitrogen cycles. Additionally, the phyllosphere inhabitants influence their hosts at the level of the individual plants. To a large extent, interest in phyllosphere microbiology has been driven by investigations on plant pathogens. Their spread, colonization, survival, and pathogenicity mechanisms have been the subject of numerous studies (2). Much less understood are nonpathogenic microorganisms that inhabit the phyllosphere. The composition of the phyllosphere microbiota has been analyzed in only a few studies by cultivation-independent methods (e.g., refs. 3–5); however, such methods are essential in light of the yet uncultivated majority of bacteria existing in nature (6), or more specifically on plant leaves (7). Not only their identity, but in particular the physiological properties of phyllosphere bacteria, their adaptations to the habitat, and their potential role (e.g., with respect to modulating population sizes of pathogens) remain largely unknown. Current knowledge on the traits important in the phyllosphere is derived from relatively few studies on gene expression and stems mostly from model bacteria cultivated on host plants under controlled conditions (8–11). However, under natural conditions, plants and their residing microorganisms are exposed to a host of diverse, highly variable environmental factors, including UV light, temperature, and water availability; moreover, individual microbes are subjected to competition with other microorganisms over resources, such as nutrients and space.

Toward a deeper understanding of phyllosphere microbiology, and in particular to learn more about the commensal majority of plant leaf colonizing bacteria, which may be of relevance for plant health and development, integrated approaches are needed. Here, we combined metagenomic and metaproteomic approaches (community proteogenomics) (12) to analyze bacterial phyllosphere communities in situ (the phyllosphere is defined here as the environment comprising both the surface and the apoplast of leaves). We studied 3 different plant species grown under standard agriculture regimes or under natural conditions. Our results provide insight into the physiology of bacteria and point toward common adaptation mechanisms among the phyllosphere populations of different plants.

Results and Discussion

The prokaryotic phyllosphere populations in our study were obtained from 2 field-grown plant species, soybean (*Glycine max*, 2 samples) and clover (*Trifolium repens*, 3 samples), as well as from a wild population of the model plant *Arabidopsis thaliana* (1 sample) (Fig. 1, Table S1). Genomic DNA and proteins of the prokaryotes were extracted from the same pools of cells. For 1 of the 6 samples, Soybean 2, 260 Mbp of metagenomic sequence reads were generated using 454 pyrosequencing technology.

Microbial Community Composition. To characterize the composition of the phyllosphere microbiota, we applied complementary approaches: phylogenetic information was derived from protein-coding marker genes in the metagenome database generated in this study, as well as from 16S rRNA gene-based clone libraries. Comparative community analyses were additionally done by denaturing gradient gel electrophoresis (DGGE) to evaluate the representativeness of the samples.

Author contributions: N.D., C.K., and J.A.V. designed research; N.D., C.K., G.I., and J.A.V. performed research; S.C., B.R., R.S., and C.v.M. contributed new reagents/analytic tools; N.D., C.K., S.C., G.I., C.v.M., and J.A.V. analyzed data; and N.D., C.K., C.v.M., and J.A.V. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: Protein databases as well as lists of identified proteins in form of Scaffold and Excel files are available at http://www.micro.biol.ethz.ch/research/vorholt. MS/MS data have been deposited in the PRIDE database (9850–9860) and gene sequences in the GenBank database [accession nos. 38721 (Metagenome), and FN421480 to FN421999 (16S rRNA)].

[1]N.D. and C.K. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: vorholt@micro.biol.ethz.ch.

This article contains supporting information online at www.pnas.org/cgi/content/full/0905240106/DCSupplemental.

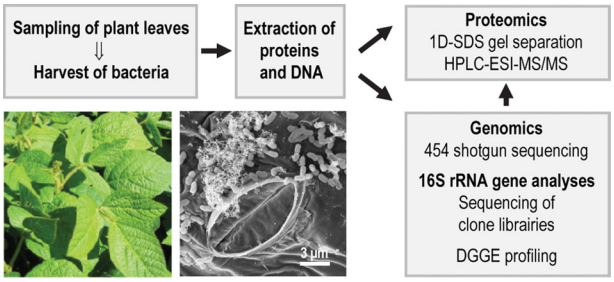


Fig. 1. Experimental strategy applied to characterize the phyllosphere microbiota. All analyses described were conducted from identical pools of cells as starting material. The photograph shows leaves of soybean plants; the electron micrograph shows the surface of an *Arabidopsis* leaf.

In a first step, the phylogenetic information contained in selected protein-coding marker genes of the metagenome data were used to analyze the composition of the microbial phyllosphere community in the Soybean 2 sample (Fig. 2). This approach gives a quantitative overview without the introduction of a PCR primer bias (13). Overall, we observed a clear dominance of Alphaproteobacteria. A relevant fraction of this group is well known to have adopted an extra- or intracellular lifestyle as plant mutualists or as plant or animal pathogens. The majority of Alphaproteobacteria in the Soybean 2 sample belonged to the families of Sphingomonadaceae (*Sphingomonas* 20.1%, *Novosphingobium* 10.1%) and Methylobacteriaceae (*Methylobacterium* 20.2%), which have been previously detected on plants (see, for example, refs. 14–16). Bacteria of the genus *Methylobacterium* and *Sphingomonas* were also detected in the Soybean 2 sample by 16S rRNA gene-based community anal-

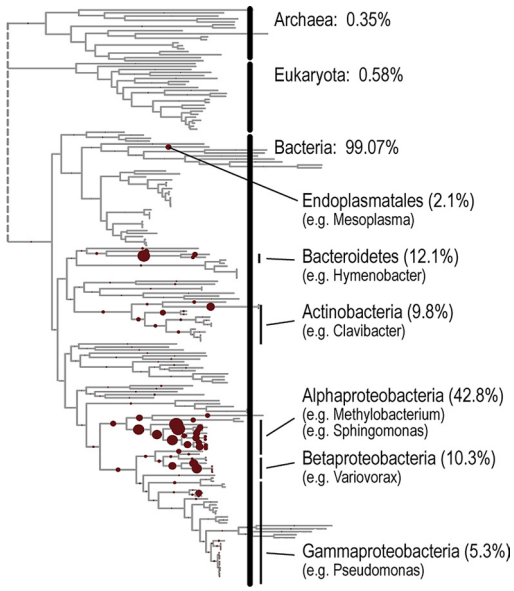


Fig. 2. Taxonomic composition of the bacterial community in the Soybean 2 sample. A phylogenetic tree calculated from informative marker genes of completely sequenced organisms serves as a reference onto which the estimated coverage of the most abundant clades present in the Soybean 2 sample is projected. Coverage is estimated based on the quantity of marker genes found in the metagenome data and is indicated by red dots (13). A selection of typical representatives of the clades is listed to the right, annotated according to the 16S rRNA gene-sequencing results (Table S2). Archaea contributed only 0.35% to the microbial community of the sample and were identified as members of the mesophilic Crenarchaeota (group 1.1b) by 16S rRNA gene sequencing. The low contribution of eukaryotes (0.58%) to the analyzed phyllosphere community in the soybean sample is in accordance with the design of the microbial harvesting procedure, which included a physical depletion step for eukaryotic cells.

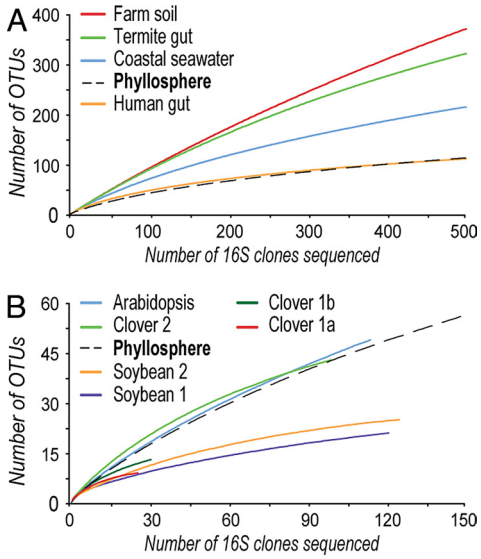


Fig. 3. Rarefaction analysis of 16S rRNA gene-sequence data to estimate microbial diversity based on a cutoff <97% sequence identity for delineation of operational taxonomic units (OTUs). (*A*) Comparison of the composite phyllosphere dataset of this study with published samples covering at least 500 sequences each: farm soil (20), termite gut (19), coastal seawater (17), human gut (18). (*B*) Rarefaction curves of the individual phyllosphere samples and the joint (composite) phyllosphere dataset.

yses as well as in the other 5 samples (Table S2). Further analysis of the clone libraries revealed that between 4% and 10% of the sequences represented unknown genera (see Table S2). Most of them were detected only sporadically, but unknown genera within the family of *Flexibacteraceae* were detected in nearly all samples. Several of the sequences that represented members of known genera were phylogenetically distinct to previously described representatives (type strains) and completely sequenced strains (Fig. S1).

Rarefaction analyses of 16S rRNA gene-sequence data from all 6 samples suggested that the bacterial diversity in the plant phyllosphere samples was lower than in soil, marine systems, or the gut of wood-feeding termites, and similar (Arabidopsis and the Clover 2 sample) or lower (Soybean, Clover 1a and b) than that of the human gut (17–20) (Fig. 3).

Based on cultivation-dependent methods, microbial communities in the phyllosphere have been described to be variable over time, in space, and across different plant species (21, 22). Therefore, DGGE analyses were performed to assess this variation in our field samples. Comparative analysis of the 6 samples showed that similar DGGE patterns were obtained for samples from the same plant species collected at different points in time, suggesting that the bacterial phyllosphere community remained rather stable over time (Fig. S2*a*). This finding was confirmed by the analysis of additional samples taken from the soybean field, which revealed that early colonizers were detectable throughout the whole growing season, while diversity increased during plant succession (Fig. S2*b*). The soybean plant leaves were colonized quite homogenously within the field, as was validated at the time points of harvest of sample material for community proteogenomic analysis (Fig. S2*c*). Taken together, the DGGE analyses showed a temporal and spatial stability of the phyllosphere communities, demonstrating the representativeness of the samples investigated in more detail in the proteome analyses described in the next section.

Comparative Metaproteome Analysis. Proteins from the microbiota of the 6 plant samples were identified after tryptic digestion, using high-accuracy MS. The proteins were processed as described in

Table 1. Identification of abundant proteins in phyllosphere bacteria

	Identifications with RefSeq	Identifications with RefSeq and metagenome	New identifications through metagenome	Gain [%]
Soybean 1	884	934	50	6
Soybean 2	561	1,047	486	87
Clover 1a	556	868	312	56
Clover 1b	442	767	325	74
Clover 2	411	548	137	33
Arabidopsis	505	751	246	49

Gain of protein identifications factored by combining the publicly available database with the generated metagenomic data.

Materials and Methods, and MS/MS spectra were searched against a database consisting of protein sequences obtained from the public RefSeq database with or without the translated metagenomic sequences mentioned above. In total, we identified 2,883 unique proteins with 12,345 peptides, originating from an extensive body of 487,304 spectra (see Table S3 for all identified bacterial proteins and Table S4 for proteins attributed to the respective host plants, soybean- or clover-mosaic viruses, as well as to fungi and oomycetes). The 2,257 bacterial proteins were considered for further interpretation, whereby protein abundance was roughly estimated by spectral counting (23).

The metagenome data significantly increased the number of identified proteins (Table 1), implying the presence of bacteria in our samples that are genetically distinct from those represented among currently sequenced genomes. As expected, the number of identifications increased most strongly for the Soybean 2 sample, from which the metagenome sequences were derived, leading to the identification of 486 additional proteins. Between 6% and 74% of new identifications were obtained for the other 5 samples (see Table 1), a finding that can be ascribed to similarities between bacterial taxa in Soybean 2 and the other samples. An overall consistency of the physiology of the microbiota present on the different plant species is evident at the level of gene expression (i.e., 75% of the proteins identified in the Soybean 2 sample were found in at least 1 of the other samples as well) (Fig. S3).

To assess the significance of similarities and differences in the proteomes and to identify shared and specifically enriched proteins with respect to the 3 different plant species, we examined the identified proteins according to their assignment to Pfam domains (24). This analysis revealed that more than 70% of all identified Pfam domains were present at roughly similar levels on the 3 different plant species (Fig. 4), confirming the overall consistency of the microbiota metaproteomes. Manual inspection of the significantly enriched Pfam-domains (E-value <0.01, P-value <0.0001) revealed that these could most likely be attributed to distinct stresses (as discussed below) or to the presence of distinct bacterial species on the various plant species (see Fig. 4).

Protein Identification in Relation to Bacterial Genera. Most identified proteins were assigned to the 3 bacterial genera *Methylobacterium*, *Sphingomonas*, and *Pseudomonas*, which profited to a different degree from metagenomic information (Table 2, and see Table S3): whereas half of the 20 most abundant proteins of *Methylobacterium* were identifiable through RefSeq and half through the metagenome database, all of the abundant proteins assigned to *Sphingomonas* were identified in various samples based on data we obtained by metagenome sequencing (see Table 2). This suggests that a certain part of the *Methylobacterium* population in the phyllosphere samples is genetically close to the completely sequenced *Methylobacterium* strains currently available in public databases (6 strains), while a major part of the *Sphingomonas* population is different from the sequenced strains (2 strains). These

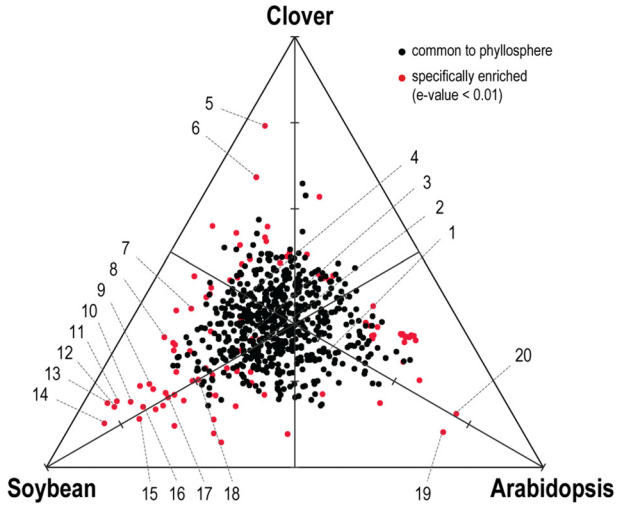


Fig. 4. Conserved and specifically enriched proteome functions (spectral counting of Pfam domains) per host-plant type. Pfam domains drawn close to a vertex are preferentially and specifically found on that respective plant. Selected examples are highlighted and discussed in the text. Examples of common phyllosphere proteome (i.e., not enriched): 1, PF00120, glutamine synthetase catalytic domain; 2, PF02469, fasciclin domain; 3, PF00593, TonB-dependent receptor; 4, PF07715, TonB-dependent receptor plug domain. Specific proteome enrichments: 5, PF00027, cyclic nucleotide-binding domain; 6, PF03328, HpcH/HpaI aldolase/citrate lyase family; 7, PF00210, ferritin-like domain (e.g., bacterioferritins); 8, PF05067, manganese containing catalase; 9, PF06823, protein of unknown function (DUF1236); 10, PF00669, bacterial flagellin N terminus; 11, PF00128, α-amylase, catalytic domain; 12, PF03413, peptidase propeptide and YPEB domain; 13, PF05443, ROS/MUCR transcriptional regulator protein; 14, PF05532, CsbD-like (general stress response); 15, PF00011, Hsp20/alpha crystallin family; 16, PF02566, OsmC-like (e.g., organic hydroperoxide detoxification); 17, PF00700, bacterial flagellin N terminus; 18, PF01584, CheW-like (chemotaxis signaling); 19, PF00532, periplasmic binding and sugar binding domain; 20, PF00502, phycobilisome protein (light harvesting).

conclusions are in agreement with our phylogenetic analysis (see above and Fig. S1). On the other extreme, we observed that all 20 dominant proteins of *Pseudomonas* spp. were identifiable based on RefSeq sequences (see Table 2). This latter observation is also in accordance with our data from 16S rRNA gene-clone library analyses, which showed a very close phylogenetic relationship of the phyllosphere-inhabiting *Pseudomonas* strains to sequenced strains (see Fig. S1*c*). In total, 77 proteins were identified on the basis of metagenome information that did not reveal significant sequence identity to any known or predicted protein (see Table S3). It can, however, not be excluded that some of these are of eukaryotic or viral origin. Notably, 8 of these proteins were found to be expressed in multiple samples among the most abundant proteins (see Table S5). These proteins are of particular interest for further characterization; however, this will most likely require assignment to their respective organisms first.

Plant-Associated Lifestyle

Transport-Related Proteins. Bacterial communities in the phyllosphere are thought to be limited by carbon availability, and it may be expected that access to carbon compounds on leaves is a major determinant of epiphytic colonization (2). There is evidence that small amounts of nutrients, such as simple sugars including glucose, fructose, and sucrose, leach from the interior of the plant (2). We specifically analyzed transport-related functions among the identified proteins to obtain indications for the type of substrates consumed by the phyllosphere microbiota. The most prominent group of transport proteins in our samples consisted of outer-membrane β-barrel proteins (i.e., porins and TonB receptors), which were consistently detected in the analyzed samples from the

Table 2. Most abundant proteins detected in *Methylobacterium*, *Sphingomonas*, and *Pseudomonas*, respectively

Protein (DB)	SY1	SY2	CL1a	CL1b	CL2	ARA
Methylobacterium						
Methanol DH-like XoxF (M)	n.d.	++	+++	+++	+++	++
Fae (M,R)	+++	++	++	++	+++	+
MucR (M)	+	+++	+++	+++	++	+
GroEL (R)	+	++	++	+++	+++	++
Hypothetical protein (R)	++	++	++	+++	++	n.d.
Nucleoside-diP kinase (M)	+	++	++	+++	++	+
Methanol DH MxaF (M,R)	+	+++	++	+++	+	n.d.
Beta-Ig-H3/fasciclin (R)	+++	+++	+	++	+	+
Cold-shock protein (M)	+	++	++	+++	++	+
Beta-Ig-H3/fasciclin (M)	++	+++	+	++	+	+
60 kDa chaperonin (M)	+	+	++	+++	++	n.d.
Phasin (R)	+++	+++	+	+	+	+
Superoxide dismutase (M,R)	+	++	++	++	++	+
Cold-shock protein (M,R)	+	++	++	+	++	+
Chaperonin Cpn10 (R)	++	+	++	++	++	+
Malyl-CoA lyase Mcl (R)	+	+	+	+++	++	+
ClpP (M)	+	+	+++	++	+	+
Surface antigen (M)	n.d.	+	+	++	+++	+
SWIB/MDM2 protein (M)	n.d.	+	++	+	++	n.d.
Invasion associated (M)	n.d.	+	++	++	++	n.d.
Sphingomonas						
OmpA/MotB (M)	+	++	++	+	+	+++
Succinyl-CoA ligase, α (M)	++	+	++	+	+	+++
EF-Tu (M)	+	+	+	++	+	++
OmpA/MotB (M)	n.d.	+	++	++	++	++
EF-Tu (M)	+	+	+	++	+	++
MotA/TolQ/ExbB (M)	+	n.d.	+	+	+	+++
TonB-dependent receptor (M)	n.d.	+	+	++	+	+
GAP dehydrogenase (M)	+	+	+	+	+	+
Histone-like protein (M)	n.d.	+	+	+	+	+
OmpA/MotB (M)	+	++	+	+	n.d.	+
Glutamine synthetase (M)	+	+	+	+	+	++
EF-G (M)	+	+	+	+	+	+
Uncharacterized protein (M)	n.d.	+	+	n.d.	n.d.	++
10 kDa chaperonin (M)	+	+	+	+	n.d.	+
Skp/OmpH (M)	+	+	+	+	n.d.	+
Uncharacterized protein (M)	+	+	n.d.	+	n.d.	+
Membrane protein (M)	+	n.d.	n.d.	+	n.d.	+
TonB-dependent receptor (M)	n.d.	n.d.	n.d.	+	n.d.	+
TonB-dependent receptor (M)	n.d.	n.d.	n.d.	+	+	+
TonB-dependent receptor (M)	+	n.d.	+	+	n.d.	+
Pseudomonas						
OprF (R)	+++	+++	+	n.d.	+	++
Single-stranded binding (R)	+++	++	+	n.d.	+	++
EF-Tu (R)	+++	+	+	n.d.	n.d.	+
Transcript. regulator (R)	+++	+	+	n.d.	n.d.	+
GroEL (R)	+++	+	+	+	+	+
DNA-binding protein (R)	++	+	+	n.d.	+	+
Unknown function DUF883 (R)	++	+	n.d.	n.d.	n.d.	+
Flagellin (R)	+++	+	n.d.	n.d.	n.d.	n.d.
OmpA (R)	++	+	n.d.	n.d.	n.d.	+
F0F1 ATP synthase, β (R)	+	+	+	+	+	+
Succinyl-CoA synth, β (R)	++	n.d.	+	n.d.	n.d.	+
Peptidoglycan lipoprotein (R)	+	+	+	n.d.	n.d.	++
Unknown function DUF883 (R)	++	+	n.d.	n.d.	n.d.	n.d.
Succinyl-CoA synth, α (R)	++	n.d.	+	n.d.	n.d.	n.d.
Chaperone Dank (R)	++	n.d.	+	n.d.	n.d.	n.d.
Glutamine synthetase (R)	++	+	+	n.d.	n.d.	+
Protein P-II (R)	+	n.d.	+	n.d.	n.d.	+
AphC (R)	+	+	+	n.d.	n.d.	+
F0F1 ATP synthase, α (R)	+	+	+	n.d.	+	+
Hsp20 (R)	++	+	n.d.	n.d.	n.d.	n.d.

Proteins were grouped if 90% identical over at least 40% of their length. Taxonomy (at the genus level) was inferred from the protein annotation. Ribosomal proteins are not reported here, but are listed in Table S3. Relative abundances are displayed with +, ++, and +++. DB, database. M (metagenome) and R (Refseq) indicate the database used for identification. n.d., not detected; SY1, Soybean 1; SY2, Soybean 2; CL1a, Clover 1a; CL1b, Clover 1b; CL2, Clover 2; ARA, *Arabidopsis*.

3 different plant hosts. Whereas the former enable passive diffusion of small molecules, the latter allow active transport of substrates greater than ≈600 Da. While we found porins to be abundantly present in various bacterial genera, including *Methylobacterium* and *Pseudomonas*, we observed an over-representation of TonB receptors and the respective plug domains among the proteins assigned to *Sphingomonas* (see Table S3 and Fig. 4). The high number and apparent divergence of the TonB systems is of particular interest, given the rapidly expanding variety of substrates known to be transported by these systems. Beyond the originally identified iron siderophore and vitamin B_{12} transport, the transport of an increasing number of carbohydrates has been reported (25). Our proteome data indicate expression of a gene for a TonB receptor in *Sphingomonas* (see Table S3, identifier Q1NFH3), which is located adjacent to a predicted sucrose hydrolase. Notably, these genes represent orthologs of XCC3358 and XCC3359. XCC3358 was recently described as one of 72 TonB-dependent receptors in the phytopathogen *Xanthomonas campestris* pv. *campestris* (Xcc) transporting sucrose with high affinity, and found to be required for full pathogenicity on *Arabidopsis* (26). Overall, the presence of multiple TonB transporters may account for the large abundance of *Sphingomonas* spp. in terms of abundance on plant leaves by scavenging various substrates present at low amounts, and may reflect a high degree of adaptiveness that can help explain the success of this alphaproteobacterial group.

We also found periplasmic compounds of ABC-transport systems for maltose, glucose, amino acids, and sucrose (see Table S3). Those proteins were more specifically observed to be expressed in *Pseudomonas*, indicating that *Pseudomonas* species could be specialized in mono- and disaccharide utilization and amino acid uptake. Remarkably, only few transporters were assigned to *Methylobacterium* spp.; these consisted mainly of ABC transporters for phosphate and sulfur compounds.

One-Carbon Metabolism. *Methylobacterium* is prominent for its methylotrophic metabolism, which allows it to use methanol, a side product of plant cell-wall metabolism, formed by pectin methyl esterases (27), as its carbon and energy source (28). The presence of this metabolic ability was suggested by numerous highly abundant proteins (see Table 2), including the large subunit of the periplasmic pyrrolo quinoline quinone-containing methanol dehydrogenase (MxaF) and a complete set of proteins of the tetrahydromethanopterin-dependent pathway (29). Moreover, proteins involved in the assimilation of methanol-derived methylene tetrahydrofolate and carbon dioxide via the serine pathway were detected, such as serine-glyoxylate aminotransferase, hydroxypyruvate reductase, and malyl-CoA lyase (30). These proteins are essential for methylotrophic growth and the encoding genes are located in a large genomic region (30), which is displayed in Fig. S4 together with identified peptides.

This genomic methylotrophy region also contains a gene for a methanol dehydrogenase-like protein (XoxF), which exhibits a sequence identity of 50% to MxaF. Under laboratory culture conditions, we were able to detect only very little of this protein in *Methylobacterium extorquens* cells and Bosch et al. (31) determined a 100-fold lower expression of *xoxF* compared to *mxaF* based on spectra counting of peptides. So far, no phenotype was observed for a *xoxF* mutant in *M. extorquens* AM1 (32) (for occurrence of *xoxF* and assumed functions in other bacteria see ref. 33). In contrast, upon plant colonization *xoxF* is highly expressed in *Methylobacterium* (see Table 2). For an approximation of expression levels, we integrated and correlated metagenomic and metaproteomic information using a 2-way fragment-recruitment approach, which revealed that the expression of *xoxF* was roughly in the same range as that for *mxaF* (Fig. S5). In the Arabidopsis sample, XoxF was even detected exclusively; that is, no MxaF was detectable. The high expression level of *xoxF* in *Methylobacterium* under environmental conditions suggests an important physiological role of XoxF during

MICROBIOLOGY

plant colonization. Further analyses of this protein, in particular with regard to substrate specificity and affinity, will be of great interest.

Overall, the detection of proteins known to be involved in methylotrophy and their assignment to *Methylobacterium* spp. suggests that facultative Methylobacteria are the dominating methylotrophs on plants, and that the large success of these bacteria in the phyllosphere can likely be attributed to specialization in carbon source utilization.

Nitrogen Metabolism. Bacteria can use various nitrogen sources, including ammonia, nitrate, dinitrogen, and a variety of amino acids and other nitrogenous organic compounds. The amino acid transporters mentioned above suggest that plant-derived nitrogen compounds are available for the bacteria. In addition, ammonia may be used as a nitrogen source, as suggested by the prominent presence of glutamine synthetase (see Fig. 4) in various bacteria, including *Sphingomonas*, *Methylobacterium*, and *Pseudomonas*. Indications for a dinitrogen fixation ability among the identified proteins of the phyllosphere microbiota inhabiting the studied plants were not found.

Stress Resistance. The phyllosphere is known as a hostile environment for the residing microorganisms (2, 9). In addition to the oligotrophic character of this habitat, physical parameters contribute to stressful conditions, such as UV radiation, temperature shifts, and the presence of reactive oxygen species. Adaptation to stressful conditions was reflected by the detection of various proteins, assigned to diverse bacterial genera and detected in all analyzed samples. Among these proteins were superoxide dismutase, catalase, DNA protection proteins, chaperones, and proteins involved in the formation of the osmoprotectant trehalose. Recently, evidence was presented that general stress response is an essential mechanism for plant colonization by *Methylobacterium* (9, 34). The regulatory system of general stress response in *Methylobacterium*, and presumably in other Alphaproteobacteria, consists of the 2-component response regulator PhyR that triggers upon activation regulation of stress-related protein functions via sigma factors of the EcfG family (35). PhyR and EcfG, respectively, were found among the detected proteins within this study (see Table S3) from members of the alphaproteobacterial genera *Methylobacterium*, *Sphingomonas*, and *Aurantimonas*, thus further emphasizing the importance of these regulatory proteins.

For *Pseudomonas*, besides the stress-response proteins, such as alkyl hydroperoxide reductase, DNA protection proteins, catalase, and the periplasmic serine protease MucD, a number of regulators were identified that are known to be related to stress response in this Gammaproteobacterium. These regulators were the oxidative stress-response regulator OxyR, and regulators such as AlgR, AlgR3, and AlgU (AlgT) (see Table S3). The latter belongs to the ECF-family of sigma factors and regulates *algD* expression. The AlgD protein, which was also detected in this study (see Table S3), is involved in biosynthesis of the exopolysaccharide alginate, which has been demonstrated to be of importance for increased epiphytic fitness, virulence, and resistance to desiccation and toxic molecules (36).

An over-representation of stress-related proteins was found in the soybean samples (see Fig. 4). This might reflect a consequence of a plant-defense response, which in turn was possibly triggered by the presence of flagellin (37) of *Pseudomonas* spp. (see below). Strains with very close relationship to the pathogen *P. syringae* pv. *glycinae* (100% sequence identity on 16S rRNA gene level) were detected on the soybean plants.

Motility. We observed a significant over-representation of flagellin in *Pseudomonas* relative to other bacteria (see Table S3, Table 2, Fig. 4 and Fig. S4). It is conceivable that *Pseudomonas* spp. rather than *Methylobacterium* spp. and *Sphingomonas* spp. have adapted a lifestyle that is predestinated to actively search for nutrients. Motility is well established as an important epiphytic fitness factor of plant colonizing *Pseudomonas* (38) and was shown to be regulated by quorum sensing (39). Apparently, *Pseudomonas* spp. are not part of the common and consistent microbiota on plants, but rather transient inhabitants probably subjected to more frequent changes in abundance (see Table S2) (see also refs. 21 and 40).

Conspicuous Proteins. Finally, we searched the metaproteomic dataset for the presence of proteins of unknown or poorly characterized function that were consistently present throughout our samples and among different bacterial species, as they may be indicative for a common trait shared by bacteria adapted to the phyllosphere. Among these proteins, "beta-Ig-H3/fasciclin" was prominent (see Table 2 and Fig. 4). Proteins of this family were detected based on genome sequence information from *Methylobacterium* (see Table 2 and Fig. S4), *Rhodopseudomonas*, *Novosphingobium*, and *Stenotrophomonas* among the most abundant proteins identified in this study (see Table S5), and from a number of other bacterial genera when considering all identified proteins (see Table S3). Homologues of this fasciclin domain protein are found in vertebrates and invertebrates and are thought to mediate cell adhesion (41). Notably, fasciclin homologues were described to be symbiotically relevant in 3 separate cases (*Nostoc*–lichens, *Rhizobium*–legume, and algae–cnidaria) (42, 43). Consequently, the fasciclin protein is a prime candidate for further investigation with regard to its importance for bacteria during the phyllospheric lifestyle and its putative role in cell-cell adhesion. Another example of a consistently detected protein in several bacterial species is given in Fig. S4 (TypA/BipA).

Conclusions

To our knowledge, this study is innovative in representing a large-scale combinatorial metagenome and metaproteome analysis from a common pool of cells. This approach allowed us to overcome limitations in protein identification that are otherwise encountered because of the absence of closely related reference genomes in publicly available databases. It also demonstrated that metagenome data, retrieved from relatively short sequence reads and with low degree of assembly, are of sufficient quality to allow protein identification of bacteria not sequenced so far. The identification of abundant proteins in the phyllosphere microbiota allowed us to detect key enzymatic functions with activities that can be expected to be relevant for global carbon and nitrogen cycles. This holds especially for the conversion of methanol, a major volatile organic compound emitted by plants (100 Tg formed per year) (27), and the assimilation of ammonia via glutamine synthetase. The latter is of relevance considering the high amount of ammonia input from agricultural sources and from industrial exhaust, as discussed in relation to the phyllosphere (44).

The identity of bacteria present in the phyllosphere in combination with the protein survey described here offers insights into strategies for phyllospheric lifestyles of bacteria on plant hosts. Our analysis revealed consistency with respect to the bacterial community composition and, in particular, the high abundance of *Sphingomonas* spp. and *Methylobacterium* spp. on the analyzed plants. Known proteins expressed in *Methylobacterium* are related, to a large extent, to one-carbon and central metabolism, as well as to stress response, whereas for *Sphingomonas* spp., the conspicuous expression of TonB-dependent receptors suggests a particularly large substrate spectrum. These adaptations contribute to the success and coexistence of these taxa in the phyllosphere. Apart from these consistently observed 2 alphaproteobacterial genera, we detected the presence of flagellated *Pseudomonas* on soybean plants and with it a number of proteins of known and unknown functions.

The survey of proteins present in situ provides a basis for targeted studies of proteins relevant in relation to the plant

environment. Strikingly, the consistent and abundant presence of some proteins of uncharacterized function in a number of different bacterial genera, of which fasciclin is one example, suggest key functions for adaptation to the phyllosphere that need to be investigated in more detail. The identity of abundant and ubiquitous commensal phyllosphere bacteria in combination with a better understanding of their physiology in this habitat will help to reveal the role of these bacteria in global carbon and nitrogen cycles, and serve as a basis to exploit them in the future with respect to a potential plant probiotic power.

Materials and Methods

Sampling of Phyllosphere Bacteria and Extraction of DNA and Protein. Bacterial cells were washed from the leaf material applying a previously published protocol (9) with slight modifications (see *SI Text*), including a centrifugation step in the presence of Percoll to deplete eukaryotic cells and dirt particles. DNA and protein extraction was performed using the AllPrep DNA/RNA/Protein Mini Kit (Qiagen). Frozen cell pellets were resuspended in 1,300 to 1,400 μl of kit-supplied RLT buffer, 1 g of 0.1-mm zirconium-silica beads was added, and cell lysis was performed in a tissue lyser (Retsch GmbH) for 3 min at maximum shaking frequency (30 s^{-1}). Cell debris and beads were pelleted for 1 min at 20,000 $\times$ *g*. The supernatant was distributed onto 2 kit-supplied columns for further extraction of the DNA and proteins according to the instructions in the kit manual.

DNA Metagenome Sequencing and Analysis. Sequencing was performed on the Genome Sequencer FLX system. All DNA sequences were assembled with the GS De Novo Assembler provided with the FLX system (Roche Applied Science and 454 Life Sciences) using default parameters for protein identification. ORFs were predicted and data annotated as outlined in the *SI Text*. Taxonomic community composition estimates based on metagenomic sequences were derived by running the software MLTreeMap on the Soybean 2 metagenomic data (13).

Microbial Community 16S rRNA Gene-Based Analysis. The bacterial and archaeal community composition of the 6 phyllosphere samples was characterized by 16S rRNA gene-clone library construction, followed by comparative sequence analysis as outlined in detail in the *SI Text*. Rarefaction curves were calculated using the Dotur software package (45).

Protein Identification and Analysis. Proteins were separated by 1-dimensional SDS/PAGE and analyzed after tryptic digestion by reversed-phase high-performance liquid-chromatography coupled to electrospray-ionization tandem mass-spectrometry. Data files obtained from high-accuracy mass spectrometers were converted to peak lists and were analyzed with 2 search algorithms and validated with Scaffold (Proteome Software Inc.). MS/MS spectra were searched against 2 different databases: one database consisting of protein sequences obtained from RefSeq (ftp://ftp.ncbi.nih.gov/refseq) and a second database built from RefSeq data plus the translated metagenomic data (see Dataset S1). For protein identification, at least 2 peptide matches were required (each having a minimum peptide identification probability of 95%; minimum required protein identification probability was 99%). The false discovery rate, as estimated by searches against a decoy database, was below 1%. Data processing and visualization were performed using custom scripts in Perl, Python, and R. Full information about all of the methods and associated references used for the analyses reported here is available in the *SI Text*.

ACKNOWLEDGMENTS. We thank Carlos Alonso Blanco (Spanish National Center for Biotechnology), Thomas Hebeisen, Daniel Suter, and Christine Herzog (Forschungsanstalt Agroscope Reckenholz-Tänikon, Switzerland) for support with the plant sampling, Marzanna Künzli (Functional Genomics Center Zurich) for support with genome sequencing, Simon Barkow-Oesterreicher (Functional Genomics Center Zurich) for support with database handling, and Roger Wepf (Electron Microscopy Center of the Eidgenössische Technische Hochschule Zurich) for support with electron microscopy imaging. We thank the Vital-IT group of the Swiss Institute of Bioinformatics for providing computational resources. The work was supported by Eidgenössische Technische Hochschule Zurich and by the University of Zurich through its Research Priority Program ''Systems Biology and Functional Genomics''.

1. Bailey MJ (2006) *Microbial Ecology of Aerial Plant Surfaces* (CABI Publishing, Wallingford).
2. Lindow SE, Brandl MT (2003) Microbiology of the phyllosphere. *Appl Environ Microbiol* 69:1875–1883.
3. Lambais MR, Crowley DE, Cury JC, Bull RC, Rodrigues RR (2006) Bacterial diversity in tree canopies of the Atlantic forest. *Science* 312:1917.
4. Redford AJ, Fierer N (2009) Bacterial succession on the leaf surface: A novel system for studying successional dynamics. *Microb Ecol* 58:189–198.
5. Yang CH, Crowley DE, Borneman J, Keen NT (2001) Microbial phyllosphere populations are more complex than previously realized. *Proc Natl Acad Sci USA* 98:3889–3894.
6. Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. *Annu Rev Microbiol* 57:369–394.
7. Leveau JHL (2006) Microbial communities in the phyllosphere. In *Biology of the Plant Cuticule*, eds Riederer M, Müller C (Blackwell, Oxford), pp 334–367.
8. Boch J, et al. (2002) Identification of *Pseudomonas syringae* pv. *tomato* genes induced during infection of *Arabidopsis thaliana*. *Mol Microbiol* 44:73–88.
9. Gourion B, Rossignol M, Vorholt JA (2006) A proteomic study of *Methylobacterium extorquens* reveals a response regulator essential for epiphytic growth. *Proc Natl Acad Sci USA* 103:13186–13191.
10. Marco ML, Legac J, Lindow SE (2005) *Pseudomonas syringae* genes induced during colonization of leaf surfaces. *Environ Microbiol* 7:1379–1391.
11. Yang S, et al. (2004) Genome-wide identification of plant-upregulated genes of *Erwinia chrysanthemi* 3937 using a GFP-based IVET leaf array. *Mol Plant Microbe Interact* 17:999–1008.
12. VerBerkmoes NC, Denef VJ, Hettich RL, Banfield JF (2009) Systems biology: Functional analysis of natural microbial consortia using community proteomics. *Nat Rev Microbiol* 7:196–205.
13. von Mering C, et al. (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* 315:1126–1130.
14. Corpe WA, Rheem S (1989) Ecology of the methylotrophic bacteria on living leaf surfaces. *FEMS Microbiol Ecol* 62:243–250.
15. Kim H, et al. (1998) High population of *Sphingomonas* species on plant surface. *J Appl Microbiol* 85:731–736.
16. Knief C, Frances L, Cantet F, Vorholt JA (2008) Cultivation-independent characterization of *Methylobacterium* populations in the plant phyllosphere by automated ribosomal intergenic spacer analysis. *Appl Environ Microbiol* 74:2218–2228.
17. Acinas SG, et al. (2004) Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* 430:551–554.
18. Eckburg PB, et al. (2005) Diversity of the human intestinal microbial flora. *Science* 308:1635–1638.
19. Hongoh Y, et al. (2005) Intra- and interspecific comparisons of bacterial diversity and community structure support coevolution of gut microbiota and termite host. *Appl Environ Microbiol* 71:6590–6599.
20. Tringe SG, et al. (2005) Comparative metagenomics of microbial communities. *Science* 308:554–557.
21. Ellis RJ, Thompson IP, Bailey MJ (1999) Temporal fluctuations in the pseudomonad population associated with sugar beet leaves. *FEMS Microbiol Ecol* 28:345–356.
22. Kinkel LL (1997) Microbial population dynamics on leaves. *Annu Rev Phytopathol* 35:327–347.
23. Liu H, Sadygov RG, Yates JR, 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* 76:4193–4201.
24. Finn RD, et al. (2008) The Pfam protein families database. *Nucleic Acids Res* 36:D281–D288.
25. Schauer K, Rodionov DA, de Reuse H (2008) New substrates for TonB-dependent transport: do we only see the 'tip of the iceberg'? *Trends Biochem Sci* 33:330–338.
26. Blanvillain S, et al. (2007) Plant carbohydrate scavenging through TonB-dependent receptors: a feature shared by phytopathogenic and aquatic bacteria. *PLoS ONE* 2:e224.
27. Galbally IE, Kirstine W (2002) The production of methanol by flowering plants and the global cycle of methanol. *J Atmosph Chem* 43:195–229.
28. Sy A, Timmers AC, Knief C, Vorholt JA (2005) Methylotrophic metabolism is advantageous for *Methylobacterium extorquens* during colonization of *Medicago truncatula* under competitive conditions. *Appl Environ Microbiol* 71:7245–7252.
29. Vorholt JA (2002) Cofactor-dependent pathways of formaldehyde oxidation in methylotrophic bacteria. *Arch Microbiol* 178:239–249.
30. Chistoserdova L, Chen SW, Lapidus A, Lidstrom ME (2003) Methylotrophy in *Methylobacterium extorquens* AM1 from a genomic point of view. *J Bacteriol* 185:2980–2987.
31. Bosch G, et al. (2008) Comprehensive proteomics of *Methylobacterium extorquens* AM1 metabolism under single carbon and nonmethylotrophic conditions. *Proteomics* 8:3494–3505.
32. Chistoserdova L, Lidstrom ME (1997) Molecular and mutational analysis of a DNA region separating two methylotrophy gene clusters in *Methylobacterium extorquens* AM1. *Microbiology* 143:1729–1736.
33. Chistoserdova L, Kalyuzhnaya MG, Lidstrom ME (2009) The expanding world of methylotrophic metabolism. *Annu Rev Microbiol* 63:477–499.
34. Gourion B, Francez-Charlot A, Vorholt JA (2008) PhyR is involved in the general stress response of *Methylobacterium extorquens* AM1. *J Bacteriol* 190:1027–1035.
35. Francez-Charlot A, et al. (2009) Sigma factor mimicry involved in regulation of general stress response. *Proc Natl Acad Sci USA* 106:3467–3472.
36. Yu J, Penaloza-Vazquez A, Chakrabarty AM, Bender CL (1999) Involvement of the exopolysaccharide alginate in the virulence and epiphytic fitness of *Pseudomonas syringae* pv. *syringae*. *Mol Microbiol* 33:712–720.
37. Boller T, He SY (2009) Innate immunity in plants: an arms race between pattern recognition receptors in plants and effectors in microbial pathogens. *Science* 324:742–744.
38. Haefele DM, Lindow SE (1987) Flagellar motility confers epiphytic fitness advantages upon *Pseudomonas syringae*. *Appl Environ Microbiol* 53:2528–2533.
39. Quinones B, Dulla G, Lindow SE (2005) Quorum sensing regulates exopolysaccharide production, motility, and virulence in *Pseudomonas syringae*. *Mol Plant Microbe Interact* 18:682–693.
40. Hirano SS, Upper CD (2000) Bacteria in the leaf ecosystem with emphasis on *Pseudomonas syringae*—a pathogen, ice nucleus, and epiphyte. *Microbiol Mol Biol Rev* 64:624–653.
41. Carr MD, et al. (2003) Solution structure of the *Mycobacterium tuberculosis* complex protein MPB70: from tuberculosis pathogenesis to inherited human corneal disease. *J Biol Chem* 278:43736–43743.
42. Oke V, Long SR (1999) Bacterial genes induced within the nodule during the *Rhizobium*-legume symbiosis. *Mol Microbiol* 32:837–849.
43. Paulsrud P, Lindblad P (2002) Fasciclin domain proteins are present in *Nostoc* symbionts of lichens. *Appl Environ Microbiol* 68:2036–2039.
44. Papen H, Gessler A, Zumbusch E, Rennenberg H (2002) Chemolithoautotrophic nitrifiers in the phyllosphere of a spruce ecosystem receiving high atmospheric nitrogen input. *Curr Microbiol* 44:56–60.
45. Schloss PD, Handelsman J (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* 71:1501–1506.

Table S1. Characterization of sampling material

Sample	Plant species	Place*	Position	Sampling date	Plant biomass, g of dry weight	Number of cells recovered†	DNA recovery, µg	Number of MS/MS spectra	Plant/leaf age
Soybean 1	*Glycine max* (Gallec)	Effretikon, Switzerland	N 47°26'46'' E 8°41'13''	16 July 2007	56.7	4.3×10^9	2.4	61,762	67 d, Developmental stage‡: R1 – R2, flowering
Soybean 2	*Glycine max* (Gallec)	„	„	28 / 29 August 2007	200.3	7.6×10^8	28.4	99,916	110 d, Developmental stage‡: R5, begin of bean development
Clover 1a	*Trifolium repens* (Tetra)	Reckenholz, Switzerland	N 47°26'27'' E 8°29'54''	13 June 2007	31.3	3.4×10^9	4.2	59,529	9 days after mowing
Clover 1b	*Trifolium repens* (breed unknown)	„	„	13 June 2007	18.3	4.7×10^9	8.6	71,204	42 days after mowing
Clover 2	*Trifolium repens* (Tetra)	„	„	3 / 4 September 2007	77.4	1.3×10^{10}	24.2	111,868	34 days after mowing
Arabidopsis	*Arabidopsis thaliana*	Ciruelos de Coca, Spain	N 41°12'48'' W 4°32'44''	28 March 2007	—§	not determined	20.6	83,025	Seed germination and leaf development started approx. half a year before sampling

*Soybean and clover samples were collected from a spot of approx. 5 m^2 in the middle of agricultural fields. The soybean field (24 × 150 m) was planted with different soybean varieties. The field from which the clover samples were collected was an experimental site (150 × 150 m) divided into 9 m^2 plots planted with different grass or clover varieties. The distance of the plots from which samples 'Clover 1a' and 'Clover 1b' were collected was approximately 60 m. *Arabidopsis* plants were collected at the edge of a pine tree forest from an area of 600 × 10 m. The individual *Arabidopsis* plants grew without direct leaf contact to other plants.

†cell numbers were roughly estimated by direct microscopic counting using a Helber counting chamber

‡classification according to Fehr, W. R., Caviness, C. E., Burmood, D. T. Pennington, J. S. (1971) Stage of development descritptions for soybeans, *Glycine-Max* (L.) Merrill. *Crop Science* 11:929-931.

§Plant dry weight was not determined for this sample; rosette leaves of 389 individual plants were used

Table S5. Most abundant proteins detected in phyllosphere bacteria not assigned to *Methylobacterium, Sphingomonas*, and *Pseudomonas* (Table 2)

Most abundant proteins	Genus	Identification	SY1	SY2	CL1a	CL1b	CL2	ARA
Porin, gram-negative type	*Verminephrobacter*	Metagenome	++	+++	++	+	+	+
Beta-Ig-H3/fasciclin	*Rhodopseudomonas*	RefSeq	+	+	++	++	++	+
Beta-Ig-H3/fasciclin	*Novosphingobium*	Metagenome	n.d.	+	+++	++	++	+
Novel	UNKNOWN	Metagenome	+	++	+++	+	n.d.	n.d.
Novel	UNKNOWN	Metagenome	+	+++	+	+	n.d.	n.d.
Beta-Ig-H3/fasciclin	*Stenotrophomonas*	Metagenome	+	++	+	++	+	++
Histone-like protein	*Agrobacterium*	RefSeq	n.d.	++	+	+	+	+
DNA-binding protein HU	*Cytophaga*	Metagenome	n.d.	+	+	+	+	++
Novel	UNKNOWN	Metagenome	+	++	+	+	+	+
Novel	UNKNOWN	Metagenome	n.d.	n.d.	+	+	+	+
Putative uncharacterized protein	*Burkholderia*	Metagenome	+	++	+	+	+	+
DNA-binding protein HU	*Clavibacter*	Metagenome	+	+	n.d.	n.d.	n.d.	++
DNA protein during starvation, Dps	*Sorangium*	Metagenome	+	+	+	n.d.	n.d.	+
ATP-dependent Clp protease	*Zymononas*	Metagenome	+	+	+	+	n.d.	+
Novel	UNKNOWN	Metagenome	+	+	+	+	+	+
Novel	UNKNOWN	Metagenome	n.d.	n.d.	+	+	+	n.d.
Novel	UNKNOWN	Metagenome	+	+	+	n.d.	n.d.	n.d.
Novel	UNKNOWN	Metagenome	+	+	+	+	+	n.d.
Enolase	*Planctomyces*	Metagenome	n.d.	+	+	+	n.d.	+
Cold-shock DNA-binding domain protein	*Arthrobacter*	RefSeq	n.d.	+	n.d.	n.d.	n.d.	++

Proteins were grouped if 90% identical over at least 40% of their length. Taxonomy (at the genus level) was inferred from the protein annotation (RefSeq/metagenome), if available. Ribosomal proteins are not reported here, but are listed in Table S3. Relative abundances are displayed with +, ++, and +++; n.d., not detected. SY1, soybean 1; SY2, soybean 2; CL1a, clover 1a; CL1b, clover 1b; CL2, clover 2; ARA, *Arabidopsis* .

Supporting Information

Delmotte et al. 10.1073/pnas.0905240106

SI Materials and Methods

Harvest of Prokaryotic Phyllosphere Cells. Plant leaf material [i.e., rosettes of thale cress (*Arabidopsis thaliana*), fully developed leaves of soybean (*Glycine max*), or fully developed trifoliates of clover (*Trifolium repens*)] was placed in 50-mL tubes and the tubes were filled up to 30 ml with sterile, precooled TE-buffer (10 mM Tris, 1 mM EDTA, pH 7.5), supplemented with 0.3 g mL^{-1} Pefabloc SC (Roche Diagnostics) and 0.2% Silwet L-77 (GE Bayer Silicones). Cells were washed from the leaf material by 3 min of alternate shaking, vortexing, and sonication. The cell suspension was separated from the leaf material by filtration through a nylon mesh (pore size, 200 μm; Spectrum Europe BV). Six milliliters of 80% Percoll (Sigma-Aldrich) was pipetted below the cell suspension, and the 50-mL tubes were centrifuged for 5 min at 800 $\times$ *g*. The bacterial cell suspension above the Percoll layer was transferred into a fresh 50-mL tube and cells were pelleted at 3,150 $\times$ *g* for 15 min. Cell pellets from multiple tubes were pooled into 1.5-ml reaction tubes and washed twice with TE-buffer plus Pefabloc SC. Cell pellets were immediately frozen at –20 °C.

Construction and Analysis of 16S rRNA Gene-Clone Libraries. Bacterial 16S rRNA genes were amplified in triplicate PCR assays (volume of 33 μl each, prepared from a master mix of 100 μl). Each 100-μl assay contained 10 μl of supplied RedAccu LA Taq Polymerase PCR buffer containing 2.5 mM of Mg^{2+} (Sigma), 1.25 mM of each deoxynucleoside triphosphate (dNTP) (Fermentas), 0.5 μM of each primer (Microsynth), 0.25 μg μL^{-1} of BSA (Roche Diagnostics), 0.05 U μL^{-1} of Red Accu LA Taq polymerase (Sigma), and 5 μl of template DNA. Primers 9f and 1492r were used for PCR amplification of bacteria (1). The PCR program consisted of initial denaturation at 94 °C for 4 min, followed by 25 cycles of denaturation at 94 °C for 45 s, annealing at 48 °C for 1 min, and elongation at 72 °C for 2 min, and then a final elongation at 72 °C for 7 min. PCR products were purified with a NucleoSpin Extract II purification kit (Machery-Nagel), and A-overlaps were replenished in an assay containing 5-μl purified PCR product, 0.6 μl of supplied Master Taq Polymerase buffer (Eppendorf), 0.3 μl of each dNTP, and 0.3 μl of Master Taq Polymerase (Eppendorf) by incubation at 72 °C for 10 min. For the detection of Archaea, a specific primer system was applied (20f + 958r) (2). PCR was performed in assays as described above using the thermal profile as described (2) with 35 reaction cycles. PCR products could be obtained in the Soybean 1, Soybean 2, Clover 2, and *A. thaliana* samples.

After cloning and sequencing with primers 9f and 1492r, the nearly full-length 16S rRNA gene sequences were obtained after assembly and aligned using the SINA webaligner of the SILVA ribosomal database project (3). Sequences were double-checked for chimeras using the Mallard program and the chimera detection program of the ribosomal database project RDP, release 8. Sequences that showed anomalies with only 1 of the 2 programs were manually checked with Pintail. Moreover, the aligned sequences were visually inspected for anomalies.

Phylogenetic trees were calculated using the maximum-likelihood algorithm PhyML, implemented in the ARB software package. Type strains for the trees shown in Fig. S1 *a–d* were selected according to "The All-Species Living Tree project," release 93 (4).

Denaturing Gradient Gel Electrophoresis. DGGE was performed as previously described (5). Briefly, primers 533f and 907r-GC were applied to PCR-amplify a fragment of the 16S rRNA gene with 35 cycles. PCR products were quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen), and equal amounts of DNA were loaded onto each gel. Acrylamide gels (6.5%) were prepared with a denaturing gradient from 35 to 65%, and gels were run at 60 °C and 70 V for 16 h. Excised bands were reamplified with 25 PCR cycles and the correct migration behavior was checked on a DGGE before sequencing. The community composition of the 6 samples used for metaproteomic analysis was additionally analyzed by using a second primer system, 357f-GC and 907r (6), which revealed a comparable clustering of samples (i.e., samples from the same plant species clustered together). DGGE patterns were compared with the GelCompar II software (Applied Maths). Cluster analysis was performed using the unweighted-pair group method using arithmetic averages algorithm based on Pearson correlation coefficients.

DNA Metagenome Sequence Analysis. Pyrosequencing was performed by GATC and at the Functional Genomics Center Zurich using an aliquot from the DNA extract of the Soybean 2 sample. Five micrograms of DNA was provided for each analysis. DNA quantity and purity (based on the ratio of absorbance at 260 and 280 nm and was 1.7 for our sample) was determined using the NanoDrop (Thermo Fisher Scientific) and the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen). Data assembly using the GS De Novo Assembler resulted in 140,550 contigs with a mean sequence length of 276 bp, or 40-bp longer than the mean length of a single read. The largest contig had a length of 12,888 bp. After assembly, different read statuses were attributed to each read by the assembler software: assembled, partially assembled (only part of the read included), singleton (no overlap with any other read), repeat (identified as a repeat region or exactly duplicated sequence; known artifact of the pyrosequencing technique), or outlier (problematic read; for example, chimera sequences). To build the metagenome database (for proteomic data annotation), singleton reads were included in the contigs file in order not to lose any information after assembly. The annotation of contigs and singleton reads was performed as follows: ORF prediction, with translation of regions between stop codons in the 6 reading frames, was done using the program *getorf* (EMBOSS package). ORFs with a minimum size of 10 aa were reported. Similarity searches for all predicted ORFs were performed using the program BLASTp with an expected (E) value cutoff of 0.0001 (and the following parameters: "-M BLOSUM62 -G 11 -E 1 -F T") against the database UniRef90. A hit was considered significant with a bitscore larger than or equal to 60. Pfam domains (7) were reported for ORFs that significantly matched UniRef90 using the mapping file *protein2ipr.dat.gz* available on the Interpro ftp Web site. A domain was reported if containing a minimum overlap of 20 aa with the contig/read. A total of 319,651 ORFs matched those criteria. All nonannotated ORFs (5,647,279) were kept in the metagenome database (total of 5,966,930 entries) for further analysis in case of identification by MS.

Preparation of Proteins for MS. The extracted protein fraction of each sample, obtained as indicated above, was processed further using the Allprep DNA/RNA/Protein kit (Qiagen). Proteins were precipitated and then dissolved in a Laemmli-related kit-supplied sample buffer. If needed, proteins were frozen and stored at −20 °C; otherwise, the proteins were diluted up to 45

μl in loading buffer and denaturated for 4 min at 95 °C. Loading buffer was prepared by mixing 125 μl of 0.5 M Tris-HCl, pH 6.8, 250 μl of glycerol, 200 μl of 10% SDS, 50 μl of 2-β-mercaptoethanol, and 1 crystal of bromophenol blue and then bringing the solution to a final volume of 2 ml with water. After cooling and centrifugation at 20,238 $\times$ *g* for 5 min, the protein sample was loaded for separation on the top of a Tris-HCl polyacrylamide gel (4–15% linear gradient, 8.6 $\times$ 6.8 cm, or 10.5–14% linear gradient, 13.3 $\times$ 8.7 cm) obtained from Bio-Rad Laboratories AG. Electrolysis buffer consisted of 25 mM Tris-HCl, pH 8.3, 192 mM glycine, and 0.1% SDS. Staining was performed for 40 min with 40% methanol, 10% acetic acid, and 0.25% Coomassie blue. Destaining was achieved overnight with 10% methanol and 10% acetic acid. For each sample, the corresponding gel lane was cut into 16 to 21 pieces. Gel pieces were destained 3 times with 50% acetonitrile and dried for 10 min under vacuum (Model SPD121P SpeedVac, Thermo Fisher Scientific). Then, proteins were reduced for 45 min at 56 °C with Tris(2-carboxyethyl)phosphine hydrochloride (2 mM in 25 mM ammonium hydrogen carbonate, pH 8.0) and carbamidomethylated for 60 min at room temperature in the dark with iodoacetamide (25 mM in 25 mM ammonium hydrogen carbonate, pH 8.0). Gel plugs were washed 3 times with 50% acetonitrile and dried for 15 min under vacuum. Finally, proteins were digested with trypsin (Promega) for 16 h at 37 °C (50 ng/gel plug) in 25-mM ammonium hydrogen carbonate, pH 8.0. Digestion was quenched with trifluoroacetic acid, digests were transferred to new vessels and solvents were evaporated. After resolubilisation in 30 μl of 3% acetonitrile and 0.1% trifluoroacetic acid, peptides were cleaned up with a C18 ZipTip supplied by Millipore Corporation.

MS Analysis. The samples were analyzed on a hybrid LTQ-Orbitrap XL mass spectrometer (Thermo Fisher Scientific) interfaced with a nanoelectrospray source. Peptides were separated by reversed-phase high-performance liquid-chromatography on an in-house packed column with 2 μm UltraHT Pro C18 packing material from YMC Co. Column dimensions were 80 $\times$ 0.75 mm inside diameter. Eluents were (A) 1% acetonitrile, 0.2% formic acid, and (B) 80% acetonitrile, 0.2% formic acid. Separation was performed by linear gradients of 3 to 10% (B) in 5 min, 10 to 40% (B) in 50 min, 40 to 97% (B) in 5 min, followed by isocratic conditions at 97% (B) for 5 min. Solvent delivery of 200 nL min^{-1} was achieved by a binary gradient pump (Model nanoLC 1D Plus, Eksigent). Peptides were loaded from a cooled (4 °C) auto sampler (Model Endurance, Spark Holland). Connection of the reversed-phase column with the ESI source was achieved by stretching the fused silica capillary at the outgoing extremity of the column.

MS detection was performed with the LTQ-Orbitrap XL mass spectrometer operating in data-dependent mode. The 4 most abundant doubly or triply charged ions from the high-accuracy survey scan with a minimum ion count of 500 were automatically taken for further MS/MS analysis at the linear ion trap. Precursor masses already taken for MS/MS were excluded for further selection for 60 s. All mass spectra were recorded in positive ion mode with an electrospray source voltage between 1.5 kV and 1.90 kV. Precursor mass spectra were acquired at the Orbitrap mass analyzer with a scan range from *m/z* 300.0 to 1,600.0 using real-time internal calibration on polydimethylcyclosiloxane background ions *m/z* 445.120025 and 429.088735, as previously described (8). Resolution was set to 60,000 at *m/z 400*. For some remeasurements, a hybrid LTQ-FTICR mass spectrometer (Model LTQ-FT Ultra, Thermo Fisher Scientific) was used. Chromatographic separation, ionization, and data acquisition were performed as described for the LTQ-Orbitrap XL mass spectrometer.

Protein Identification and Determination of False-Discovery Rate. Mass spectra processing was performed with Xcalibur 2.0.7 (Thermo Fisher Scientific). Peak list generation for database searches was performed with Mascot Distiller 2.1.1.0 (Matrix Science). Database searches were performed against 3 different databases. The first database (DB1), containing 5,195,116 protein sequences, consisted of RefSeq Release 28 and was downloaded from the NCBI ftp Web site (ftp://ftp.ncbi.nih.gov/refseq/release). The second database (DB2) was a concatenation of DB1 with 5,966,930 sequences issued from the metagenomics part of the project and had a total of 11,162,046 entries. The third database (DB3) had 15,285 entries and consisted of all of the protein sequences from 3 reference complete genomes, *Methylobacterium extorquens* PA1, *Sphingomonas wittichii* RW1, and *Pseudomonas syringae* pv. *phaseolicola* 1448A, downloaded from the NCBI ftp Web site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). The selection of reference genomes was based on number of identified proteins for different species within the genus and was thus based on results shown in Table S3.

For database searches, a first computation was performed with Mascot 2.2 (Matrix Science) based on the MOWSE algorithm (9). The following search parameters were applied: taxonomy, all entries; fixed modification, cysteine carbamidomethylation; variable modifications, methionine oxidation; enzyme, trypsin; maximum number of missed cleavages, 1; peptide tolerance, $\pm$ 5 ppm; MS/MS tolerance, $\pm$ 0.5 Da. We were able to set a low peptide tolerance (5 ppm) because of the high accuracy of the Orbitrap mass spectrometer and the use of internal lock-mass calibration with the polydimethylcyclosiloxane background ions. By acquiring the data with high accuracy we were able to obtain peptide matches with high MOWSE scores (high quality), which are above identity cutoffs computed by Mascot (typically 40 with huge databases). A second database search was performed by using the X!Tandem database searching program (10). Results from both algorithms were validated with Scaffold 2.1 (Proteome Software Inc.). Peptide identifications were accepted if they could be established at greater than 95% probability as specified by the Peptide Prophet algorithm (11). Probability [in the sense of the Protein Prophet algorithm (12)] greater than 99% was required to validate protein identifications. One-hit wonders were removed (only proteins identified with at least 2 peptides were considered) and proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principle of parsimony.

To check the quality of our validation process, we prepared a synthetic mixture of 10 bacteria occurring in environmental samples. Both gram-positive and gram-negative species were represented and protein concentrations varied over 2 orders of magnitude between the different species. The protein mixture was processed as described for the real samples. Mass spectra were searched against DB1. Less than 1% of the hits we obtained were false-positive.

We also computed the false-discovery rate by testing experimental mass lists against a composite version of database DB2, created by concatenating the target protein sequences with reversed sequences (total of 22,324,092 sequences, target-decoy searches) as described by Elias and Gygi (13). Because we searched the mass lists against a database containing forward and reverse sequences, the number of identified reverse hits was multiplied by 2 and divided by the total number of identifications. We computed a false-discovery rate lower than 1%.

Contrary to classical proteomics, for which protein assignment to a given organism is obvious, protein assignment to a given taxon in community proteomics may remain uncertain (see also ref. 14).

Spectral Counting. Given the redundancy and diversity of identified proteins using the database DB2, we performed a clustering of the corresponding sequences to facilitate the interpretation of the results and to be able to roughly estimate protein expression by spectral counting (15). A single linkage clustering based on sequence identity was performed using the program BLASTCLUST and the following parameters "-p T -e F -L .4 -b T -S 90." Sequences aligning at least 40% of their length and with an identity superior or equal to 90% were clustered together. For a given sample, the cluster spectral count (the sum of spectral counts for all of the proteins in a given cluster) was normalized according to the total number of spectra acquired for this sample. Because longer proteins have a greater chance to be detected via MS, we also normalized cluster spectral counts by the longest protein length present in a cluster. Finally, we report the normalized spectral counts as +++, ++, and + for values ≥ 1.7, ≥ 0.9 and <1.7, and <0.9, respectively.

To better characterize clusters, biologically and functionally, we annotated them using the Gene Ontology database (http://www.geneontology.org/) using precomputed annotation available for Uniprot proteins in the GOA database (http://www.ebi.ac.uk/GOA/) and the online Protein Identifier Cross-Reference Service (http://www.ebi.ac.uk/Tools/picr/).

Differential Proteome Composition. The similarity between plant-sample proteomes was analyzed based on expressed Pfam protein domains. Spectral counting was performed to semiquantitatively estimate protein abundance. For each known protein domain (Pfam), these abundances were then aggregated based on the protein/domain mappings (in this case, we used the whole length of the identified protein; that is, even for cases where a given domain was not itself covered by peptides, it clocked counts based on the annotated occurrence of that domain in the protein). To investigate which protein domains were consistently expressed or plant-specifically enriched, we pooled the samples according to the plant species. A triangular representation was used to visualize the specific enrichments of domains detected on each plant species. Each protein domain is represented by 1 dot within the triangle, whereby the position of the dot signifies the relative enrichment of the domain in one or several of the samples. Domains that are equally frequent on all 3 plants appear in the middle of the triangle. Domains that appear in 1 of the corners of the triangle are found primarily on 1 of the plants, and domains that appear along 1 of the edges of the triangle are found primarily in 2 of the 3 sample pools, but are largely absent from the third. For each protein domain, the relative counts for the 3 habitats were normalized to add up to 1 (after addition of pseudocounts to select against rare domains). This permitted the display of 3-dimensional data in 2 dimensions (using 3 axes at 120° angles). Statistical significance assessment was performed using a Monte-Carlo method (comparison to randomized data). For more details on the method, see Tringe et al. (16).

Two-Way Fragment Recruitment. The fragment recruitment analysis was developed using a custom Python script to integrate the data and generate fragment recruitment plots. The DNA short reads recruitment was performed using the program BLAST (17) and the following parameters "a 3 -F 'L;m;' -e 0.0001 -G 5 -E 2 -r 2." All 454 reads were searched for similarity against a database (DB3) containing the 3 reference genomes (*Methylobacterium extorquens* PA1, *Sphingomonas wittichii* RW1, and *Pseudomonas syringae phaseolicola* 1448A) and their respective plasmid sequences downloaded from the RefSeq database. Best hits on a given genome were defined by the best bitscore and a bitscore cutoff superior or equal to 50. For the postanalysis and genus-taxa encoding level estimations, an identity cutoff of 90% was applied to select reads assigned to a given genome. The read coverage of a gene was defined as the sum of the aligned length of each read respecting these cutoffs, expressed in nucleotide.

The peptide recruitment was based on the Mascot score reported by the Scaffold software when searching the database DB3 as for the DNA recruitment; no other cutoff was applied to identify peptides assigned to a reference genome. To compare relative expression between genes (*mxaF* and *xoxF*) (see Fig. S5), we defined the expression level of a given gene using the following calculation: (number_of_spectra/gene_length)/read_coverage. The relative expression ratio of 2 genes is the ratio of their gene relative expression.

1. Weisburg W, Barns G, Dale SM, Pelletier DA, Lane DJ (1991) 16S ribosomal DNA amplification for phylogenetic study. *J Bacteriol* 2:697–703.
2. Ochsenreiter T, Selezi D, Quaiser A, Bonch-Osmolovskaya L, Schleper C (2003) Diversity and abundance of Crenarchaeota in terrestrial habitats studied by 16S RNA surveys and real time PCR. *Environ Microbiol* 5:787–797.
3. Pruesse E, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35:7188–7196.
4. Yarza P, et al. (2008) The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst Appl Microbiol* 31:241–250.
5. Henckel T, Friedrich M, Conrad R (1999) Molecular analyses of the methane-oxidizing microbial community in rice field soil by targeting the genes of the 16S rRNA, particulate methane monooxygenase, and methanol dehydrogenase. *Appl Environ Microbiol* 65:1980–1990.
6. Green SJ, Minz D (2005) Suicide polymerase endonuclease restriction, a novel technique for enhancing PCR amplification of minor DNA templates. *Appl Environ Microbiol* 71:4721–4727.
7. Finn RD, et al. (2008) The Pfam protein families database. *Nucleic Acids Res* 36:D281–D288.
8. Olsen JV, et al. (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol Cell Proteomics* 4:2010–2021.
9. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20:3551–3567.
10. Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20:1466–1467.
11. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 74:5383–5392.
12. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 75:4646–4658.
13. Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 4:207–214.
14. VerBerkmoes NC, Denef VJ, Hettich RL, Banfield JF (2009) Systems biology: Functional analysis of natural microbial consortia using community proteomics. *Nat Rev Microbiol* 7:196–205.
15. Zhang B, et al. (2006) Detecting differential and correlated protein expression in label-free shotgun proteomics. *J Proteome Res* 5:2909–2918.
16. Tringe SG, et al. (2005) Comparative metagenomics of microbial communities. *Science* 308:554–557.
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.

a

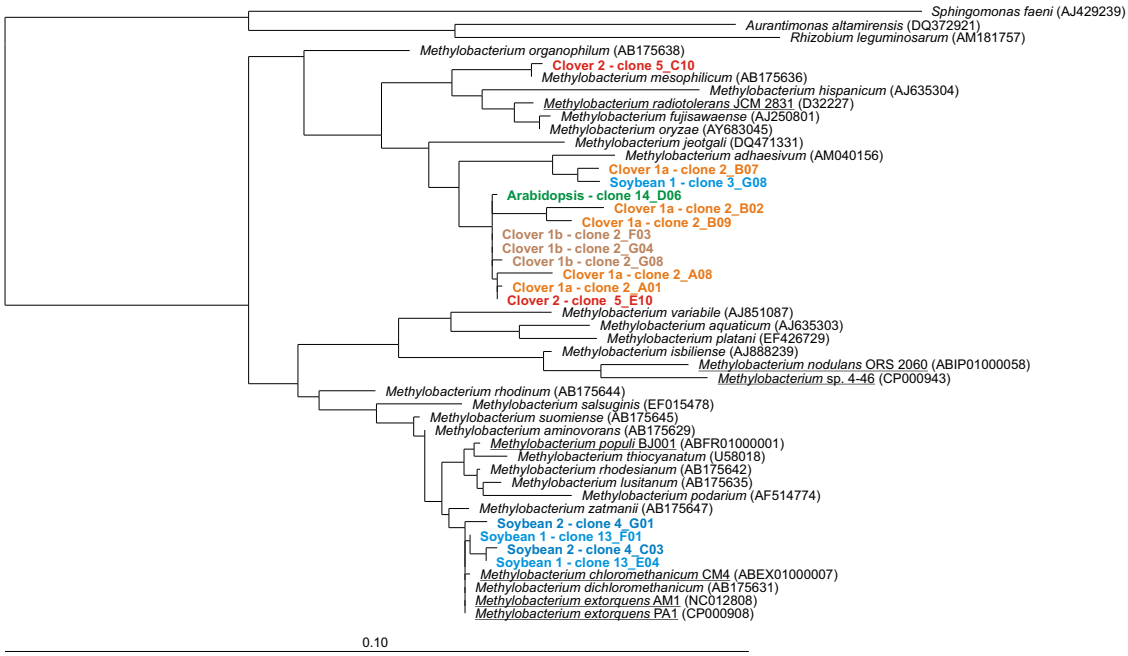


Fig. S1. Phylogenetic trees of 16S rRNA gene sequences obtained from clone libraries of all 6 samples. Phylogenetic relationship of nearly full-length 16S rRNA gene sequences to sequences of type strains and genome-sequenced strains (*underlined*) of the genera (*a*) *Methylobacterium*, (*b*) *Sphingomonas*, and (*c*) *Pseudomonas*. (*d*) The phylogenetic position of all other sequences detected in the clone libraries is shown with regard to the most closely related sequence and to sequences of cultivated reference organisms. All trees were constructed based on 1,388 nucleotide positions with the maximum likelihood algorithm PhyML. The bar represents 10% sequence divergence.

Fig. S1. Continued.

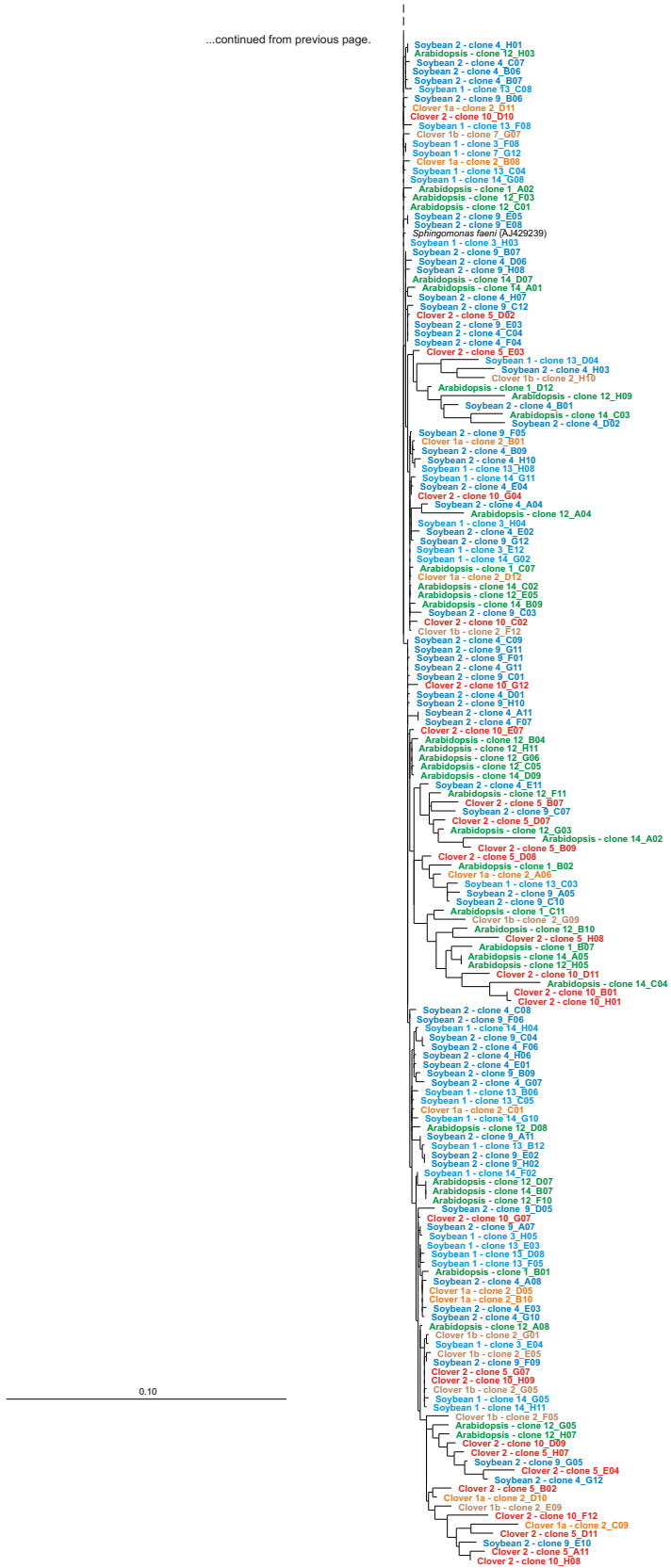


Fig. S1. Continued.

Fig. S1. Continued.

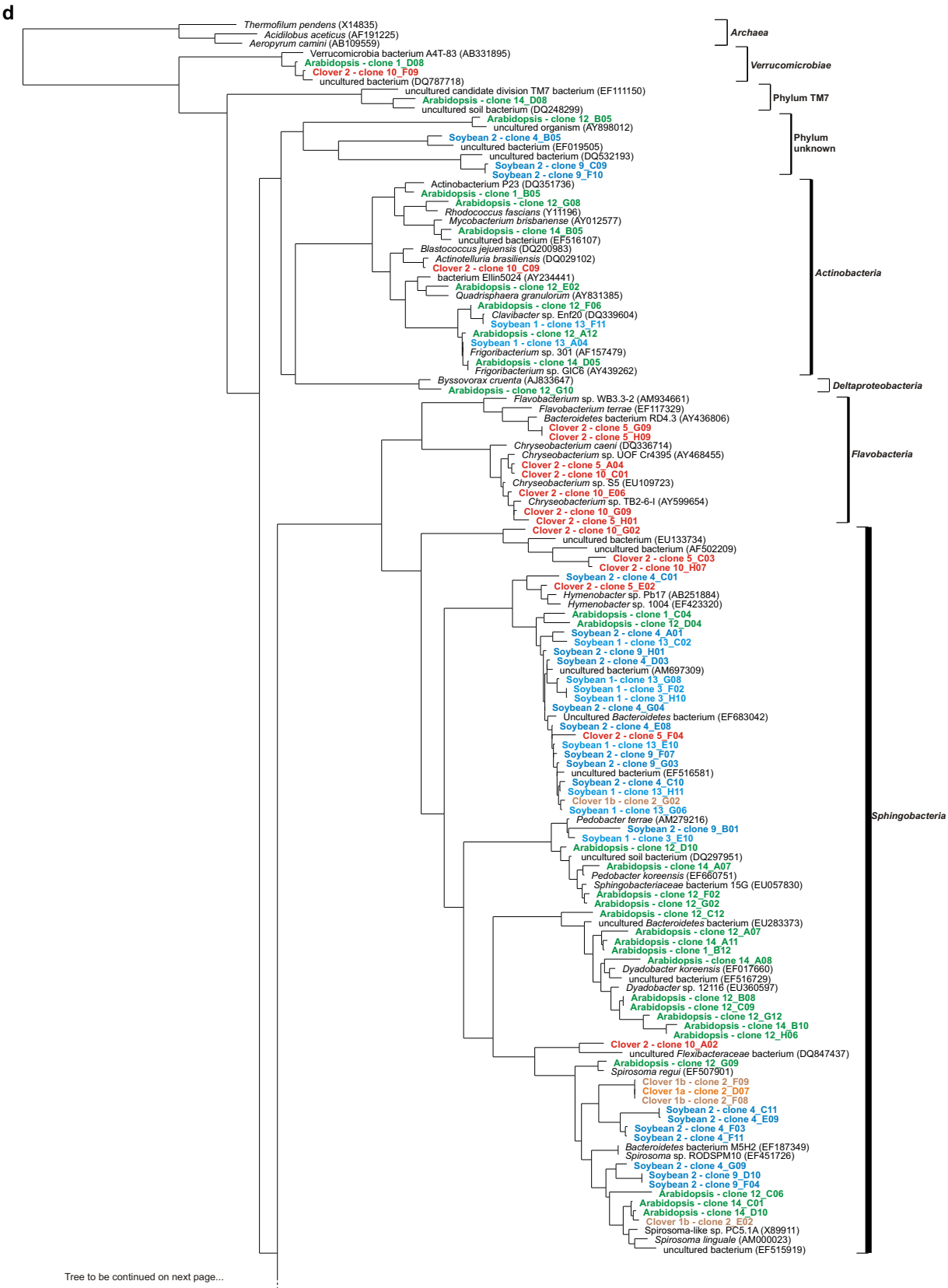


Fig. S1. Continued.

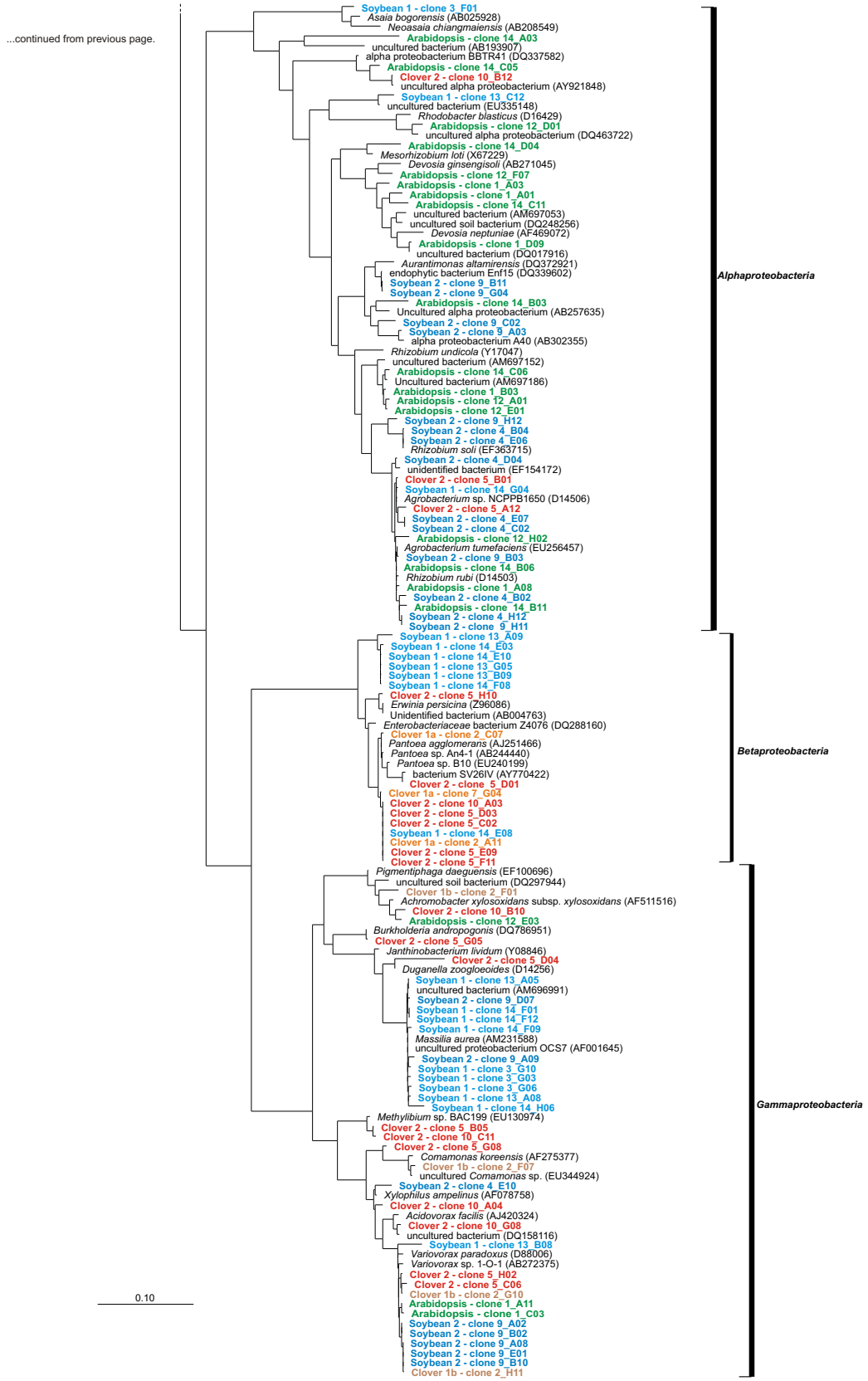


Fig. S1. Continued.

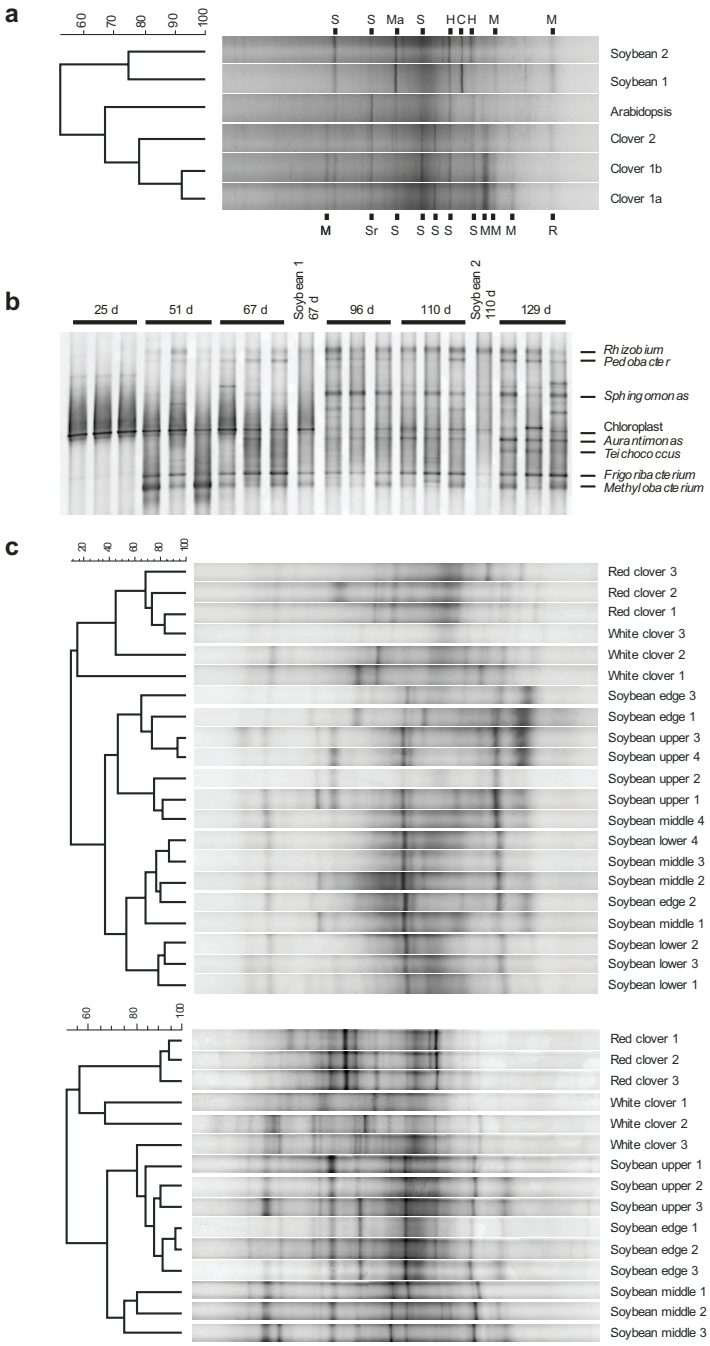


Fig. S2. Composition, similarities, and dissimilarities of microbial phyllosphere communities analyzed by denaturing gradient gel electrophoresis of 16S rRNA gene-amplification products. Cluster analyses of DGGE profiles were done using the unweighted-pair group method using arithmetic averages algorithm based on Pearson correlation coefficients. (*a*) Community composition of the 6 phyllosphere samples as seen based on DGGE banding patterns and sequencing of representative bands. The identity of bands from soybean and clover samples is indicated as follows: (H) *Hymenobacter*, (M) *Methylobacterium*, (Ma) *Massilia*, (R) *Rhodococcus*, (S) *Sphingomonas*, (Sr) *Spirosoma*, (C) Chloroplast. (*b*) Temporal succession of the bacterial phyllosphere community on leaves of soybean plants. Young leaves were collected at the beginning, while older leaves from the middle part of the plant were collected from day 67 on. The fingerprints suggest that soybean leaves were colonized by a rather stable bacterial community over the growing period. Early colonizers were still detectable at later developmental stages. Samples collected for the proteogenomic analysis ($n > 300$ leaves) reflect the community composition of the subsamples ($n = 3$ to 4 leaves per replicate) taken at the same time. (*c*) Comparison of microbial communities on soybean leaves taken from the upper, middle, and lower parts of the plant and, therefore, of different age. Additional samples were included from soybean grown at the edge of the field (from the middle part of the plant) and adjacent clover plants (white and red clover). The analysis was performed on material taken on days on which samples for metaproteomic analyses were taken: for example, day 67 (*Upper*) and day 110 (*Lower*). Bacterial communities on leaves of the soybean plants were in general more similar to each other than to communities on the clover plants, suggesting a plant species-specific phyllosphere colonization. Most of the dominant bands were detected on all different soybean samples, regardless of leaf age, leaf position, or plant position in the field (center versus edge), suggesting the presence of a spatially homogeneous bacterial phyllosphere population.

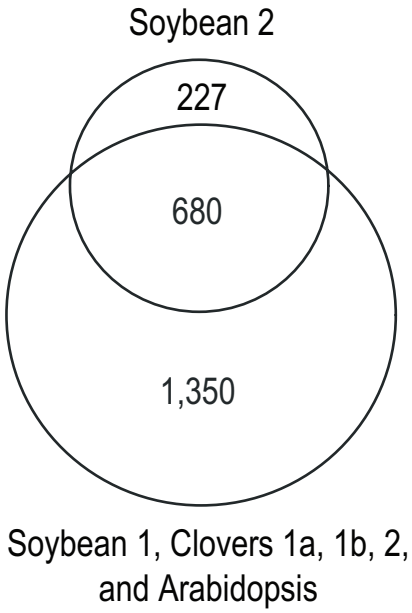


Fig. S3. Venn diagram showing the overlap of proteins identified in the Soybean 2 sample relative to the other plant phyllosphere samples. Peak lists of each sample were searched against DB2 and validated with Scaffold. The 6 resulting Scaffold files were merged together and only proteins assigned to bacteria were displayed.

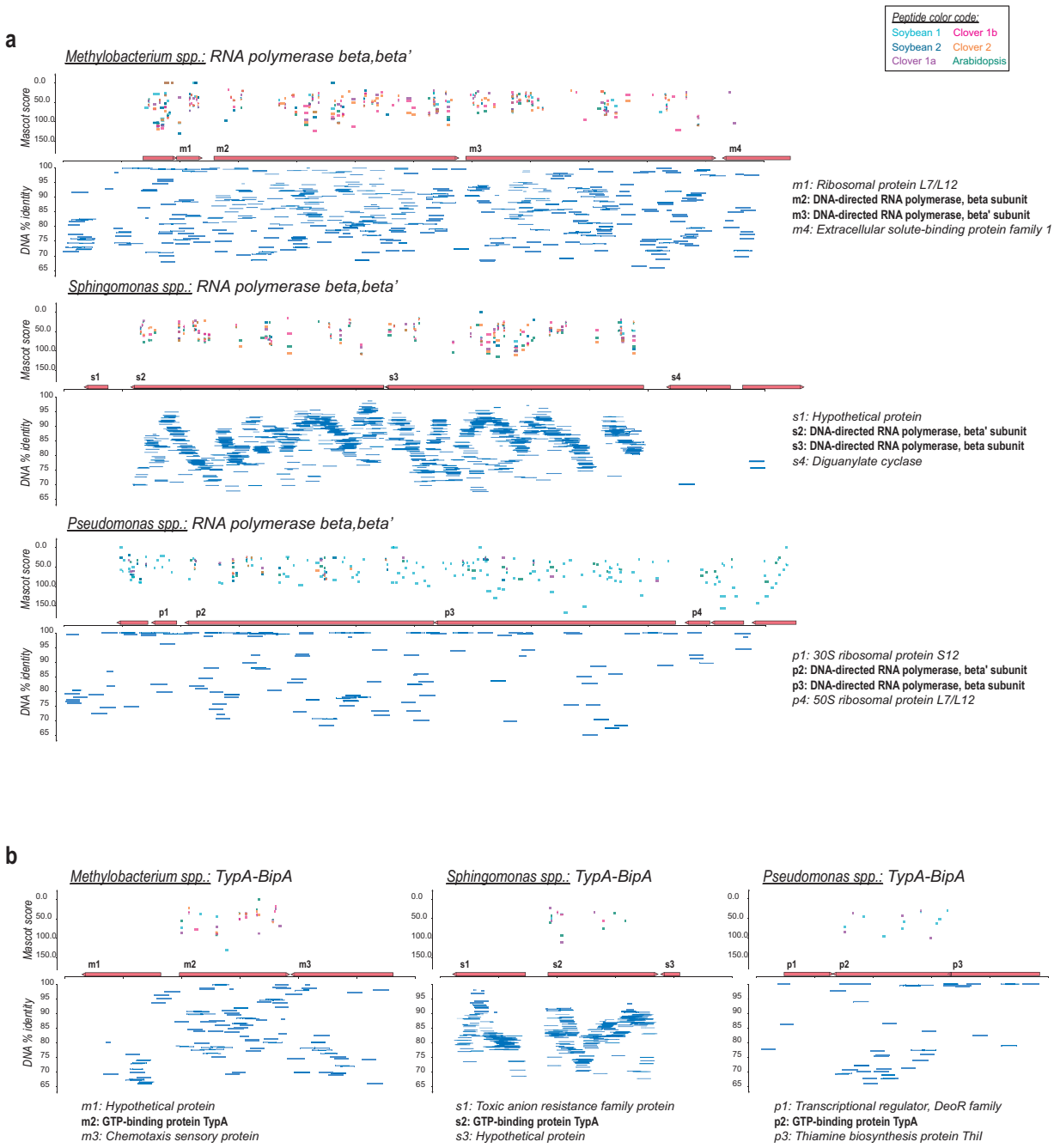
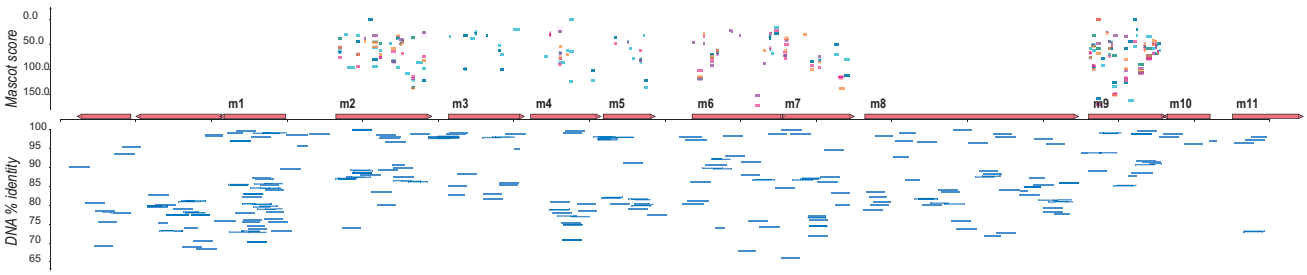


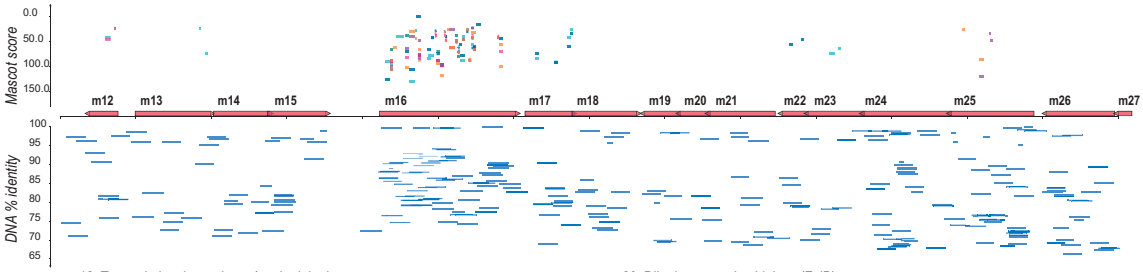
Fig. S4. Diverse fragment recruitment plots for proteins and functional clusters of particular interest. Both metagenomic (*Below x*-axis) and metaproteomic (*Above x*-axis) data are recruited on a reference genome. The color code indicates the sample source of a given peptide. (*a*) Exemplarily for a housekeeping protein, RNA polymerase beta and beta′ within the reference genera *Methylobacterium*, *Sphingomonas*, and *Pseudomonas* is shown. (*b*) GTP-binding protein TypA/BipA within the 3 reference genera is shown. This protein belongs to the ribosome-binding family of GTPases and is widely distributed in bacteria and plants. It has been shown to regulate multiple cell surface and virulence-associated components in *Escherichia coli* (1), and it is required for growth at low temperatures in Enterobacteria and *Sinorhizobium meliloti* (2, 3). (*c*) A major methylotrophy cluster in *Methyloacterium* containing genes for tetrahydromethanopterin-linked 1-carbon dissimilation and the serine cycle involved in 1-carbon assimilation, as well as genes for cofactor biosynthesis of tetrahydromethanopterin and pyrrolo quinoline quinone (4). Please note that *xoxF* (Fig. S5) is also encoded in this genomic region. (*d*) Motility cluster of *Pseudomonas*, covering ≈67kb. (*e*) The gene-annotated Beta-Ig-H3/fasciclin from *M. extorquens* PA1 attracted both DNA reads and peptides.
1. Farris M, Grant A, Richardson TB, O'Connor CD (1998) BipA: a tyrosine-phosphorylated GTPase that mediates interactions between enteropathogenic *Escherichia coli* (EPEC) and epithelial cells. *Mol Microbiol* 28:265–279. 2. Kiss E, Huguet T, Poinsot V, Batut J (2004) The *typA* gene is required for stress adaptation as well as for symbiosis of *Sinorhizobium meliloti* 1021 with certain *Medicago truncatula* lines. *Mol Plant Microbe Interact* 17:235–244. 3. Pfennig PL, Flower AM (2001) BipA is required for growth of *Escherichia coli* K12 at low temperature. *Mol Genet Genomics* 266:313–317. 4. Chistoserdova L, Chen SW, Lapidus A, Lidstrom ME (2003) Methylotrophy in *Methylobacterium extorquens* AM1 from a genomic point of view. *J Bacteriol* 185:2980–2987.

C

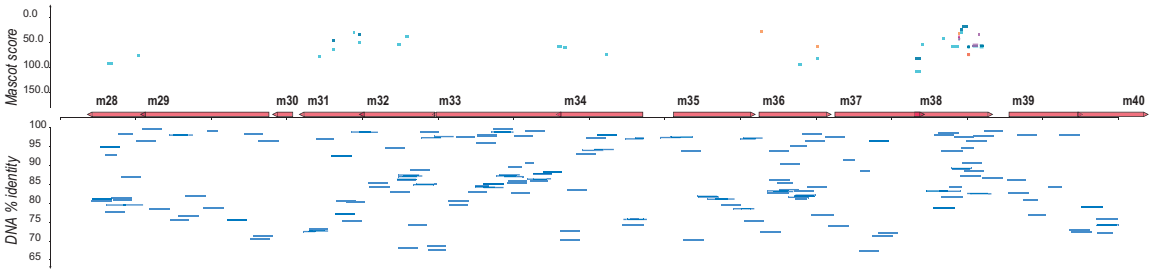
Methylobacterium spp.: Methylotrophy cluster



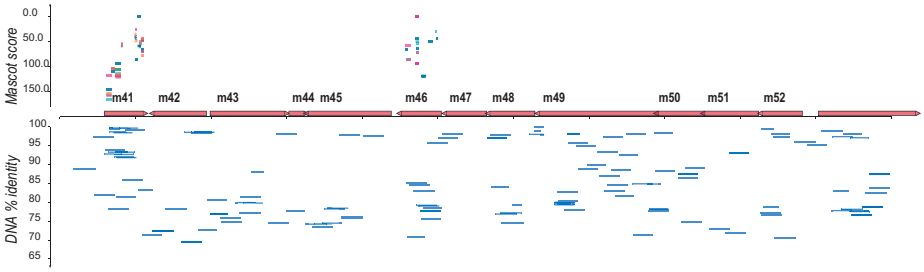
m1: Metallophosphoesterase
m2: Serine-glyoxylate aminotransferase (SGAT)
m3: Hydroxypyruvate reductase (HprA)
m4: Methylene tetrahydromethanopterin/tetrahydrofolate dehydrogenase (MtdA)
m5: Methenyl tetrahydrofolate cyclohydrolase (Fch)
m6: Malate thiokinase, large subunit (MtkA)
m7: Malate thiokinase, small subunit (MtkB)
m8: Phosphoenolpyruvate carboxylase (PpC)
m9: Malyl-CoA lyase (Mcl)
m10: Hypothetical protein
m11: Carboxymethylenebutenolidase



m12: Transcriptional coactivator/pterin dehydratase
m13: Putative ABC transporter periplasmic solute-binding protein
m14: Binding-protein-dependent transport systems inner membrane component
m15: ABC transporter related
m16: Homolog of PQQ-dependent methanol dehydrogenase large subunit (XoxF)
m17: Cytochrome C (XoxG)
m18: Conserved exported of unknown function (XoxJ)
m19: 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase (FolK)
m20: Dihydroneopterin aldolase (FolB)
m21: Dihydropteroate pyrophosphorylase (FolP)
m22: Protein of unknown function
m23: Protein of unknown function
m24: Pyrroloquinoline quinone (PQQ) biosynthesis protein E (PqqE)
m25: Bifunctional coenzyme pyrroloquinoline quinone (PQQ) biosynthesis protein C/D (PqqC/D)
m26: Pyrroloquinoline quinone (PQQ) biosynthesis protein B (PqqB)
m27: Pyrroloquinoline quinone (PQQ) biosynthesis protein A (PqqA)



m28: Response regulator MxbM
m29: Sensor kinase (MxbD)
m30: Protein of unknown function
m31: Formyltransferase/hydrolase complex Fhc subunit C (FhcC)
m32: Formyltransferase/hydrolase complex subunit D (FhcD)
m33: Formyltransferase/hydrolase complex Fhc subunit A (FhcA)
m34: Formyltransferase/hydrolase complex Fhc subunit B (FhcB)
m35: Ribofuranosylaminobenzene 5'-phosphate (RFAP) synthase (H4MPT biosynthesis) (MptG)
m36: Methylene tetrahydromethanopterin dehydrogenase (MtdB)
m37: OrfY
m38: Methenyl tetrahydromethanopterin cyclohydrolase (Mch)
m39: Orf5; m40: Orf7 (cofactor biosynthesis)

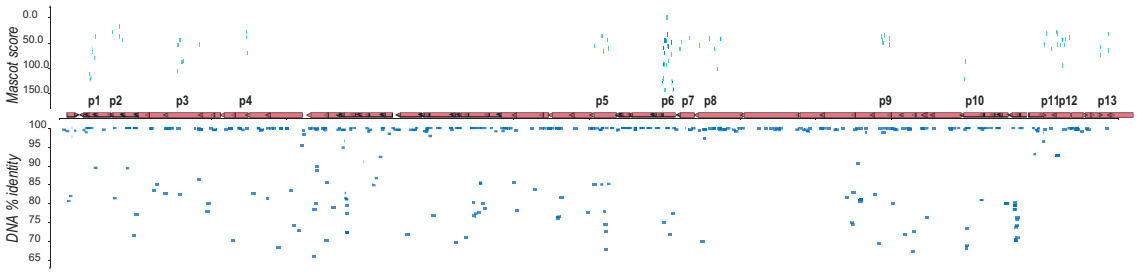


m41: Formaldehyde activating enzyme (Fae)
m42: Orf17 (Cofactor biosynthesis)
m43: Orf9 (Cofactor biosynthesis)
m44: Protein of unknown function
m45: Protein of unknown function
m46: Protein of unknown function
m47: Protein of unknown function
m48: Orf19 (cofactor biosynthesis)
m49: Orf20 (cofactor biosynthesis)
m50: Orf21 (cofactor biosynthesis)
m51: Orf22 (cofactor biosynthesis)
m52: Protein of unknown function

Fig. S4. Continued.

d

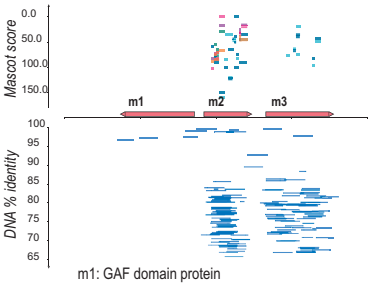
Pseudomonas spp.: Motility cluster *



p1 (71734681): Chemotaxis protein CheW
p2 (71737980): ParA family protein
p3 (71734503): Chemotaxis sensor histidine kinase CheA
p4 (71736626): Flagellar synthesis regulator FleN
p5 (71734449): Flagellar regulator FleQ
p6 (71733403): Flagellin
p7 (71737685): 3-oxoacyl-(acyl-carrier-protein) synthase III, putative
p8 (71735871): Glycosyl transferase, group 2 family protein
p9 (71735315): Flagellar basal body P-ring protein
p10 (71736953): Flagellar hook protein FlgE
p11 (71737331): Chemotaxis protein methyltransferase CheR
p12 (71734029): Chemotaxis protein CheV, putative
p13 (71735292): Hypothetical protein
** here only proteins (RefSeq PID between brackets) detected are reported.*

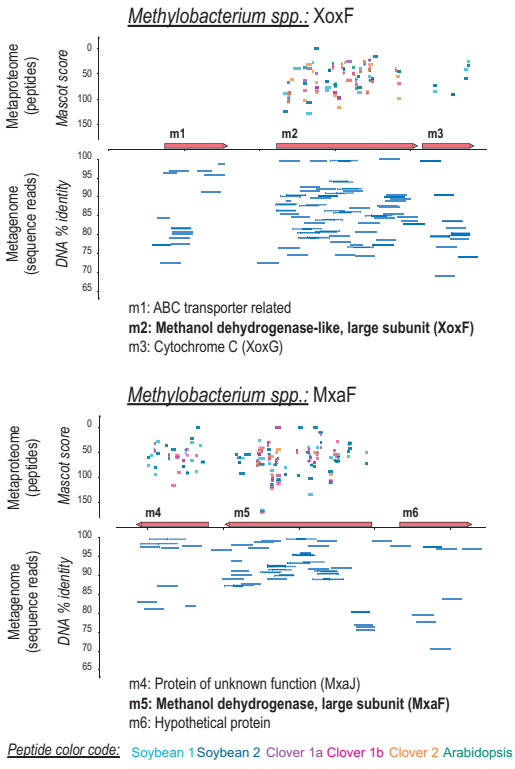
e

Methylobacterium spp.: Fasciclin



m1: GAF domain protein
m2: Beta-Ig-H3/fasciclin
m3: Hypothetical protein

Fig. S4. Continued.



protein name	gene length (a.a.)	# reads identity >90%	read coverage (nt)	# peptides	# spectra
XoxF	601	23	5,843	91	97
MxaF	626	44	10,912	368	407

Relative expression ratio computed for Soybean 2: XoxF/MxaF = 0.5
Relative expression ratio computed for all samples: XoxF/MxaF = 1.4

Fig. S5. Fragment recruitment plots of MxaF and XoxF. Recruitment of both metagenomic (*Below* *x*-axis) and metaproteomic (*Above* *x*-axis) data on a *Methylobacterium* reference genome (*M. extorquens* PA1). Relative expression levels of *mxaF* and *xoxF* were calculated to compare expression strength of both genes. The color code indicates the sample source of a given peptide.

Other Supporting Information Files

Table S1
Table S2
Table S3
Table S4
Table S5
Dataset S1