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.

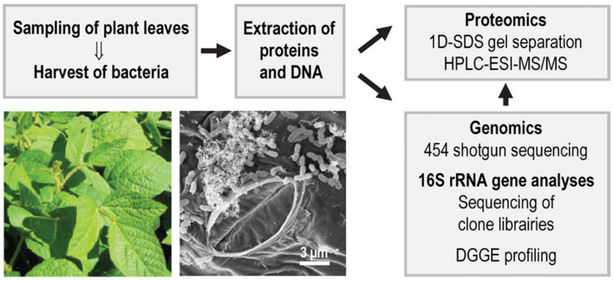

Fig. 1. Experimental strategy applied to characterize the phyllosphere microbiota. All analyses described were conducted from identical pools of cells as starting material. The photograph shows leaves of soybean plants; the electron micrograph shows the surface of an *Arabidopsis* leaf.

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 analyses 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).

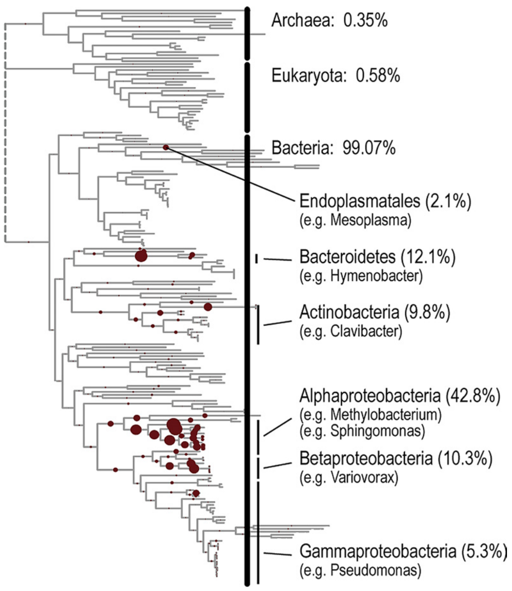

Fig. 2. Taxonomic composition of the bacterial community in the Soybean 2 sample. A phylogenetic tree calculated from informative marker genes of completely sequenced organisms serves as a reference onto which the estimated coverage of the most abundant clades present in the Soybean 2 sample is projected. 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.

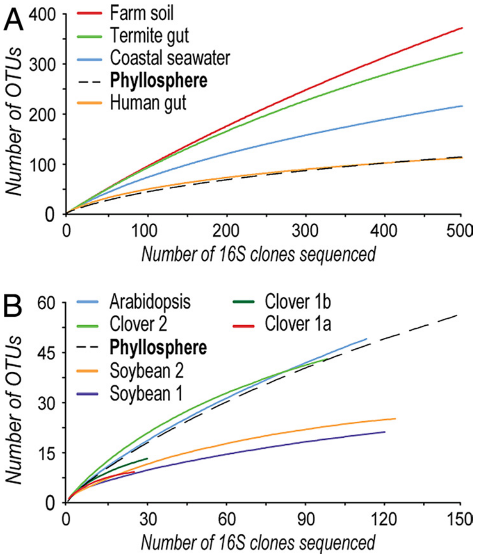

Fig. 3. Rarefaction analysis of 16S rRNA gene-sequence data to estimate microbial diversity based on a cutoff <97% sequence identity for delineation of operational taxonomic units (OTUs). (*A*) Comparison of the composite phyllosphere dataset of this study with published samples covering at least 500 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.

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 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.

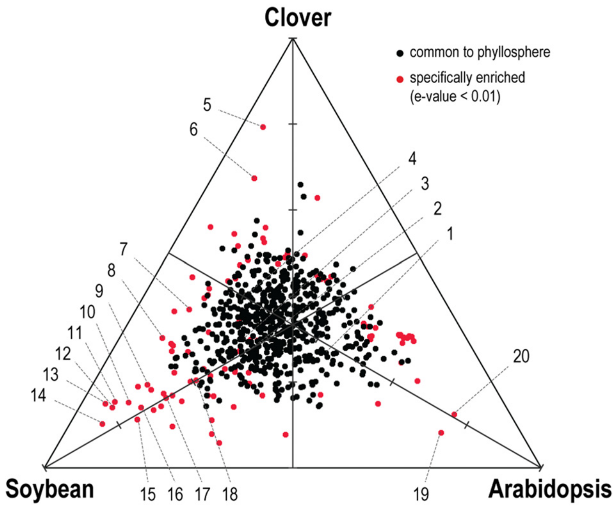

Fig. 4. Conserved and specifically enriched proteome functions (spectral counting of Pfam domains) per host-plant type. Pfam domains drawn close to a vertex are preferentially and specifically found on that respective plant. Selected examples are highlighted and discussed in the text. Examples of common 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).

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 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 × *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.