MPMI Vol. 27, No. 7, 2014, pp. 664–677. http://dx.doi.org/10.1094/MPMI-12-13-0368-R

e-Xtra*

Metabolic Environments and Genomic Features Associated with Pathogenic and Mutualistic Interactions Between Bacteria and Plants

Tatiana V. Karpinets,[1] Byung H. Park,[2] Mustafa H. Syed,[1] Martin G. Klotz,[3] and Edward C. Uberbacher[1]

[1]Biosciences Division, and [2]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, U.S.A.; [3]Department of Biological Sciences, University of North Carolina, Charlotte, NC 28223, U.S.A.

Submitted 13 December 2013. Accepted 12 February 2014.

Genomic characteristics discriminating parasitic and mutualistic relationship of bacterial symbionts with plants are poorly understood. This study comparatively analyzed the genomes of 54 mutualists and pathogens to discover genomic markers associated with the different phenotypes. Using metabolic network models, we predict external environments associated with free-living and symbiotic lifestyles and quantify dependences of symbionts on the host in terms of the consumed metabolites. We show that specific differences between the phenotypes are pronounced at the levels of metabolic enzymes, especially carbohydrate active, and protein functions. Overall, biosynthetic functions are enriched and more diverse in plant mutualists whereas processes and functions involved in degradation and host invasion are enriched and more diverse in pathogens. A distinctive characteristic of plant pathogens is a putative novel secretion system with a circadian rhythm regulator. A specific marker of plant mutualists is the co-residence of genes encoding nitrogenase and ribulose bisphosphate carboxylase/oxygenase (RuBisCO). We predict that RuBisCO is likely used in a putative metabolic pathway to supplement carbon obtained heterotrophically with low-cost assimilation of carbon from CO_2. We validate results of the comparative analysis by predicting correct phenotype, pathogenic or mutualistic, for 20 symbionts in an independent set of 30 pathogens, mutualists, and commensals.

Plant–microbe symbiotic interactions are very diverse. In general terms, these interactions can be mutualistic, when it is beneficial for both organisms; parasitic, when only the microbe benefits at the expense of the host or host damage; and commensal, when one organism benefits and the other is not affected. Mutualistic and commensal bacteria in association with plants are either so-called ectophytes or endophytes, if their location is outside or within plant tissues, respectively. Parasitic plant-associated bacteria are considered pathogens if they damage the host during or after successful colonization, thereby causing symptoms of disease (Newton et al. 2010). Although parasitic and mutualistic interactions represent extreme outcomes in the diversity of plant–microbe interactions (Bulgarelli et al. 2013), the assignment of a bacterial symbiont to pathogen or mutualist cohorts is not straightforward. Parasitic bacteria can change the type of interaction with the host during the host developmental stage (Newton et al. 2010) or may escape classification as pathogens when expression of their pathogenicity and virulence factors needs host-specific triggers (Mi et al. 2012). Successful colonization of a plant host by any symbionts requires the evasion of plant innate immunity; therefore, processes involved in the evasion (Hogenhout et al. 2009; Soto et al. 2009), such as surface polysaccharides and quorum-sensing signals, may be found in genomes of plant symbionts regardless of their mutualistic or pathogenic interaction phenotype. Yet, a subset of plant-pathogenic bacteria causes devastating diseases to their host, such as *Xylella fastidiosa*, causing chlorosis of grapevines (Hopkins and Purcell), or *Xanthomonas campestris* pv. *campestris*, causing bacterial spot on pepper and tomato (Dow and Daniels 1994). In contrast, nonparasitic interactions, such as associations of nitrogen-fixing rhizobia with legumes (Long 1989), are well-known examples of resource-based mutualism, when one type of resource produced by bacteria is traded for another resource produced by plants. Although some of these plant–bacteria interactions have been studied in detail at the genetic, biochemical, and physiological levels and the genomes of several plant-associated bacteria have been sequenced, a large-scale comparative genome analysis searching for genomic markers and characteristics underlying the distinct mutualistic and parasitic interaction phenotypes is not available in the literature to date. Individual molecular functions, metabolic pathways, and biological processes may exist or be enriched in microbial genomes of one versus the other interaction phenotype. Consequently, revealing these genomic differences may have important implications for distinguishing one phenotype from the other, for diagnostics, for controlling plant diseases, and for increasing crop productivity.

In this study we used comparative analysis of sequenced genomes from bacteria in well-known mutualistic or parasitic symbioses with plants to explore genomic features underlying these interaction phenotypes. We calculated enrichments of the genomes with protein functions and enzymes and used metabolic network models to predict substrates consumed by symbionts when they live inside or outside the host. We inferred metabolic pathways and cellular processes that discriminate one phenotype from the other and employed literature, published experimental data, and annotations of 3,062 complete

Corresponding author: T. V. Karpinets; E-mail: karpinetstv@ornl.gov or k2n@ornl.gov

*The *e*-Xtra logo stands for "electronic extra" and indicates that nine supplementary datasets and eight supplementary figures are published online.

bacterial genomes to analyze in more detail two genomic hallmarks of parasitic and mutualistic phenotypes: a putative secretion system in pathogens and a putative ribulose bisphosphate carboxylase/oxygenase (RuBisCO)-based metabolic pathway for carbon assimilation. We validated the predictive power of identified pathogen- and mutualist-specific protein domains using two machine-learning techniques and an independent set of recently sequenced symbionts with validated pathogenic, mutualistic, and commensal phenotypes.

RESULTS AND DISCUSSIONS

Sequenced plant symbionts and related limitations of the analysis.

Bacteria characterized by a symbiotic interaction with plants were manually identified using metadata from the Genomes OnLine Database (GOLD) (Liolios et al. 2010) and the literature. An extensive search of the data revealed only 64 plant symbionts with finished genomes and with well-documented mutualistic (28 organisms) or pathogenic (36 organisms) phenotypes. The majority of these bacteria were classified in the phylum *Proteobacteria*, with representatives in the orders *Acholeplasmatales*, *Burkholderiales*, *Enterobacteriales*, *Pseudomonadales*, *Rhizobiales*, and *Xanthomonadales*. Several of the bacteria were strains and isolates in the same genus or species. Most of such "duplicates", except *Pseudomonas syrinage* DC3000 and *P. syrinage* B728a, were removed, which resulted in an equal number of either phenotype (27 pathogens and 27 mutualists) for further comparative analysis (Fig. 1). Although the cohort of mutualists was dominated by members of the class *Alphaproteobacteria*, members of *Gammaproteobacteria* dominated among pathogens; nevertheless, both cohorts included bacteria from either class. Initial analysis of the metadata revealed that mutualists had, on average, larger genome sizes than pathogens and a higher genomic GC content. A significant difference between groups, however, was not observed when obligate symbiotic *Phytoplasma* spp. were excluded from the comparison. The majority of mutualists were annotated as "nitrogen fixation positive", whereas only one pathogen had this annotation. Most bacteria in both subsets were motile, free-living mesophiles, had varying plant hosts, and were assigned aerobic or facultative anaerobic catabolism.

There are several limitations of the comparative study of the symbionts that must be acknowledged. The number of plant symbionts with completely sequenced genomes is rather small. Although the number of sequenced genomes is quickly increasing due to new sequencing technologies, the majority of these genomes are not finished or strains belong to already represented genera. Additionally, phenotypic information of sequenced plant symbionts is rather incomplete, which precludes an automated selection of organisms for computational analyses.

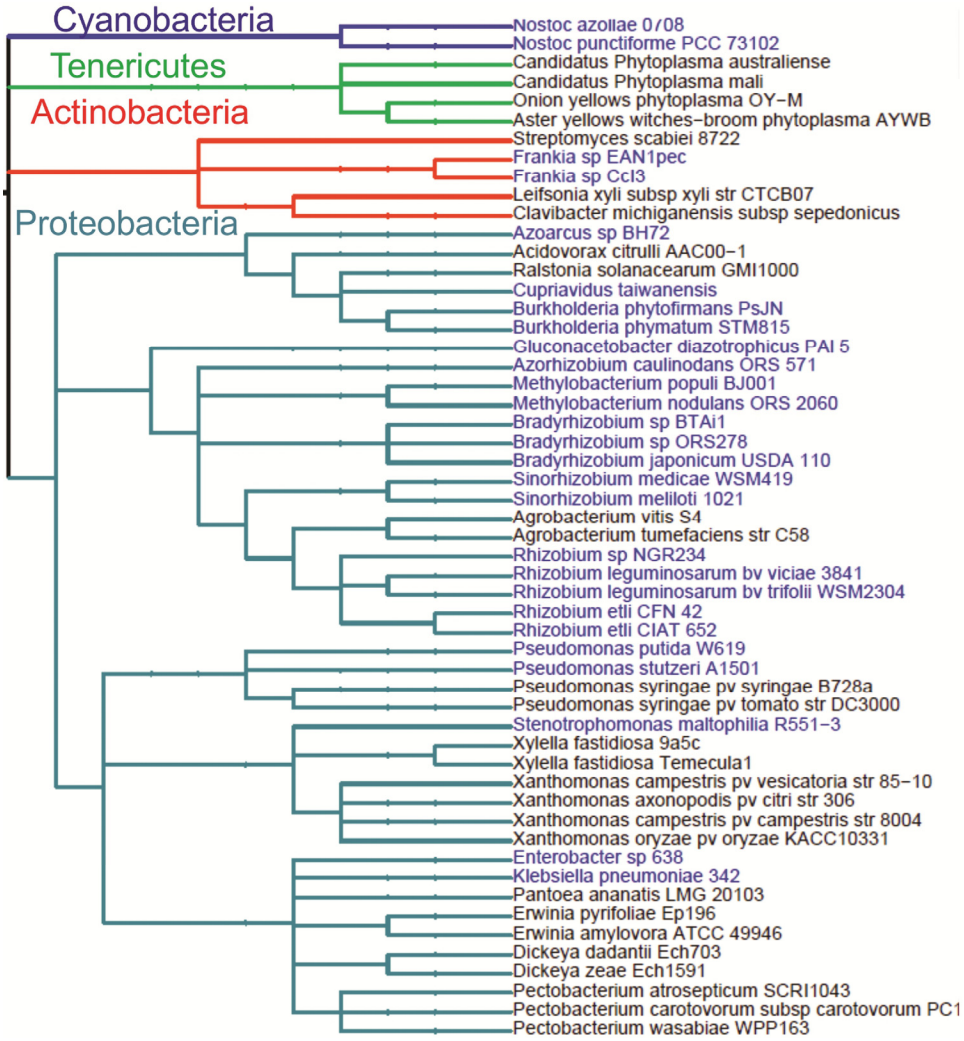


Fig. 1. Taxonomic relationship of plant symbionts selected for a comparative analysis of their genomes. Names of mutualists are given in light blue and names of pathogens are given in black.

The study is focused on two broad phenotypes: parasitic and mutualistic. Both phenotypes, however, are represented by organisms with different mechanisms of interaction with the plant host. Biotrophic mechanisms, for example, can be found in pathogens and mutualists. Moreover, some symbionts are not characterized by a single mechanism; they change the type of interaction during their life cycles. The dynamics may preclude the detection of general processes underlying pathogenesis and mutualism of the symbionts.

Another limitation to this analysis is the uneven phylogenetic distribution of plant symbionts with a mutualistic or pathogenic lifestyle (Fig. 1). Members of classes Gammaproteobacteria and Alphaproteobacteria constitute 66 and 75%, respectively, of the analyzed organisms, with more pathogens (88%) being classified as Gammaproteobacteria and more mutualists (88%) belonging to Alphaproteobacteria. This bias is the outcome of practical interest of researchers regarding genomic sequences of agriculturally important strains, such as root-nodulating nitrogen-fixing mutualists (in order *Rhizobiales*) and devastating necrotrophic plant pathogens (in orders *Enterobacteriales* and *Xanthomonadales*) (Mansfield et al. 2012). Similar associations between phylogeny and lifestyle, however, may exist in a natural setting. A recent study of naturally occurring microbial communities from roots and leaves of *Arabidopsis thaliana* (Bodenhausen et al. 2013) revealed that members of class *Gammaproteobacteria* were the most abundant (34.9%) in epiphytic microbial communities. In contrast, the endophytic communities had only 13.5% of sequences from this class, whereas members of *Alphaproteobacteria* as well as *Betaproteobacteria* were abundant in all endophytic leaf samples, as well as in all root samples. These results indicate that lifestyle and phylogeny may be naturally related. Most mutualists with sequenced genomes were isolated as root or leave endophytes. Most pathogens were isolated from stems and leaves of the infected plants in their epiphytic stage. They often enter plant tissues by wounds and then multiply and spread. More sequenced plant symbionts with a known lifestyle are necessary to explore further the link of the mutualistic or pathogenic nature of symbionts with their phylogeny and habitat.

Host and nonhost biochemical environments of symbionts.

The environment occupied by a microorganism is a major evolutionary force shaping the genome. Extreme genome reduction in obligate pathogens leading to complete dependence on the plant host is well documented (McCutcheon and Moran 2012). However, most plant symbionts can live not only inside the host but also in nonhost environments, which can be rather diverse. Nonhost and host environments of symbionts may contribute by selection to contemporary genomic differences between pathogenic and mutualistic phenotypes. Experimental characterization of the microbial external environments, however, is rather challenging. We used a "reverse-ecology" approach to predict these environments computationally using metabolic network models of the symbiont and of the plant.

First, we found compounds (potential substrates) that are likely acquired by each symbiont exogenously from either the host or nonhost environment (Supplementary Dataset S7A). The employed computational tools (Borenstein et al. 2008; Carr and Borenstein 2012; Kreimer et al. 2012) can predict "potential substrates" from the topology of a metabolic network of the organism. Briefly, a graph-theory algorithm is used to separate compounds into those that can be synthesized from other compounds in the network and those that cannot be synthesized and, therefore, can be considered as potential substrates acquired exogenously to drive bacterial metabolism. We subdivided the potential substrates into two subsets. The first subset, termed "host substrates", includes compounds that symbionts can uptake directly from the plant host. The second subset, termed "nonhost substrates", includes the remaining potential substrates that are not produced by plants and, therefore, must be acquired from sources outside of the host. It has been demonstrated that the predicted potential substrates, which are also referred to in the literature as "seed"-compounds, can provide a proxy of the metabolic environment in habitats of the bacteria (Borenstein et al. 2008). By separating the potential substrates into host substrates and nonhost substrates, we create proxies of metabolic environments for free living and for symbiosis, respectively. Although the computational approach provides a unique opportunity for the comparison of host and nonhost environments of mutualists and pathogens, it has several important limitations. We considered the model plant organism *A. thaliana* as the host for all studied symbionts and used the metabolic model of this plant to predict host and nonhost substrates for each symbiont. We found only small differences between the pools of metabolites predicted from the genome annotation of *A. thaliana* and from other sequenced plant genomes (details below). The finding, however, may be biased by the incomplete knowledge of other plant genomes, because their annotations are typically based on comparison of the predicted genes with *A. thaliana* orthologs. More complete information on specific genomic characteristics of different plant species is critical to make more accurate computational predictions of the host and nonhost environments of plant symbionts. In addition, the reconstruction of a metabolic model for plants, including *A. thaliana*, is significantly more challenging than reconstructions of bacterial metabolism. Although the *A. thaliana* metabolic model (Mintz-Oron et al. 2012) used in the analysis is based on tissue-specific proteomics data of the plant and has been validated by gene-expression studies, only a small subset of predicted metabolites has been verified by a targeted metabolomics analysis. Therefore, we were not able to quantify how many metabolites are under- or overpredicted in the plant host.

Keeping in mind the aforementioned limitations, we compared host and nonhost environments of the symbionts using predicted host substrates and nonhost substrates. We hypothesized that an increased portion of nonhost substrates might be linked to more diverse environments of symbionts when they live outside the plant, and an increased portion of host substrates might be linked to a greater dependence on compounds produced by plants. Therefore, we calculated the ratio of the number of nonhost substrates to the number of host substrates to compare host and nonhost environments and to characterize the ecological independence of the symbiont on the host (Fig. 2). We define this ratio as the host independence (HI) score with greater HI scores indicating a lesser dependence of the symbiont on the plant host and greater ecological flexibility, reflecting a higher probability of survival in diverse biochemical environments outside the host. The calculated HI scores of the symbionts ranged from approximately 0 in obligate pathogens to an average value of 1.5 in the rest (Fig. 2). The latter indicates a diverse nonhost environment of nonobligate plant symbionts, whether pathogens or mutualists. This metabolic versatility may contribute to the dynamics of microbe–host interactions mentioned in the introduction (Newton et al. 2010).

We did not find a significant correlation of the HI score with the size or the GC content of the analyzed symbiont genomes. When excluding obligate pathogens, the ecological versatility and the host independence of studied plant symbionts did not associate with their pathogenic or mutualistic phenotype or their phylogenetic relationships (Fig. 2). We identified several soft-rot plant pathogens, including three *Erwinia* spp., two *Pectobacterium* spp., *Dickeya dadantii*, and *Pantoea ananatis* as symbionts with high HI scores (HI > 1.6). The ability of these

pathogens to occupy a diverse range of environments, including soil and water, is well documented (Coutinho and Venter 2009; Glasner et al. 2008; Kikumoto 1980; van der Wolf et al. 2007) and is consistent with their high HI scores. Most pathogens are actually opportunistic (Pérombelon 2002); they damage plants only when host immunity is compromised. We also identified a mutualistic stem-nodulating bacterium, *Azorhizobium caulinodans*, among symbionts with relatively low HI scores. This symbiont is phenotypically different from other rhizobia and can support only saprophytic growth (Ladha et al. 1989).

Differences in bacterial phenotypes are more pronounced when environments of symbionts are compared in terms of substrates required for their growth outside (nonhost substrates) and inside (host substrates) the host. For both phenotypes, we built a hierarchy of symbionts considering pairwise similarities between profiles of nonhost substrate abundances as well as host substrate abundances (Supplementary Fig. S1A and B). These hierarchies demonstrate a better separation of pathogens and mutualists. However, our analysis of nonhost substrates revealed that caution should be taken in using them for comparison of the growth environments. These substrates are supposed to represent compounds consumed by bacteria but are not produced by the plant host. In some cases, however, these compounds are not predicted because the plant metabolic model is incomplete and fails to predict all metabolites produced by the plant. The compounds identified as host substrates are more accurate because they are inferred from reactions and pathways verified for *Arabidopsis thaliana*, the model plant organism. Many of these reactions are involved in central metabolism of plants and, thus, are common across most plant species. A comparison of host substrates revealed that structural components of plant cell walls and their intermediates, such as cellulose, raffinose, galacturonate, and rhamnulose, are common host substrates for plant pathogens, whereas plant-stress-related compounds such as biotin, stachyose, 5-oxoproline, phenylacetonitrile, cystathionine, and histidinol are common host substrates for plant mutualists. This observation suggests that different metabolic environments shaped genomes of mutualists and pathogens inside the host, which may be the cause of the observed distinct genomic characteristics associated with pertinent phenotypes. These characteristics were a target of our further analysis.

"Constructive" genomic signatures and metabolic diversity in plant mutualists versus "destructive" genomic signatures and diversity of host invasion mechanisms in plant pathogens.

To identify differences in cellular processes and in molecular functions that associate with mutualistic and parasitic phenotypes of plant symbionts, we compared genomes representing the phenotypes in terms of enzymes, pathways, CAZy families, and Pfam domains. All the comparisons revealed similar genomic signatures separating the phenotypes; processes and functions involved in biosynthesis are enriched and more diverse in plant mutualists, whereas processes and functions involved in degradation and host invasion are enriched and diverse in pathogens.

Comparative analysis of metabolic enzymes reveals approximately 1.4 times less the total number of predicted enzymes in the genomes of pathogens than mutualists (P value 0.0007; Supplementary Dataset S5A) and a 10-fold smaller number of enzymes that are significantly ($P < 0.001$) enriched in their genomes (10 unique enzymes in pathogens versus 105 in mutualists). Significant differences remained even when obligate pathogens were excluded from the comparison ($P = 0.0096$). Hierarchical clustering of the enzymes (Fig. 3) clearly demonstrates that genomes of mutualists are essentially more enriched with metabolic activities than genomes of pathogens, with most of the predicted activities participating in biosynthetic and respiration reactions. In the case of pathogens, the enriched metabolic activities are more common across organisms; many of them are represented by hydrolases (EC3) and lyases (EC4) involved in degradation of carbohydrates (Fig. 3).

None of the MetaCyc or Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways is significantly enriched in pathogen genomes, and only one MetaCyc pathway, adenosylcobalamin biosynthesis II (late cobalt incorporation), is enriched in mutualist genomes (false discovery rate [FDR] q value = 0.18 and P value = 0.005). This aerobic route of coenzyme B_{12} (adenosylcobalamin) biosynthesis, including cobalt-containing enzymes, emerged as a characteristic marker for mutualists. The importance of cobalt for *Rhizobium* spp. was experimentally established five decades ago (Burton and Lochhead 1952; Nicholas et al. 1962), which is consistent with our computational predictions.

A comparison of parasitic and mutualistic phenotypes in terms of abundances of predicted CAZy families in their genomes (Supplementary Dataset S6; Supplementary Fig. S2) produced results consistent with the destructive and constructive nature, respectively, of the corresponding phenotypes. The analysis revealed 12 CAZy families that are more abundant in one cohort than the other ($P < 0.01$), with three families of polysaccharide lyases (PL1, PL3, and PL4) and three families of glycoside hydrolases (GH28, GH5, and GH53) being significantly enriched in pathogens. Many of these enzymes are involved in the rotting of plant tissue and in plant cell wall degradation. Six CAZy families enriched in mutualists include one family of GHs (GH102), three families of glycosyl transferases (GT4, GT4, and GT51), and two families of carbohydrate esterases (CE9 and CE4). According to annotations of the families (Cantarel et al. 2009), these enzymes are involved in synthesis of oligo- and polysaccharides rather than in their degradation.

A comparison of Pfam domains found in genomes of the symbionts revealed that plant mutualists are twice as functionally diverse as pathogens. None of the identified Pfam domains, however, is shared by all pathogens or by all mutualists. Many of the Pfam domains are phylum or genus specific and shared by two to three symbionts. Most cellular processes identified as enriched in or specific for plant pathogens (Table 1) represent putative molecular mechanisms utilized by pathogenic bacteria to invade plants. In contrast, none of the clustered mutualist-specific or -enriched protein functions (Table 1) associated with processes of plant invasion. Cellular processes represented by Pfam domains conserved in nine to 25 genomes of mutualists are involved in biosynthetic processes; many of them are well characterized. Importantly, the most common plant pathogen-specific protein functions (Table 1) are represented by a conserved genomic locus with several hypothetical proteins that appear to be a part of a putative secretion system. Furthermore, the most common and specific protein functions encoded in the genomes of mutualists, including nonphotosynthetic symbionts living in root nodules (Table 1), represent RuBisCO (type I), the most abundant enzyme on Earth (Andrews and Lorimer 1987). These two genomic features are the most prominent discriminators of phenotypically different plant symbionts and, thus, they were analyzed in more detail. To this end, we have searched genomes of 3,062 bacteria in the Integrated Microbial Genomes database (IMG) (Mavromatis et al. 2009) for the presence of these discriminating protein functions in the genomes of other pathogens and mutualists. This genome query was complemented by searching the published literature on experimental studies of host–microbe interactions

under in vivo conditions to verify that the genes encoding these discriminating functions are actually expressed during the symbiosis.

A novel putative secretion system as a distinctive characteristic of plant pathogens.

The most common pathogen-specific protein functions (the first cluster in Table 1) are encoded by genes that reside in a predicted operon (Supplementary Fig. S3). These functions are conserved in the genomes of 12 pathogens from six different genera and include nine Pfam domains, with only two of them found in known proteins characterized as components of the type IV secretion system (Audette et al. 2007). These two known secretion-specific protein functions and the co-localization of genes in a conserved operon suggest that the cluster may represent an as-yet-unknown pathogen-specific secretion system and, potentially, a set of unknown virulence factors participating in colonization of the plant host. A search of bacterial genomes in IMG for Pfam domains representing this putative secretion system has identified 52 organisms, 31 (60%) of which are known pathogens, including 17 human pathogens, nine plant pathogens (17%), and also pathogens of insects, rodents, and fish. None of the 52 organisms are characterized as mutualists.

A search of published experimental data in order to confirm expression of this secretion system during pathogenesis did not identify any pertinent functional genomics studies with plant pathogens under in vivo conditions. However, we have identified a study of a human pathogen with the same putative secretion system, *Salmonella enterica* serotype Typhi, the cause of typhoid fever (Sheikh et al. 2011). In this study, an RNA capture and amplification technique followed by microarray hybridization was used to identify expression levels of *S. enterica* Typhi transcripts in the blood of five humans infected with *S. enterica* Typhi compared with expression levels in in vitro cultures. Analysis of the experimental data (Supplementary Fig. S4) revealed that transcripts of five adjacent genes of the putative secretion system—*sty4575* (DUF1525), *sty4576* (DUF1527), and *sty4577* as well as *sty4572* and *sty4573* with a TraC_F_IV domain—were identified in the blood of three to five patients infected with *S. enterica* Typhi. Moreover, *sty4575* (DUF1525) and *sty4576* (DUF1527) transcripts were at significantly higher levels in vivo versus in vitro. Especially high expression (19- to 207-fold increase) was found for *sty4575* (DUF1525). Domain DUF1525 belongs to the clan of thioredoxins. The DUF1525 domain protein STY4575 is small and has a thioredoxin fold; therefore, it potentially regulates activity of host proteins by moving from bacterial cells to plant or animal cells. According to a recent study, the circadian clock is crucial for induction of a pro-inflammatory response to *S. enterica* Typhi infection in mice (Bellet et al. 2013). Therefore, the presence of a virulence factor that may affect the host circadian machinery in the secretion system is not surprising. Well-known proteins with a thioredoxin fold are KaiB and peroxiredoxin involved in circadian regulation by a post-translational mechanism in cyanobacteria (Johnson et al. 2008) and eukary-

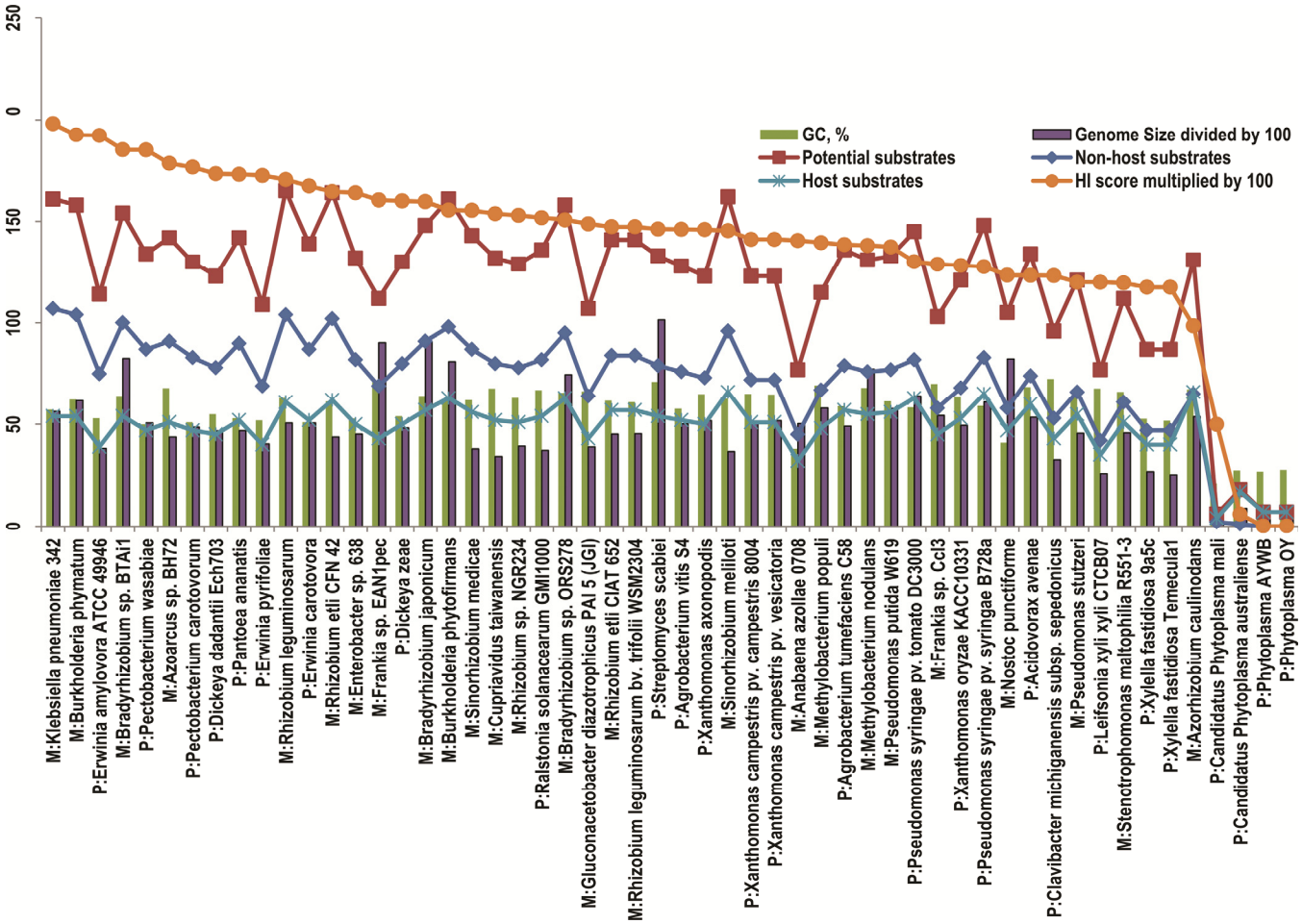


Fig. 2. Characteristics of host and nonhost biochemical environments of symbionts. Number of all predicted potential substrates (rectangles) acquired from the host (stars) and from the nonhost environment (diamonds) in symbionts ordered by the host independence (HI) score (ovals) are shown by lines with corresponding markers. Genome sizes and genomic GC contents are shown for comparison by purple and green columns, respectively.

otes (O'Neill et al. 2011), respectively. Our analysis found *KaiB* in five mutualists in a cluster of genes involved in the bacteriochlorophyll biosynthesis (Table 1) but not in any pathogen. We found 86 genomes with the KaiB protein domain in IMG, of which only *Legionella* and *Microcystis* spp. were pathogenic. Therefore, we propose that the conserved DUF1525 domain protein found in 11 pathogens but not in mutualists represents an alternative to the clock protein KaiB found in mutualists. The high abundance of the DUF1525 protein in blood cells infected with *S. enterica* Typhi suggests an involvement of the protein in colonization of the host.

Nitrogenase and RuBisCO (type I) genes as the most common distinctive characteristic of plant mutualists.

The most common biological process inventory found in 25 mutualists and only in two pathogens is represented by genomic loci encoding nitrogen-fixation proteins (Table 1). Only in genomes of the studied mutualists, however, did the nitrogen fixation genomic loci coexist with genes encoding Pfam domains of RuBisCO, type I. This enzyme provides a key activity for autotrophic CO_2 fixation in the Calvin Benson Bassham (CBB) cycle (Joshi and Tabita 1996), and the importance of the enzyme for CO_2 fixation is well documented for autotrophically growing bacteria, including some plant mutualists such as the cyanobacteria *Nostoc punctiforme* and *Anabaena azollae* (Ekman et al. 2006; Olivares et al. 2013). Autotrophic growth based on the CBB pathway, however, consumes a lot of energy and reductant, and the CBB pathway is predominantly found in phototrophic and chemolithotrophic primary producers. It is intriguing that RuBisCO is also common in genomes of nitrogen-fixing mutualists growing in root nodules and that the enzyme appears to be a specific genomic marker for mutualistic relationships with the plant host in the studied symbionts.

To evaluate associations of RuBisCO with nitrogen fixation and mutualistic phenotypes in other sequenced organisms, we have searched bacterial genomes in IMG (Mavromatis et al. 2009) for protein functions representing RuBisCO (Pfam domains Rubisco_large_N and Rubisco_small) and the nitrogenase (EC 1.18.6.1 and Pfam Oxidored_nitro). We have identified 219 finished genomes in IMG as nitrogenase positive, of which 35 are annotated as genomes of symbionts; however, none of the nitrogenase-positive genomes are annotated as that of a parasite. We also have found 112 genomes in the database that encoded RuBisCO, with approximately half of them (49 genomes) also encoding nitrogenase. Among 112 RuBisCO-positive genomes, not a single one is annotated with a parasitic symbiotic relationship. However, we found 14 known plant symbionts, 13 of which are annotated as mutualistic plant symbionts expressing both nitrogenase and RuBisCO activities. Thus, the IMG dataset suggests that RuBisCO, type I, is not encoded in the genomes of parasites and often co-occurs with nitrogenase-encoding genes in the genomes of plant mutualists.

We further used transcriptomics data from two in vivo studies of *Bradyrhizobium japonicum* (114 microarrays) (Lindemann et al. 2007; Pessi et al. 2007) and an in vivo study of *Sinorhizobium meliloti* (nine microarrays) (Sheikh et al. 2011) to evaluate expression patterns of nitrogenase and RuBisCO genes during a colonization of the plant host. We hypothesized that, if RuBisCO was critical to symbiotic relationships, we would detect RuBisCO gene expression in nodules together with an

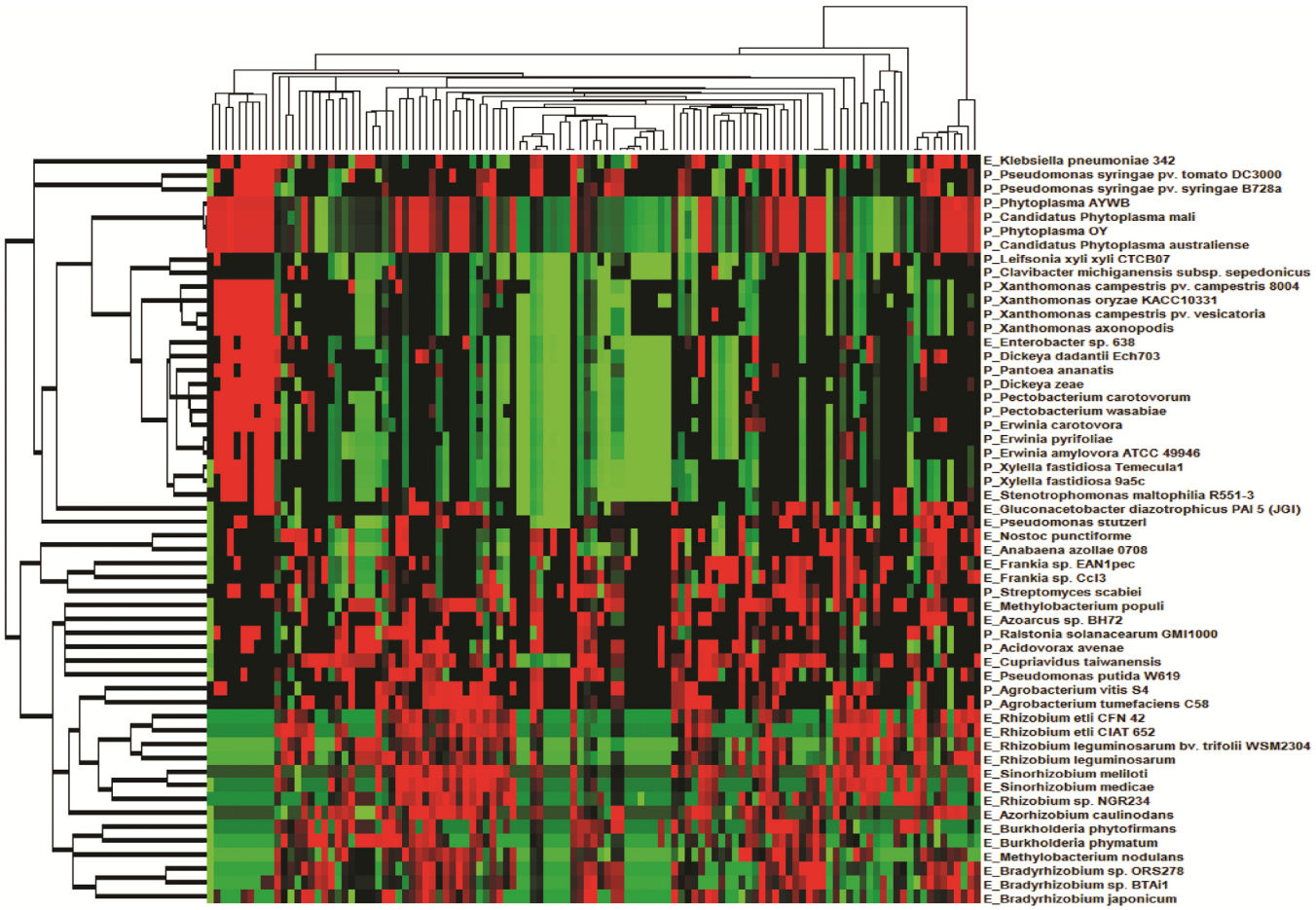


Fig. 3. Hierarchical clustering of number of enzymes predicted in genomes of the symbionts. Only enzymes that are significantly ($P < 0.001$) enriched in one of the groups, pathogens or mutualists, are included in the analysis.

upregulation of nitrogenase gene expressions. We have analyzed transcriptional profiles of the nitrogenase genes (*nifDKENX*) and RuBisCO genes (*rbcL* and *cbbS*) in *B. japonicum* and *Sinorhizobium meliloti* across all microarrays and found similar coexpression patterns of these genes during in vivo growth for both bacteria (Supplementary Fig. S5).

Although the role of RuBisCO in CO_2 assimilation by aerobic autotrophs is well documented, there is presently no literature that implicates RuBisCO in carbon metabolism of nonautotrophic bacteria, including nitrogen-fixing plant symbionts. To understand a potential role of RuBisCO in the metabolism of aerobic, nonautotrophic nitrogen-fixing plant symbionts, the gene neighborhoods of RuBisCO subunit-encoding genes (*cbbS* and *rbcL*) in mutualist genomes were investigated. We found that both subunit genes are co-localized in a conserved seven-member gene operon (Supplementary Fig. S6) that encodes several enzymes of a pathway known in plants as "the RuBisCO shunt" (Schwender et al. 2004). This metabolic pathway is employed by plants during the formation of oil in developing green seed to increase carbon conversion efficiency. Therefore, we propose that plant mutualists may use RuBisCO in their metabolism in a similar context, which would result in less consumption of fixed carbon taken from the host accompanied by maintenance of the fixed-carbon pool at affordable metabolic cost. We considered a set of chemical reactions catalyzed by enzymes of the conserved RuBisCO operon and were able to predict a putative RuBisCO-based pathway for carbon fixation from CO_2 in mutualists (Fig. 4). The pathway, which we propose to call "RuBisCO-amended glycolysis", may provide a straightforward metabolic route to production of pyruvate from D-fructose-6-phosphate, with supplemental CO_2 utilization by RuBisCO at the expense of only one hydrolyzed ATP. The overall stoichiometry of the pathway is similar to the RuBisCO shunt of plants (5D-fructose-6-phosphate = 12acetyl-CoA + $6CO_2$). Both pathways are approximately 20% more efficient in producing acetyl-CoA when compared with Embden-Meyerhof glycolysis (5D-fructose-6-phosphate = 10acetyl-CoA + $6CO_2$) and reduce production of CO_2 by 50%. Importantly, reactions of the pathway require only one additional ATP and, therefore, can be implemented in organoheterotrophic organisms. The mutualist still has to uptake fixed carbon from the plant; however, the uptake will be decreased and dicarboxylic acids such as malate or succinate, which are the main carbon source for nitrogen fixation in culture conditions, can be replaced by glucose or fructose. Activation of the pathway, however, may require a special atmosphere, enriched in CO_2 and depleted in O_2, for expressing RuBisCO, because the enzyme is sensitive to changing CO_2 and O_2 levels (Parry

Table 1. Biological processes represented by differentially abundant or specific Pfam domains in plant pathogens and plant mutualists

Biological process[a]	Species	Pfam domains
Plant pathogens		
Putative secretion system	*Acidovorax avenae, Pseudomonas syringae* pv. *syringae* B728a, *Xanthomonas campestris* pv. *campestris* 8004, *X. campestris* pv. *vesicatoria, Dickeya zeae, X. axonopodis, Erwinia pyrifoliae, Pectobacterium wasabiae, Pseudomonas syringae* pv. *tomato* DC3000, *E. carotovora, X. oryzae* PXO99A, *P. putida* W619	DUF1527,TraC_F_IV, DUF2895, DUF2976, DUF3438, DUF3487, Plasmid_RAQPRD, DUF1525,TraG_N
Group A virulence factors associated with the Hrp protein secretion system	*D. zeae, E. amylovora* ATCC 49946, *E. carotovora E. pyrifoliae, Pectobacterium carotovorum, Pseudomonas syringae* pv. *syringae* B728a, *P. syringae* pv. *tomato* DC3000, *Ralstonia solanacearum* GMI1000	HrpF, HrpE, DspF, AvrE, HrpJ, Hairpins, MxiH
Group B virulence factors associated with the Hrp protein secretion system	*A. avenae, R. solanacearum* GMI1000, *X. axonopodis, X. campestris* pv. *campestris* 8004, *X. campestris* pv. *vesicatoria, X. oryzae* KACC10331	HrpB2, HpaP, HrpB4, HrpB7, HrpB1_HrpK
Plant cell wall degradation	*X. campestris* pv. *campestris 8004, X. campestris* pv. *vesicatoria, X. axonopodis, X. oryzae* KACC10331	Glu_cyclase_2, Cdd1, Glyco_hydro_67C, Glyco_hydro_67M, Glyco_hydro_67N, NAGidase
Phytoplasma-specific virulence factors associated with the cobalt transport system	Phytoplasma AYWB, '*Candidatus* Phytoplasma australiense', Phytoplasma OY	DUF2779, DUF1393, DUF3744, DUF2963
Plant oncogene	*Acidovorax avenae, Agrobacterium tumefaciens* C58	RolB_RolC, Agro_virD5
Unknown	*Xylella fastidiosa* 9a5c, *X. fastidiosa Temecula1*	DUF2669, DUF2913, DUF2815, DUF1566, X_fast-SP_rel, DUF769
Bacillus thuringiensis toxin	*D. dadantii* Ech703, *D. zeae*	Bac_thur_toxin, Neur_chan_LBD, MbeD_MobD
Putative colonization system	*P. syringae pv. syringae* B728a, *P. syringae pv. tomato* DC3000	AvrPtoB-E3_ubiq, AvrPto , DUF1523, DUF2513, DUF3010, DUF3335, WavE, HrpA_pilin

(*continued on next page*)

[a] Processes are named according to functional description of Pfam domains.

[b] Group A and group B effectors are encoded in 13 pathogens by two different conserved operons; each operon is found in a distinct set of pathogens. Although the *hrp* genes are known, the fact that the effectors are found in two non-overlapping sets of many studied pathogens and represent a very discriminative feature of the pathogenic phenotype may be used for classification of plant pathogens and for selection of specific pesticides for each group.

[c] In addition to the clustered domains, a set of 11 glucosyl hydrolases and two cellulose-binding modules (CBM_6 and CBM_3) were encoded in the genomes of 11 pathogens but not in genomes of mutualists. These domains do not constitute a single cluster; instead, different domains were found in different pathogens. The difference likely evolved because each pathogen has a specific host plant and is challenged to degrade a different cell wall structure. The most common functions, found in three to six pathogens, are Glyco_hydro_30, Glyco_hydro_67C, Glyco_hydro_67M, Glyco_hydro_67N, Glyco_hydro_98C, Glyco_hydro_98M, CBM_6, and CBM_3. Both carbohydrate-binding domains (CBM_3 and CBM_6) have the cellulose-binding function but CBM_6 is often found in plant pathogens within complex enzymes in combination with Glyco_hydro_98C and Glyco_hydro_98M domains. CBM_3 is found in combination with the cellulase domain. Domains Glyco_hydro_67N, Glyco_hydro_67M, and Glyco_hydro_67C are also found within one enzyme in four *Xanthomonas* spp. Genomes of four analyzed *Xanthomonas* spp. and of *Streptomyces scabiei* were especially enriched with the glucosyl hydrolases, indicating the importance of cell wall degradation during plant colonization by these species.

[d] Although auxin is produced by some pathogens and, actually, best characterized in the phytopathogenic bacterium *Agrobacterium tumefaciens*, nitrile hydratase is not encoded in any of the investigated pathogen genomes. In *Agrobacterium tumefaciens* C58, auxin is synthesized by a different metabolic route from L-tryptophan by tryptophan 2-monooxygenase (EC 1.13.12.3).

et al. 2013). This requirement is satisfied for endophytic symbionts living in root nodules. The importance of the proposed pathway for the symbiosis is also supported by experiments with mutants of *Bradyrhizobium* sp. strain ORS278. In a study by Bonaldi and associates (2010), mutations in genes encoding the RuBisCO subunit *rbcL* and the ribulose-phosphate 3-epimerase (*rpe*) affected the formation of nodules. In a study by Gourion and associates (Gourion et al. 2011), a *cbbL* deletion mutant strain had an impaired nitrogen-fixation activity. The requirement of CO_2 for growth of legume-nodulating bacteria is also well documented (Lowe and Evans 1962). Additionally, the use of an alternative carbon source was indicated in a study by Bergersen and Turner (1990). These authors reported that only low concentrations of dicarboxylic acids stimulated N_2 fixation but already modestly increased concentrations were inhibitory.

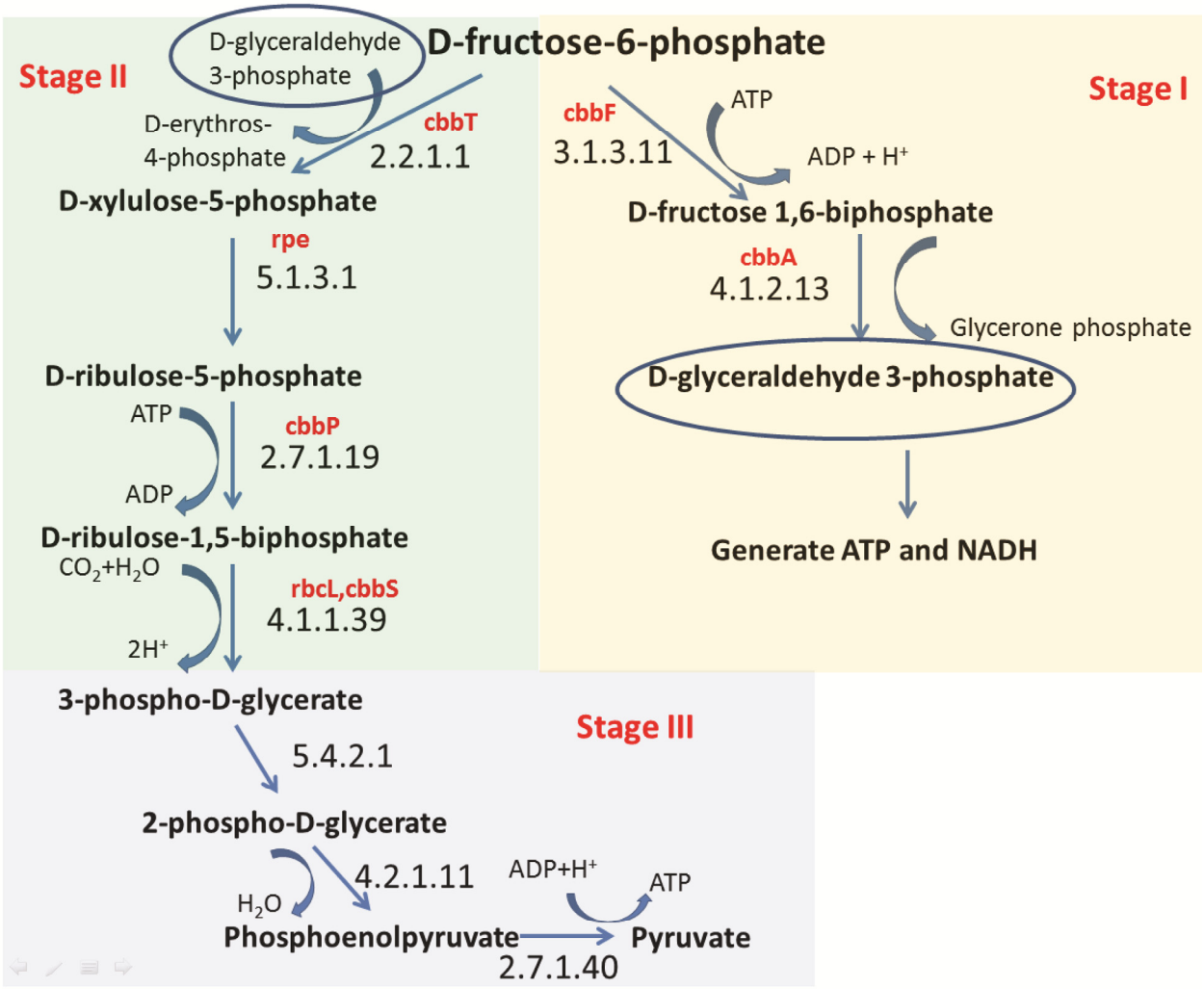
Predicting the plant–microbe interaction phenotype using mutualist and pathogen specific protein family domains.

Specific protein functions encoded in the genomes of pathogens and mutualists and identified by our analysis provided an opportunity to develop a statistical model for predicting the phenotype (mutualistic versus pathogenic) from the genomic content of newly sequenced plant symbionts. We explored the opportunity using two supervised machine-learning techniques, a naive Bayes classifier (John and Langley 1995) and Support Vector Machines (SVM) (Cortes and Vapnik 1995). In both cases, we used abundances of 135 pathogen- and mutualist-specific domains encoded in the genome of 54 studied symbionts to train the model. To validate the predictive power of the models, we created a validation set of 30 plant symbionts with recently sequenced (draft or complete) genomes. The set included 10 pathogens, 10 mutualists, and 10 commensals, which served as the outgroup. Gram-positive bacteria were represented in the set by three pathogens, one mutualist, and one commensal; and mutualists included not only rhizobia but also symbionts of *Datisca glomerata* (a perennial herb)*, Pinus silvestris* (pine), and *Oryza sativa* (rice). Detailed genomic and phenotypic characteristics of the organisms are provided in Supplementary Dataset S1 and the predicted Pfam domains are listed in Supplementary Dataset S8. Both developed models assigned the correct phenotype to pathogens and mutualists in the validation set (Supplementary Dataset S9). The classification of commensal organisms, however, was inconsistent. The naïve Bayes classifier predicted five mutualists and five pathogens from the set of 10 commensals, whereas the SVM model predicted eight mutualists and two pathogens. Nevertheless, the interaction phenotype assignment to *Erwinia tasmaniensis* Et1/99 and *Pantoea agglomerans* 299R as pathogens and the classification of two strains of *Methylobacterium extorquens* and two strains of *Pseudomonas fluorescens* as mutualists was consistent between the models. We believe that the difference

Table 1. (*continued from preceding page*)

Biological process[a]	Species	Pfam domains
Streptomyces scabiei specific proteins	*Streptomyces scabiei*	61 domains (17 DUF)
Plant viruses replication protein	Phytoplasma AYWB	Rep_2, Gemini_AL1, Methyltransf_5, Peptidase_M22
Putative toxin haemolysin-type	*X. fastidiosa* 9a5c (*xfa*), *R. solanacearum* GMI1000, *X. fastidiosa temecula* 1 (*xft*), *A. tumefaciens* C58, *P. syringae* pv. *syringae* B728a, *Xanthomonas axonopodis*	HCBP_related
DUF3396	*A. tumefaciens* C58, *A. vitis* S4, *P. syringae* pv. *syringae* B728a, *P. syringae* pv. *tomato* DC3000, *Pectobacterium wasabiae, Xanthomonas oryzae KACC10331*	DUF3396
Chitin synthase	*A. vitis* S4, *E. carotovora, D. dadantii* Ech703, *Pectobacterium carotovorum, P. wasabiae*	Chitin_synth_1
Plant mutualists		
Nitrogen fixation	Domains found in 23 mutualists (except *Burkholderia phytofirmans* and *Enterobacter* sp. strain 638) and only in two pathogens (*D. dadantii* Ech703 and *E. carotovora*)	NifW, DUF3364, Fer4_NifH, Oxidored_nitro, NifZ, Nitro_FeMo-Co
Nitrogenase biosynthesis	19 species of mutualists have DUF683 and 15 species have DUF269	DUF683, DUF269
IAA biosynthesis[d]	*Azorhizobium caulinodans, B. phytofirmans, Klebsiella pneumoniae* 342, *Nostoc punctiforme, Bradyrhizobium* spp. (*Bradyrhizobium* sp. strain BTAi1*, Bradyrhizobium japonicum, Bradyrhizobium* sp. strain *ORS278*), *Methylobacterium* spp. (*Methylobacterium nodulans, M. populi*), *Rhizobium* spp. (*Rhizobium* sp. strain NGR234, *Rhizobium leguminosarum, R. etli* CFN 42, *R. etli* CIAT 652, *R. leguminosarum* bv. *trifolii* WSM2304), *Sinorhizobium* spp. (*Sinorhizobium meliloti, S. medicae*)	NHase_beta NHase_alpha
CO_2 fixation using Rubisco	*Anabaena azollae* 0708, *Bradyrhizobium* sp. strain BTAi1, *B. japonicum, Burkholderia phymatum, Bradyrhizobium* sp. strain ORS278, *M. nodulans, N. punctiforme, Cupriavidus taiwanensis, S. meliloti, S. medicae*	Rubisco_small Rubisco_large_N
Synthesis of selenocysteine	*Azorhizobium caulinodans, B. phymatum, B. phytofirmans, Enterobacter* sp. strain 638, *K. pneumoniae* 342, *Pseudomonas putida* W619, *P. stutzeri, S. meliloti, Stenotrophomonas maltophilia* R551-3	SelB-wing_2 Se-cys_synth_N SelB-wing_3
Formate-dependent hydrogen production	*Anabaena azollae* 0708, *N. punctiforme, P. stutzeri, R. etli* CFN 42, *R. etli* CIAT 652, *Sinorhizobium meliloti, S. medicae*	FrhB_FdhB_N FrhB_FdhB_C
Oxidative deamination of primary amines to the corresponding aldehydes	*R. leguminosarum* bv. *viciae* 3841, *Burkholderia phytofirmans, Bradyrhizobium* sp. strain BTAi1, *Bradyrhizobium* sp. strain ORS278, *Frankia* sp. strain EAN1pec, *K. pneumoniae* 342, *N. punctiforme*	Cu_amine_oxidN2 Cu_amine_oxid Cu_amine_oxidN3
The phenol monooxygenase system	*Azoarcus* sp. strain BH72, *Bradyrhizobium* sp. strain *BTAi1, Bradyrhizobium japonicum, Bradyrhizobium* sp. strain ORS278, *Cupriavidus taiwanensis, Frankia* sp. strain CcI3	Phenol_Hydrox MmoB_DmpM
Bacteriochlorophyll biosynthesis	*Bradyrhizobium* sp. strain BTAi1, *Bradyrhizobium* sp. strain ORS278, *M. populi, A. azollae* 0708, *N. punctiforme*	DUF3479, Mg-por_mtran_C, PCP_red, PUCC, KaiB, Photo_RC
Unknown	*A. azollae* 0708, *Frankia* sp. strain CcI3, *Frankia* sp. strain EAN1pec, *N. punctiforme*	Saccharop_dh_N, ATP-sulfurylase, SAM_adeno_trans
Exopolysaccharide synthesis and secretion	*R. etli* CFN 42, *R. etli* CIAT 652, *R. leguminosarum* bv. *viciae* 3841, *R. leguminosarum* bv. *trifolii* WSM2304	HTH_CodY, Alg14
Formate utilization	*Burkholderia phymatum, M. nodulans, M. populi*	FTR, DUF447, FTR_C, MCH, Mpt_N DUF556

A



B

Gene	EC	Enzyme Name	Locus	Reaction	Stage
cbbF	3.1.3.11	phosphofructokinase	blr2581	D-fructose 6-phosphate + ATP => D-fructose 1,6-bisphosphate + ADP+H^+	I
cbbA	4.1.2.13	fructose-1,6-bisphosphate aldolase	blr2584	D-fructose 1,6-bisphosphate => glycerone phosphate + D-glyceraldehyde 3-phosphate	I
cbbT	2.2.1.1	transketolase	blr2583	D-glyceraldehyde-3-phosphate + D-fructose-6-phosphate = D-erythros-4-phosphate + D-xylulose-5-phosphate	II
rpe	5.1.3.1	ribulose-phosphate 3-epimerase	blr2588	D-xylulose 5-phosphate => D-ribulose 5-phosphate	II
cbbP	2.7.1.19	phosphoribulokinase	blr2582	ATP + D-ribulose 5-phosphate => ADP + D-ribulose 1,5-bisphosphate	II
rbcL, cbbS	4.1.1.39	RuBisCO	blr2585, blr2586	D-ribulose 1,5-bisphosphate + CO_2 + HO_2 => 3-phospho-D-glycerate + $2H^+$	II

C

Fig. 4. Computationally predicted ribulose bisphosphate carboxylase/oxygenase-amended glycolysis in mutualists. **A,** The metabolic pathway was reconstructed from **B,** enzymes encoded in the genome of mutualists by **C,** a conserved operon. Reactions catalyzed by enzymes of the operon implement stage I and stage II of the metabolic pathway. The input compound of both stages is D-fructose-6-phosphate. In stage I of the pathway, two genes (*cbbF* and *cbbA*) of the operon are used to produce D-glyceraldehyde-3-phosphate that is further used as an input compound in stage II. During the second stage, the rest of the five genes (*cbbT, rpe, cbbP, rbsL*, and *cbbS*) encoded by the operon synthesize D-xylulose-5-phosphate and convert it into D-ribulose 1,5-bisphosphate to produce 3-phospho-D-glycerate. In stage III of the pathway, 3-phospho-D-glycerate can be converted into pyruvate via the later steps of Emden-Meyerhof glycolysis. Three enzymes involved in stage III are encoded separately in the genome of *Bradyrhizobium japonicum*. Pyruvate can be further metabolized to fatty acids or energy through acetyl-CoA.

in predictions between the models for commensal microbes is not surprising because of their phenotypic variance. Similar to pathogens, commensal bacteria with sequenced genomes are usually ectophytes that occupy plant surfaces in the rhizosphere or phylosphere; however, they do not cause symptoms of disease. Instead, they are used as biocontrol agents to protect plant hosts from pathogenic bacteria. Some commensals, however, are also reported as plant mutualists or endophytes (Rosenblueth and Martinez-Romero 2006). The ambiguity in the interaction phenotype of commensals might actually originate from the fact that they were considered commensals because they could not be directly linked to a plant disease or a beneficial effect on the host. Such outcomes are difficult to establish experimentally.

To explore associations of commensal organisms with mutualists and pathogens in more detail, we used two unsupervised computational techniques: principal coordinates analysis (PCoA) (Gower 1966) and association network (ANET) (Karpinets et al. 2012). We combined the organisms used for training and validation into one set and employed the computational methods to visualize distances (PCoA) and associations (ANET) between 84 plant symbionts in terms of pathogen- and mutualist-specific Pfam domains. In the PCoA analysis, the domain abundances were used to build a dissimilarity matrix and them to calculate distances between each pair of the symbionts and to display the strains in the two-dimensional space (PCoA plot; Supplementary Fig. S8). In the ANET technique, we considered co-occurrences of symbionts for each Pfam domain and then built a network from similar co-occurrence profiles for pairs of the strains (Fig. 5). Both methods (Fig. 5) support a transitional position of the commensal interaction phenotype (between pathogenic and mutualistic) and the conclusion about potentially incorrect assignments of commensal bacteria as pathogens or mutualists in terms of the identified pathogen- or mutualist-specific Pfam domains. The association network analysis (Fig. 5) also revealed that most commensal bacteria, in spite of their similarity to pathogens and mutualists, tend to cluster together and can be assigned to

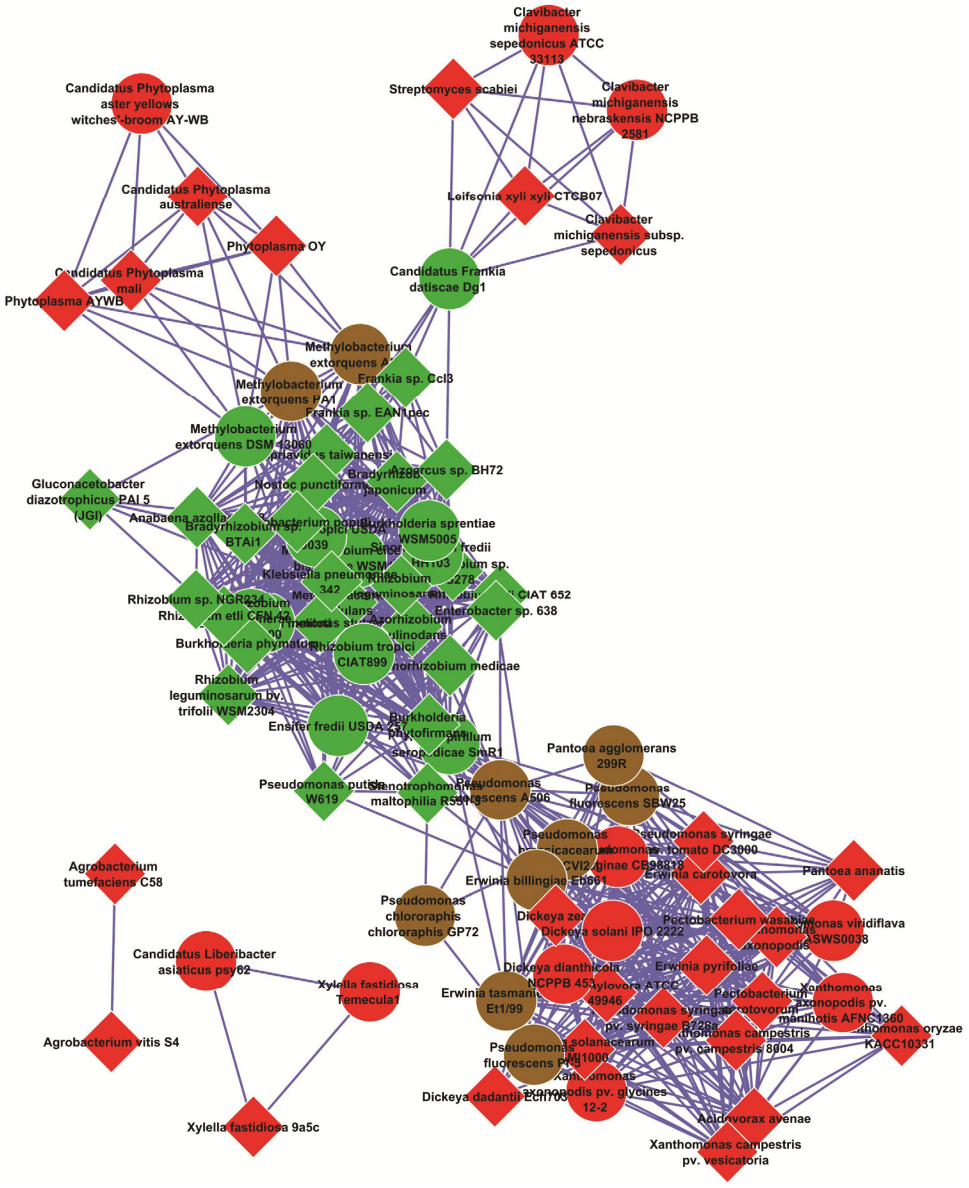


Fig. 5. Association network of 84 plant symbionts used in the study. Each node in the network represents a symbiont, and an edge between nodes indicates similar co-occurrence profiles of the symbionts in terms of the mutualist- and pathogen-specific Pfam domains. The node color indicates pathogenic (red), mutualistic (green), or commensal (brown) phenotype of the symbiont. The node shape shows whether the genome annotation of the symbiont was used for training of the classification models (diamond) or for their validation (circles).

a distinct cohort computationally. This suggests that it is crucial to include commensal organisms in the training set of classification models in order to make more accurate predictions of the interaction phenotypes of plant symbionts.

In conclusion, this study provided a comparative characterization of pathogenic and mutualistic interactions between bacteria and plants based on a diverse set of computational annotations and predictions inferred from genomes of sequenced symbionts. The comparison extends beyond well-known genomic annotations, such as protein family domains, enzymes, and pathways, but also includes characteristics inferred from metabolic network models of the interacting organisms, such as metabolites consumed inside and outside the host and the level of metabolic independence of symbionts on the host. The use of metabolic models for interaction phenotype predictions may be especially valuable for uncultured sequenced symbionts whose ecological characteristics cannot be measured experimentally.

The results of our comparative analysis suggest that nonobligatory symbionts of both phenotypes might be similar in their ability to occupy diverse metabolic environments outside the host; nevertheless, both phenotypes can be distinguished when considering the metabolite environment and specific compounds utilized by symbionts. Each interaction phenotype is characterized by a set of functional markers in the genome that encode enzymes, especially carbohydrate-active enzymes (CAZymes), and other protein functions. Genomes of mutualists have an increased metabolic diversity and an increased number of enzymes involved in synthesis of oligo- and polysaccharides in biosynthetic and respiratory reactions. Genomes of pathogens are enriched with functions involved in host invasion, with catabolic enzymes and with CAZymes involved in degradation of plant cell walls. Using two machine-learning techniques, we demonstrated that pathogen- and mutualist-specific protein functions identified in the analysis can provide a foundation for development of a predictive computational tool for classification of plant symbionts. However, the inclusion of outgroup organisms such as commensals, saprophytes, free-living nitrogen-fixing bacteria, and a greater number of complete genomes representing these phenotypes are important conditions for increasing specificity and sensitivity of the prediction.

Two genomic characteristics are found to be most discriminating between pathogenic and mutualistic phenotypes of the symbionts and reveal some interesting insights into the nature of mutualism and parasitism. A putative unknown secretion system expressed during pathogenesis and associated with a circadian rhythm regulator appears to be the most discriminating genomic feature of plant-pathogenic bacteria; therefore, this system may be a target of antibacterial drugs or pesticides. A distinctive genomic feature of nitrogen-fixing plant mutualists is the operon encoding enzymes of a plant pathway known as the RuBisCO shunt. To this point, the bacterial RuBisCO-amended glycolysis shunt is a computational prediction. Once confirmed experimentally, this pathway may enhance our understanding of how two critical resources, nitrogen and carbon, are traded between bacterial mutualists and their plant hosts. Further experimental evidence, however, is needed for the predicted RuBisCO-containing pathway and for the putative secretion system.

MATERIALS AND METHODS

Protein family domains.

Protein sequences of the genomes selected for analysis were downloaded from the National Center for Biotechnology Information (NCBI) in January 2011 and annotated by protein family domains using Pfamscan and hidden Markov models from the Pfam protein families database (v.24) (Finn et al. 2008). A table of Pfam enrichment profiles was generated, with each row representing a domain and each column representing an organism, by counting the number of each Pfam domain for each organism (Supplementary Dataset S2). The table was analyzed using a computational framework presented in Supplementary Fig. S7. Two statistical tests were used to characterize specificity of each Pfam domain in the table for each subset of organisms and whether the domain was differentially abundant in one subset versus the other. Differentially abundant Pfam domains in each subset were selected using the nonparametric *t* test in combination with the Fisher's exact test, as described (White et al. 2009). Using this method, we estimated the FDR (q value) for each Pfam domain or the proportion of false positives within the set of predicted differentially abundant domains. We identified groups of protein functions that likely represent pathways or contribute to biological processes specific for each phenotype using a previously described approach (Karpinets et al. 2010) by calculating enrichment profiles of each Pfam domain among the genomes, correlating each pair of the profiles, and clustering significantly correlated profiles using the Markov clustering algorithm (MCL) (Enright et al. 2002). Evidence for synteny of identified clusters of Pfam domains was established by selecting the most confident groups of relating protein functions. Statistical tests were further used to characterize each cluster as specific for either pathogens or mutualists. Spearman correlation coefficient (R) was used to calculate similarity between each pair of Pfam profiles. A threshold value of the coefficient ($R = 0.90$) was used to find significantly correlated pairs of Pfam profiles among organisms. The correlated profiles were clustered using MCL with the inflation value 1.8. Results of this analysis, including all identified clusters of Pfam domains, their names, synteny scores, domain abundances in the studied organisms, and results of statistical tests for each domain, are shown in Supplementary Datasets S3 and S4. A manual curation of the information collected for each cluster was used to compile a table of cellular processes that are specific for plant pathogens or mutualists (Table 1).

Enzymes and metabolic pathways.

Enzyme annotations (EC numbers) of the genomes were downloaded from KEGG (Kanehisa 2002) and combined into one table to create profiles with the number of each enzyme across all organisms. Statistical significance for enrichment of each enzyme in the group (pathogens or mutualists) was estimated using statistical tests, as described in the previous sections. Only enzymes that were statistically significantly enriched in one of the groups ($P < 0.001$) were selected for hierarchical clustering, which was implemented using Gene Cluster v.3 (de Hoon et al. 2004) and the Spearman correlation coefficient as a similarity metric and centroid linkage (distance) between the two clusters. The pathway enrichment analysis of pathogens versus mutualists was implemented using gene set enrichment analysis (Subramanian et al. 2005). We compared the phenotypes in terms of their enrichment with MetaCyc (Krieger et al. 2004) and KEGG (Kanehisa 2002) pathway annotations downloaded from websites of the respective databases. We used a signal-to-noise ratio to score the difference between phenotypes for each EC number and to rank the numbers to calculate the enrichment score, P value, and q value for each pathway predicted by KEGG and MetaCyc.

CAZymes

We used the CAZymes Analysis Toolkit (Park et al. 2010) to predict CAZy families in the genome of each symbiont. Then, we calculated the number of predicted CAZymes in each family and used the resulting table to select differentially abundant CAZy families in each subset of organisms by calculating P

and q values (White et al. 2009). The selected families were used for hierarchical clustering of symbionts. Different thresholds for q and P values as well as for similarity measures were tested to get the best separation of mutualists and pathogens.

Host and nonhost biochemical environments.

We used NetCmpt (a network-based tool for calculating the metabolic competition) (Borenstein et al. 2008; Carr and Borenstein 2012; Kreimer et al. 2012) to infer a set of exogenously acquired compounds, also referred to as the potential substrates, from the list of reactions predicted for each symbiont. The predicted potential substrates were then compared with the compounds available from the plant host using a global metabolic reconstruction of model plant organism *Arabidopsis thaliana* (Mintz-Oron et al. 2012). Those compounds in the list of symbiont-potential substrates that overlapped with the plant compounds were considered as the host substrates of the symbiont and the remaining compounds were considered as the nonhost substrates. R scripts were used for all calculations and are available upon request. The hierarchical clustering was used to cluster symbionts and their nonhost seed compounds, as described above.

To justify the use of *A. thaliana* as the plant host for predicting host and nonhost substrates of the plant symbionts, we compared metabolites inferred from the *A. thaliana* genome with genomes of five different plants: *Populus trichocarpa* (poplar) (Tuskan et al. 2006), *O. sativa* (rice) (Ouyang et al. 2007), *Zea mays* (corn) (Schnable et al. 2009), *Vitis vinifera* (grape) (Jaillon et al. 2007), and *Solanum lycopersicum* (tomato) (Sato et al. 2012). These plants represent hosts of six mutualists and seven pathogens in the studied symbionts. We downloaded annotations of the genomes from Phytozome V9 and used the KEGG orthology annotation of proteins from each genome to extract all associated reactions from the KEGG database (Kanehisa 2002). Furthermore, we linked the reactions to metabolites in the KEGG database to identify a unique set of metabolites for each plant and compared this set with metabolites found in *A. thaliana*. The analysis shows that the number of predicted metabolites in the genomes varied from 1,483 in *Z. mays* to 1,553 in *A. thaliana*. The number of plant-specific metabolites (that are not found in *A. thaliana*), however, was rather small, ranging from 15 in *Solanum lycopersicum* to one in *Z. mays* and *O. sativa*. None of the plant-specific metabolites is found in the list of potential substrates predicted from the genome of the studied symbionts. This result reveals limitations for discrimination of different plant hosts in terms of the predicted metabolites based on the present knowledge of plant metabolism.

Transcriptomics data.

Five previously published transcriptomics datasets (Capela et al. 2006; Lindemann et al. 2007; Pessi et al. 2007; Sheikh et al. 2011) were used to confirm the in vivo activity (during the symbiosis) of the two most discriminative genomic features of pathogenic and mutualistic phenotypes: genes of a putative secretion system and genes encoding nitrogenase and RuBisCO. Coexpression of nitrogenase- and RuBisCO-encoding genes was confirmed during the in vivo growth of *B. japonicum* and *Sinorhizobium meliloti*. Datasets of microarray experiments from studies of *B. japonicum* (Lindemann et al. 2007; Pessi et al. 2007) were downloaded from the Gene Expression Omnibus database (accession numbers GDS3120, GDS3121, and GDS3122). The datasets were combined into one table, normalized and $\log^2$ transformed. The resulting expression values of RuBisCO and nitrogenase genes were used to explore coexpression patterns of the enzymes. Data from a microarray study with *Sinorhizobium meliloti* were downloaded from the online supplement of Capela and associates (2006). The expression of the unknown putative secretion system was confirmed during in vivo growth of *Salmonella enterica* serotype Typhi. Experimental data available in the online supplement of a published study by Sheikh and associates (2011) were used to select expression levels of genes of the putative secretion system.

Computational prediction of the interaction phenotype.

We used the R package e1071 to build two classification models, a naive Bayes classifier and SVM, based on the training set of 135 Pfam protein domain abundances in genomes of the studied symbionts. Only pathogen- and mutualist-specific Pfam domains shared by at least two pathogens or mutualists from different genera were selected to train the models. To validate the predictive power of the models, we searched the literature and databases (IMG, GOLD, and NCBI) and identified 30 recently sequenced plant symbiotic bacteria representing three different interaction phenotypes with the plant host, 10 pathogens, 10 mutualists (endophytes), and 10 commensals. All annotations of the organisms, including Pfam annotation, were downloaded from IMG. The R environment was used to predict the interaction phenotypes for bacteria in the validation set using the classifiers. The R packages ape and vegan were used to implement the PCoA with the default parameters. The ANET program developed by us and described before (Karpinets et al. 2012) was employed to construct the association network using Spearman correlation (Rs) as the similarity measure and Rs = 30 as the threshold. The Cytoscape V 2.8.1 (Smoot et al. 2011) was used to visualize the network.

ACKNOWLEDGMENTS

This research was sponsored by the Plant Microbe Interface Project of the Genomic Science Program, U.S. Department of Energy (DOE), Office of Science, Biological, and Environmental Research. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. DOE under contract DE-AC05-00OR22725. The work of M. H. Syed to adapt certain tools was supported by The BioEnergy Science Center (BESC). BESC is a U.S. DOE Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. M. G. Klotz was supported by incentive funds from the University of North Carolina at Charlotte. We thank anonymous reviewers of the manuscript for thoughtful suggestions and comments on the study.

LITERATURE CITED

Andrews, T. J., and Lorimer, G. H. 1987. Rubisco: Structure, mechanisms, and prospects for improvement. Pages 77-83 in: The Biochemistry of Plants: A Comprehensive Treatise. Photosynthesis: Physiology and Metabolism. R. C. Leegood, T. D. Sharkey, and S. von Caemmerer, eds. Kluwer Academic Publilshers, New York.

Audette, G. F., Manchak, J., Beatty, P., Klimke, W. A., and Frost, L. S. 2007. Entry exclusion in F-like plasmids requires intact TraG in the donor that recognizes its cognate TraS in the recipient. Microbiology 153:442-451.

Bellet, M. M., Deriu, E., Liu, J. Z., Grimaldi, B., Blaschitz, C., Zeller, M., Edwards, R. A., Sahar, S., Dandekar, S., Baldi, P., George, M. D., Raffatellu, M., and Sassone-Corsi, P. 2013. Circadian clock regulates the host response to *Salmonella*. Proc. Natl. Acad. Sci. U.S.A. 110:9897-9902.

Bergersen, F. J., and Turner, G. L. 1990. Bacteroids from soybean root-nodules—Respiration and N2-fixation in flow-chamber reactions with oxyleghemoglobin. Proc. R. Soc. Lond. Ser. B Biol. Sci. 238:295-320.

Bodenhausen, N., Horton, M. W., and Bergelson, J. 2013. Bacterial communities associated with the leaves and the roots of *Arabidopsis thaliana*. PLoS One 8:e56329. Published online.

Bonaldi, K., Gourion, B., Fardoux, J., Hannibal, L., Cartieaux, F., Boursot, M., Vallenet, D., Chaintreuil, C., Prin, Y., Nouwen, N., and Giraud, E. 2010. Large-scale transposon mutagenesis of photosynthetic *Bradyrhizobium* sp. strain ORS278 reveals new genetic loci putatively important for nod-independent symbiosis with *Aeschynomene indica*. Mol. Plant-Microbe Interact. 23:760-770.

Borenstein, E., Kupiec, M., Feldman, M. W., and Ruppin, E. 2008. Large-scale reconstruction and phylogenetic analysis of metabolic environ-

ments. Proc. Natl. Acad. Sci. U.S.A. 105:14482-14487.
Bulgarelli, D., Schlaeppi, K., Spaepen, S., van Themaat, E. V. L., and Schulze-Lefert, P. 2013. Structure and functions of the bacterial microbiota of plants. Annu. Rev. Plant Biol. 64:807-838.
Burton, M. O., and Lochhead, A. G. 1952. Production of vitamin B-12 by *Rhizobium* species. Can. J. Bot. Rev. Can. Bot. 30:521-524.
Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. 2009. The carbohydrate-active enzymes database (CAZy): An expert resource for glycogenomics. Nucleic Acids Res. 37:D233-D238.
Capela, D., Filipe, C., Bobik, C., Batut, J., and Bruand, C. 2006. *Sinorhizobium meliloti* differentiation during symbiosis with alfalfa: A transcriptomic dissection. Mol. Plant-Microbe Interact. 19:363-372.
Carr, R., and Borenstein, E. 2012. NetSeed: A network-based reverse-ecology tool for calculating the metabolic interface of an organism with its environment. Bioinformatics 28:734-735.
Cortes, C., and Vapnik, V. 1995. Support-vector networks. Machine Learning 20:273-297.
Coutinho, T. A., and Venter, S. N. 2009. *Pantoea ananatis*: An unconventional plant pathogen. Mol. Plant Pathol. 10:325-335.
de Hoon, M. J., Imoto, S., Nolan, J., and Miyano, S. 2004. Open source clustering software. Bioinformatics 20:1453-1454.
Dow, J. M., and Daniels, M. J. 1994. Pathogenicity determinants and global regulation of pathogenicity of *Xanthomonas campestris* pv. *campestris*. Curr. Top. Microbiol. Immunol. 192:29-41.
Ekman, M., Tollback, P., Klint, J., and Bergman, B. 2006. Protein expression profiles in an endosymbiotic cyanobacterium revealed by a proteomic approach. Mol. Plant-Microbe Interact. 19:1251-1261.
Enright, A. J., Van Dongen, S., and Ouzounis, C. A. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575-1584.
Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L., and Bateman, A. 2008. The Pfam protein families database. Nucleic Acids Res. 36:D281-D288.
Glasner, J. D., Marquez-Villavicencio, M. , Kim, H. S., Jahn, C. E., Ma, B., Biehl, B. S., Rissman, A. I., Mole, B., Yi, X., Yang, C. H., Dangl, J. L., Grant, S. R., Perna, N. T., and Charkowski, A. O. 2008. Niche-specificity and the variable fraction of the *Pectobacterium* pan-genome. Mol. Plant-Microbe Interact. 21:1549-1560.
Gourion, B., Delmotte, N., Bonaldi, K., Nouwen, N., Vorholt, J. A., and Giraud, E. 2011. Bacterial RuBisCO is required for efficient *Bradyrhizobium/Aeschynomene* symbiosis. PLoS One 6. Published online.
Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325-338.
Hogenhout, S. A., Van der Hoorn, R. A., Terauchi, R., and Kamoun, S. 2009. Emerging concepts in effector biology of plant-associated organisms. Mol. Plant-Microbe Interact. 22:115-122.
Hopkins, D., and A. Purcell. 2002. Xylella fastidiosa: Cause of Pierce's disease of grapevine and other emergent diseases. Plant disease 86:1056-1066.
Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyere, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pe, M. E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A. F., Weissenbach, J., Quetier, F., and Wincker, P. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463-467.
John, G. H., and Langley, P. 1995. Estimating continuous distributions in Bayesian classifiers. Pages 338-345 in: Proceeding of the Eleventh Conference Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., Burlington, MA, U.S.A.
Johnson, C. H., Mori, T., and Xu, Y. 2008. A cyanobacterial circadian clockwork. Curr. Biol. 18:R816-R825.
Joshi, H. M., and Tabita, F. R. 1996. A global two component signal transduction system that integrates the control of photosynthesis, carbon dioxide assimilation, and nitrogen fixation. Proc. Natl. Acad. Sci. U.S.A. 93:14515-14520.
Kanehisa, M. 2002. The KEGG database. Novartis Found Symp. 247:91-101; discussion 101-103, 119-128, 244-152.
Karpinets, T. V., Obraztsova, A. Y., Wang, Y., Schmoyer, D. D., Kora, G. H., Park, B. H., Serres, M. H., Romine, M. F., Land, M. L., Kothe, T. B., Fredrickson, J. K., Nealson, K. H., and Uberbacher, E. C. 2010. Conserved synteny at the protein family level reveals genes underlying *Shewanella* species' cold tolerance and predicts their novel phenotypes. Funct. Integr. Genomics 10:97-110.
Karpinets, T. V., Park, B. H., and Uberbacher, E. C. 2012. Analyzing large biological datasets with association networks. Nucleic Acids Res. 40:e131-e131.
Kikumoto, T. 1980. Ecological aspects of the soft rot bacteria. Rep. Inst. Agric. Res. Tohoku Univ. 31:19-41.
Kreimer, A., Doron-Faigenboim, A., Borenstein, E., and Freilich, S. 2012. NetCmpt: A network-based tool for calculating the metabolic competition between bacterial species. Bioinformatics 28:2195-2197.
Krieger, C. J., Zhang, P., Mueller, L. A., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S. Y., and Karp, P. D. 2004. MetaCyc: A multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res. 32:D438-D442.
Ladha, J. K., Garcia, M., Miyan, S., Padre, A. T., and Watanabe, I. 1989. Survival of *Azorhizobium caulinodans* in the soil and rhizosphere of wetland rice under *Sesbania rostrata*–rice rotation. Appl. Environ. Microbiol. 55:454-460.
Lindemann, A., Moser, A., Pessi, G., Hauser, F., Friberg, M., Hennecke, H., and Fischer, H. M. 2007. New target genes controlled by the *Bradyrhizobium japonicum* two-component regulatory system RegSR. J. Bacteriol. 189:8928-8943.
Liolios, K., Chen, I. M., Mavromatis, K., Tavernarakis, N., Hugenholtz, P., Markowitz, V. M., and Kyrpides, N. C. 2010. The Genomes On Line Database (GOLD) in 2009: Status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 38:D346-D354.
Long, S. R. 1989. *Rhizobium*–legume nodulation—Life together in the underground. Cell 56:203-214.
Lowe, R. H., and Evans, H. J. 1962. Carbon dioxide requirement for growth of legume nodule bacteria. Soil Sci. 94:351.
Mansfield, J., Genin, S., Magori, S., Citovsky, V., Sriariyanum, M., Ronald, P., Dow, M., Verdier, V., Beer, S. V., Machado, M. A., Toth, I., Salmond, G., and Foster, G. D. 2012. Top 10 plant pathogenic bacteria in molecular plant pathology. Mol. Plant Pathol. 13:614-629.
Mavromatis, K., Chu, K., Ivanova, N., Hooper, S. D., Markowitz, V. M., and Kyrpides, N. C. 2009. Gene context analysis in the Integrated Microbial Genomes (IMG) data management system. PLoS One 4:e7979. Published online.
McCutcheon, J. P., and Moran, N. A. 2012. Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. 10:13-26.
Mi, X., Swenson, N. G., Valencia, R., Kress, W. J., Erickson, D. L., Perez, A. J., Ren, H., Su, S. H., Gunatilleke, N., Gunatilleke, S., Hao, Z., Ye, W., Cao, M., Suresh, H. S., Dattaraja, H. S., Sukumar, R., and Ma, K. 2012. The contribution of rare species to community phylogenetic diversity across a global network of forest plots. Am. Nat. 180:E17-30.
Mintz-Oron, S., Meir, S., Malitsky, S., Ruppin, E., Aharoni, A., and Shlomi, T. 2012. Reconstruction of *Arabidopsis* metabolic network models accounting for subcellular compartmentalization and tissue-specificity. Proc. Natl. Acad. Sci. U.S.A. 109:339-344.
Newton, A. C., Fitt, B. D., Atkins, S. D., Walters, D. R., and Daniell, T. J. 2010. Pathogenesis, parasitism and mutualism in the trophic space of microbe-plant interactions. Trends Microbiol. 18:365-373.
Nicholas, J. D., Wilson, P. W., and Kobayashi, M. 1962. Cobalt requirement for inorganic nitrogen metabolism in microorganisms. Proc. Natl. Acad. Sci. U.S.A. 48:1537-1542.
Olivares, J., Bedmar, E. J., and Sanjuan, J. 2013. Biological nitrogen fixation in the context of global change. Mol. Plant-Microbe Interact. 26:486-494.
O'Neill, J. S., van Ooijen, G., Dixon, L. E., Troein, C., Corellou, F., Bouget, F. Y., Reddy, A. B., and Millar, A. J. 2011. Circadian rhythms persist without transcription in a eukaryote. Nature 469:554-558.
Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R. L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J., and Buell, C. R. 2007. The TIGR Rice Genome Annotation Resource: Improvements and new features. Nucleic Acids Res. 35:D883-D887.
Park, B. H., Karpinets, T. V., Syed, M. H., Leuze, M. R., and Uberbacher, E. C. 2010. CAZymes Analysis Toolkit (CAT): Web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. Glycobiology 20:1574-1584.
Parry, M. A. J., Andralojc, P. J., Scales, J. C., Salvucci, M. E., Carmo-Silva, A. E., Alonso, H., and Whitney, S. M. 2013. Rubisco activity and regulation as targets for crop improvement. J. Exp. Bot. 64:717-730.
Pérombelon, M. 2002. Potato diseases caused by soft rot erwinias: An overview of pathogenesis. Plant Pathol. 51:1-12.
Pessi, G., Ahrens, C. H., Rehrauer, H., Lindemann, A., Hauser, F., Fischer, H. M., and Hennecke, H. 2007. Genome-wide transcript analysis of *Bradyrhizobium japonicum* bacteroids in soybean root nodules. Mol. Plant-Microbe Interact. 20:1353-1363.
Rosenblueth, M., and Martinez-Romero, E. 2006. Bacterial endophytes and

their interactions with hosts. Mol. Plant-Microbe Interact. 19:827-837.
Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., Kaneko, T., Nakamura, Y., Shibata, D., Aoki, K., Egholm, M., Knight, J., Bogden, R., Li, C. B., Shuang, Y., Xu, X., Pan, S. K., Cheng, S. F., Liu, X., Ren, Y. Y., Wang, J., Albiero, A., Dal Pero, F., Todesco, S., Van Eck, J., Buels, R. M., Bombarely, A., Gosselin, J. R., Huang, M. Y., Leto, J. A., Menda, N., Strickler, S., Mao, L. Y., Gao, S., Tecle, I. Y., York, T., Zheng, Y., Vrebalov, J. T., Lee, J., Zhong, S. L., Mueller, L. A., Stiekema, W. J., Ribeca, P., Alioto, T., Yang, W. C., Huang, S. W., Du, Y. C., Zhang, Z. H., Gao, J. C., Guo, Y. M., Wang, X. X., Li, Y., He, J., Li, C. Y., Cheng, Z. K., Zuo, J. R., Ren, J. F., Zhao, J. H., Yan, L. H., Jiang, H. L., Wang, B., Li, H. S., Li, Z. J., Fu, F. Y., Chen, B. T., Han, B., Feng, Q., Fan, D. L., Wang, Y., Ling, H. Q., Xue, Y. B. A., Ware, D., McCombie, W. R., Lippman, Z. B., Chia, J. M., Jiang, K., Pasternak, S., Gelley, L., Kramer, M., Anderson, L. K., Chang, S. B., Royer, S. M., Shearer, L. A., Stack, S. M., Rose, J. K. C., Xu, Y. M., Eannetta, N., Matas, A. J., McQuinn, R., Tanksley, S. D., Camara, F., Guigo, R., Rombauts, S., Fawcett, J., Van de Peer, Y., Zamir, D., Liang, C. B., Spannagl, M., Gundlach, H., Bruggmann, R., Mayer, K., Jia, Z. Q., Zhang, J. H., Ye, Z. B. A., Bishop, G. J., Butcher, S., Lopez-Cobollo, R., Buchan, D., Filippis, I., Abbott, J., Dixit, R., Singh, M., Singh, A., Pal, J. K., Pandit, A., Singh, P. K., Mahato, A. K., Dogra, V., Gaikwad, K., Sharma, T. R., Mohapatra, T., Singh, N. K., Causse, M., Rothan, C., Schiex, T., Noirot, C., Bellec, A., Klopp, C., Delalande, C., Berges, H., Mariette, J., Frasse, P.,. Vautrin, S., Zouine, M., Latche, A., Rousseau, C., Regad, F., Pech, J. C., Philippot, M., Bouzayen, M., Pericard, P., Osorio, S., del Carmen, A. F., Monforte, A., Granell, A., Fernandez-Munoz, R., Conte, M., Lichtenstein, G., Carrari, F., De Bellis, G., Fuligni, F., Peano, C., Grandillo, S. , Termolino, P., Pietrella, M., Fantini, E., Falcone, G., Fiore, A., Giuliano, G., Lopez, L., Facella, P., Perrotta, G., Daddiego, L., Bryan, G., Orozco, M., Pastor, X., Torrents, D., van Schriek, K. N. V. M. G. M., Feron, R. M. C., van Oeveren, J., de Heer, P., daPonte, L., Jacobs-Oomen, S., Cariaso, M., Prins, M., van Eijk, M. J. T., Janssen, A., van Haaren, M. J. J., Jo, S. H., Kim, J., Kwon, S. Y., Kim, S., Koo, D. H., Lee, S., Hur, C. G., Clouser, C., Rico, A., Hallab, A., Gebhardt, C., Klee, K., Jocker, A., Warfsmann, J., Gobel, U., Kawamura, S., Yano, K., Sherman, J. D., Fukuoka, H., Negoro, S., Bhutty, S., Chowdhury, P., Chattopadhyay, D., Datema, E., Smit, S., Schijlen, E. W. M., van de Belt, J., van Haarst, J. C., Peters, S. A., van Staveren, M. J., Henkens, M. H. C., Mooyman, P. J. W., Hesselink, T., van Ham, R. C. H. J., Jiang, G. Y., Droege, M., Choi, D., Kang, B. C., Kim, B. D., Park, M., Kim, S., Yeom, S. I., Lee, Y. H., Choi, Y. D., Li, G. C., Gao, J. W., Liu, Y. S., Huang, S. X., Fernandez-Pedrosa, V., Collado, C., Zuniga, S., Wang, G. P., Cade, R., Dietrich, R. A., Rogers, J., Knapp, S., Fei, Z. J., White, R. A., Thannhauser, T. W., Giovannoni, J. J., Botella, M. A., Gilbert, L., Gonzalez, R., Goicoechea, J. L., Yu, Y., Kudrna, D., Collura, K., Wissotski, M., Wing, R., Schoof, H., Meyers, B. C., Gurazada, A. B., Green, P. J., Mathur, S., Vyas, S., Solanke, A. U., Kumar, R., Gupta, V., Sharma, A. K., Khurana, P., Khurana, J. P., Tyagi, A. K., Dalmay, T., Mohorianu, I., Walts, B., Chamala, S., Barbazuk, W. B., Li, J. P., Guo, H., Lee, T. H., Wang, Y. P., Zhang, D., Paterson, A. H., Wang, X. Y., Tang, H. B., Barone, A., Chiusano, M. L., Ercolano, M. R., D'Agostino, N., Di Filippo, M., Traini, A., Sanseverino, W., Frusciante, L., Seymour, G. B., Elharam, M., Fu, Y., Hua, A., Kenton, S., Lewis, J., Lin, S. P., Najar, F., Lai, H. S., Qin, B. F., Qu, C. M., Shi, R. H., White, D., White, J., Xing, Y. B., Yang, K. Q., Yi, J., Yao, Z. Y., Zhou, L. P., Roe, B. A., Vezzi, A., D'Angelo, M., Zimbello, R., Schiavon, R., Caniato, E., Rigobello, C., Campagna, D., Vitulo, N., Valle, G., Nelson, D. R., De Paoli, E., Szinay, D., de Jong, H. H., Bai, Y. L., Visser, R. G. F., Lankhorst, R. M. K., Beasley, H., McLaren, K., Nicholson, C., Riddle, C., Gianese, G., and Tomato Genome Consortium 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635-641.
Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., Ying, K., Yeh, C. T., Emrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A. P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J. M., Deragon, J. M., Estill, J. C., Fu, Y., Jeddeloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A., and Wilson, R. K. 2009. The B73 maize genome: Complexity, diversity, and dynamics. Science 326:1112-1115.
Schwender, J., Goffman, F., Ohlrogge, J. B., and Shachar-Hill, Y. 2004. Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. Nature 432:779-782.
Sheikh, A., Charles, R. C., Sharmeen, N., Rollins, S. M., Harris, J. B., Bhuiyan, M. S., Arifuzzaman, M., Khanam, F., Bukka, A., Kalsy, A., Porwollik, S., Leung, D. T., Brooks, W. A., LaRocque, R. C., Hohmann, E. L., Cravioto, A., Logvinenko, T., Calderwood, S. B., McClelland, M., Graham, J. E., Qadri, F., and Ryan, E. T. 2011. In vivo expression of *Salmonella enterica* serotype Typhi genes in the blood of patients with typhoid fever in Bangladesh. PLoS Negl. Trop. Dis. 5:e1419. Published online.
Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., and Ideker, T. 2011. Cytoscape 2.8: New features for data integration and network visualization. Bioinformatics 27:431-432.
Soto, M. J., Dominguez-Ferreras, A., Perez-Mendoza, D., Sanjuan, J., and Olivares, J. 2009. Mutualism versus pathogenesis: The give-and-take in plant-bacteria interactions. Cell. Microbiol. 11:381-388.
Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U.S.A. 102:15545-15550.
Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G. L., Cooper, D., Coutinho, P. M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J. C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D. R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C. J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y., and Rokhsar, D. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596-1604.
van der Wolf, J., Speksnijder, A., Velvis, H., van de Haar, J., and van Doorn, J. 2007. Why is *Erwinia chrysanthemi* (*Dickeya* sp.) taking over?—The ecology of a blackleg pathogen. Pages 32-33 in: New and old pathogens of potato in changing climate. Proceedings of the EAPR Pathology Section seminar, Hattula, Finland.
White, J. R., Nagarajan, N., and Pop, M. 2009. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput. Biol. 5:e1000352. Published online.

AUTHOR-RECOMMENDED INTERNET RESOURCES

The Genomes On Line Database (GOLD): genomesonline.org/cgi-bin/GOLD/index.cgi
The Integrated Microbial Genomes (IMG) data management system: img.jgi.doe.gov
Kyoto Encyclopedia of Genes and Genomes or (KEGG): www.genome.jp/kegg
The Carbohydrate-Active enZYmes (CAZy) database: www.cazy.org
The MetaCyc database of metabolic pathways and enzymes: metacyc.org
The BioCyc collection of Pathway/Genome databases: biocyc.org