Proteomics Analysis of *Psychotria* Leaf Nodule Symbiosis: Improved Genome Annotation and Metabolic Predictions

Aurelien L. Carlier,¹ Ulrich Omasits,² Christian H. Ahrens,² and Leo Eberl¹

¹Institute of Plant Biology and ²Institute of Molecular Life Sciences, University of Zurich, CH-8008 Zürich

Submitted 28 May 2013. Accepted 15 July 2013.

Several plant species of the genus *Psychotria* (Rubiaceae) harbor *Burkholderia* sp. bacteria within specialized leaf nodules. The bacteria are transmitted vertically between plant generations and have not yet been cultured outside of their host. This symbiosis is considered to be obligatory because plants devoid of symbionts fail to develop into mature individuals. The genome of ‘*Candidatus Burkholderia kirkii*’ has been sequenced recently and has revealed evidence of reductive genome evolution, as shown by the proliferation of insertion sequences and the presence of numerous pseudogenes. We employed shotgun proteomics to investigate the expression of ‘*Ca. B. kirkii*’ proteins in the leaf nodule. Drawing from this dataset and refined comparative genomics analyses, we designed a new pseudogene prediction algorithm and improved the genome annotation. We also found conclusive evidence that nodule bacteria allocate vast resources to synthesis of secondary metabolites, possibly of the C₇N aminocyclitol family. Expression of a putative 2-epi-5-valiolone synthase, a key enzyme of the C₇N aminocyclitol synthesis, is high in the nodule population but downregulated in bacteria residing in the shoot apex, suggesting that production of secondary metabolites is particularly important in the leaf nodule.

Several hundred species of the flowering plant family Rubiaceae genera *Pavetta*, *Sericanthe*, and *Psychotria* form symbiotic relationships with bacteria, resulting in visible nodules in the leaf lamina. Numerous *Psychotria* understory shrub species from continental Africa and Madagascar possess characteristic nodules or galls on the surface of their leaves. The presence of bacteria within the nodules was first demonstrated by Zimmermann (1902) and was described in more detail by von Faber (1912). Leaf nodule symbiosis is relatively rare, affecting only approximately 500 flowering plant species. Although only a few examples have been studied thus far, they all share characteristics that make leaf nodule symbiosis unique among land plants (Miller 1990). First, evidence has been presented that the bacteria are not acquired by the plants from the environment but are transmitted hereditarily to the progeny via colonization of the developing seed (Lemaire et al. 2012). There

is some indication that *Psychotria* symbionts can also be acquired horizontally by unknown means but vertical transmission seems to be the rule rather than the exception (Lemaire et al. 2012). Although vertical transmission of bacterial endosymbionts from one generation to the next is common among insect–microbe symbioses (Sachs et al. 2011; Saridaki and Bourtzis 2010), it is unique in higher plants. Second, as the seeds germinate, bacteria in the shoot buds infect developing leaves through open stomata, which later develop into leaf nodules (Whitmoyer and Horner 1970; Miller et al. 1983). Killing the bacteria causes degeneration of the apical bud, which eventually provokes growth arrest. Seeds cured of the symbionts by antibiotic or heat treatment germinate normally but the resulting bacteria-free plants fail to reach maturity, wither, and die within 2 years (*our own observations*). The obligatory character of the interaction is common to all leaf nodule symbioses studied thus far (Miller 1990).

Multiple attempts have been made to cultivate the *Psychotria*-symbiotic bacteria on various media and under various conditions but all have failed thus far (Miller 1990; Van Oevelen et al. 2003; *our unpublished data*). Culture-independent molecular techniques have assigned the symbiotic bacteria to the genus *Burkholderia* (Lemaire et al. 2012; Van Oevelen et al. 2002, 2004). More recently, we have determined the genome sequence of *Burkholderia kirkii* and found evidence of genome erosion (Carlier and Eberl 2012). This is manifested by a proliferation of insertion element sequences and the presence of an inordinate number of pseudogenes. We estimated that only 2,274 genes, spanning 48% of the 4-Mb genome sequence, coded for proteins, whereas more than 4,000 were pseudogenes or remnants of pseudogenes. The results of our analysis did not support previous speculations about the nature of the symbiotic relationship, including nitrogen fixation in the nodules and cytokinin production (Centifanto and Silver 1964; Edwards and Lamotte 1975; Van Oevelen et al. 2003). However, we have found a cluster of genes located on the 140-kb pKIR01 plasmid, which had no known homologues in other *Burkholderia* spp. These genes are likely involved in the biosynthesis of secondary metabolites of the C₇N aminocyclitol family. Natural products of the C₇N aminocyclitol family are exclusively produced by microorganisms, usually soil bacteria belonging to the order Actinomycetes. Their chemical structures commonly share a valienamine moiety. The C₇N aminocyclitols display various biological activities, ranging from antifungal (e.g., validamycin) to antibacterial (adiposins, oligostatins, and pyralomicins), insecticidal (validoxylamine A), and glycosidase inhibitors with medical applications (acarbose, amylostatins, and cetoniacytone A) (Mahmud 2003).

Discriminating between genes and pseudogenes in the ‘*Ca. B. kirkii*’ genome is an enormous challenge. Usually, the anno-

Corresponding author: A. Carlier; Telephone: +41 44 634 82 27; Fax: +41 44 634 82 04; E-mail: aurelien.carlier@uzh.ch

*The e-Xtra logo stands for “electronic extra” and indicates that five supplementary tables and two supplementary figures are published online.

tation of pseudogenes relies on comparative genomics methods between closely related organisms (de Souza et al. 2009; Ran et al. 2010; Schrimpe-Rutledge et al. 2012; Zhang et al. 2006). '*Ca. B. kirkii*' was the first organism of the Rubiaceae endosymbiont clade whose genome was sequenced at the time of our analysis (Lemaire et al. 2011) and, consequently, no closely related sequences were available. Moreover, traditional pseudogene detection methods did not work well because of the unusually large gene repertoire (i.e., the pan genome) of the genus *Burkholderia* (Ussery et al. 2009). We reasoned that a majority of the genes encoded in a drastically reduced genome should have orthologs in species of the same evolutionary lineage. Therefore, we developed a novel approach that relied on the comparison of '*Ca. B. kirkii*' genes to those of a database of putative orthologs to identify intact genes (Carlier and Eberl 2012).

Accurate genome annotation is of prime importance regardless of the organism studied, as indicated by the vast resources dedicated to curation and verification of annotations (e.g., the Ensembl project). It is arguably especially critical for uncultured organisms such as *Ca. B. kirkii*, because verification of protein predictions is obviously extremely difficult for organisms that cannot be manipulated. In this study, we used shotgun proteomics methods to identify bacterial proteins expressed in the nodule of *Psychotria kirkii*.

We used this inventory to estimate the accuracy of our previous genome annotation and designed a new, automated algorithm that allowed us to minimize the number of falsely identified pseudogenes. Our semiquantitative data also allowed verification of our genome-based predictions with respect to the bacterial metabolic pathways required for colonization of the *P. kirkii* nodule. Quantitative reverse-transcription polymerase chain reaction (RT-PCR) experiments were performed to confirm that genes participating in synthesis of secondary metabolites are highly expressed in the leaf nodule.

RESULTS AND DISCUSSION

Determination and analysis of the '*Ca. B. kirkii*' nodule proteome.

One of the main challenges of the genome analysis of *Ca. B. kirkii*, and of heavily eroded genomes in general, is to distinguish genuine protein coding genes from spurious coding sequence (CDS) annotations. Our original approach relied heavily upon available genomic information from free-living relatives of *Ca. B. kirkii*, without the possibility of experimental validation. Although the notion that pseudogenes are translationally silent is open to debate (Castellana et al. 2008; Gupta et al. 2007; Schrimpe-Rutledge et al. 2012), we assumed that only intact genes are translated and, therefore, that the mapping of the '*Ca. B. kirkii*' proteome may help to identify pseudogenes that had been misannotated. The first step in our work flow was to enrich our samples in bacteria to maximize detection of low-abundance bacterial proteins. To that purpose, we separated bacterial cells from plant cells by density gradient centrifugation. This approach would not allow identification of secreted proteins; however, only 26 CDS were predicted to encode secreted proteins, most of which were of housekeeping (e.g., cell wall biosynthesis) or hypothetical function and were presumably expressed in the leaf nodule (Supplementary Table S1). We hypothesize that the intractability of secreted proteins by our approach does not significantly impact our findings. We subsequently fractionated our samples by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) prior to trypsin digestion. Peptides were analyzed by liquid chromatography tandem mass spectrometry and mass spectra were searched against a database consisting of protein products of

predicted genes (including predicted pseudogenes) from the previously published '*Ca. B. kirkii*' genome (Carlier and Eberl 2012), as well as translated open reading frames (ORF) from intergenic regions and predicted *P. kirkii* chloroplastic proteins (*unpublished data*). Our method resulted in protein identifications with a very low false discovery rate (FDR) of 1.2%. Of the peptide-spectrum matches (PSM) that fulfilled our stringent criteria (discussed below), 96.2% could be attributed to '*Ca. B. kirkii*' proteins, with *P. kirkii* chloroplast proteins being the main contaminant (1.9% of total PSM). This indicated that the enrichment strategy was successful.

We could unambiguously identify mass spectra corresponding to 1,184 proteins spread over three biological replicates in our database. Of those, 1,106 proteins were encoded by the reference '*Ca. B. kirkii*' genome (ENA accession number CAFE000000000, version 1), representing nearly 50% of all CDS features (pseudogenes excluded) of the genome. Although the proportion of predicted proteins that we could detect was high, several lines of evidence indicated that we did not obtain the complete expressed proteome of the organism. Logistic regression analysis indicated that the probability of detection varied according to a protein's molecular weight, isoelectric point, and predicted membrane localization. Variables such as the presence of a signal peptide or grand average of hydropathy value did not significantly influence the probability of detection (data not shown). A detailed analysis revealed that small, basic proteins were underrepresented among detected proteins (Supplementary Fig. S2), similar to observations made by others (Brunner et al. 2007).

Novel ORF and misannotated pseudogenes.

We detected a significant number of peptides (1.3% of total) that were not part of the predicted '*Ca. B. kirkii*' proteome. However, we found good spectral matches for those peptides to 15 new ORF that had not been predicted in the original '*Ca. B. kirkii*' genome annotation (Supplementary Table S2). Most of the newly identified CDS were short, with an average length of 170 amino acids (aa). We also observed that eight of these newly annotated CDS overlapped with longer genes that had been erroneously predicted. Peptides mapping in the predicted 5' untranslated region of 65 predicted genes indicated a misannotated N-terminus and annotations were corrected accordingly. Our data confirm observations by others that no single gene prediction software is completely error proof, and that combining gene calls from various sources helps improve accuracy (Ansong et al. 2011; Armengaud 2009; Castellana and Bafna 2010).

The original '*Ca. B. kirkii*' genome annotation contains 4,000 putative CDS predicted to be pseudogenes. Of these, we found evidence for the expression of 54 genes (Supplementary Table S3a). These translated putative pseudogenes (TPP) represented only a small proportion of the total coding potential of the genome, and accounted for less than 1.6% of total assigned spectra. However, these TPP contained precious information that could help develop a more accurate pseudogene prediction algorithm for the annotation of the *Ca. B. kirkii*. Of the 54 TPP, 22 possessed an obvious frame shift or null mutation and were most likely bona fide pseudogenes subjected to residual translation. Thirty-two TPP were putative functional genes erroneously tagged as pseudogenes.

Among the 32 cases of false TPP, we could identify several scenarios that led to an erroneous functional assignment: i) the protein sequence contained low-complexity regions that were masked in the blast search strategy, ii) a strict ortholog was not found but the sequence showed high homology to sequences in the BKH9 database we used as a putative orthologs database, iii) the protein is a fusion protein without a true ortholog but with high homology to proteins in the BKH9 database, iv) the

predicted protein is short (<100 aa) and without an ortholog in the BKH9 database, v) the predicted protein sequence has no high scoring homologue in the National Center for Biotechnology Information (NCBI) RefSeq database, and vi) mistakes were made during manual annotation of the genome.

Improved algorithm for the prediction of '*Ca. B. kirkii*' pseudogenes.

We reasoned that the accuracy of our pseudogene prediction could be greatly improved by using more precise alignment methods and greater automation to eliminate errors during the manual curation process. We designed an algorithm to more accurately identify functional genes in the '*Ca. B. kirkii*' genome (details can be found below and in Supplementary Figure

S1). With our new algorithm, we predicted that 2,069 CDS, from a total of 6,274, were functional, compared with 2,274 in the original annotation. We then compared the features of the updated '*Ca. B. kirkii*' genome sequence to our dataset of proteins detected by shotgun proteomics. All but two false TPP identified above were predicted to be functional by our new work flow, whereas only three new cases of false TPP arose. In all cases, the false TPP coded for short products (<100 amino acid residues) with limited homology to protein sequences in the NCBI RefSeq database.

Small, nonconserved CDS are difficult to call accurately and are often found in proteogenomics studies, even in well-annotated genomes (Armengaud 2009; Christie-Oleza et al. 2012; Zhao et al. 2011).

Fig. 1. Functional analysis of the proteomics dataset. **A**, Distribution of genes in COG (clusters of orthologous groups) functional categories. The number of intact genes in each category was counted in the observed (black bars) and the undetected (white bars) proteome. Asterisks denote significant differences in the distributions of genes belonging to a functional category between the observed and undetected proteomes (Pearson χ^2 test). **B**, Relative abundance of proteins per COG category. The y-axis represents the percentage of normalized mass spectrometry spectra for each category in the observed (dark gray) and expected (light gray) proteome. Expected abundances were calculated by dividing total abundance by the percentage of genes in each category. Error bars represent the standard deviation. Column C = energy production and conversion; column D = cell cycle control, cell division, and chromosome partitioning; column E = amino acid transport and metabolism; column F = nucleotide transport and metabolism; column G = carbohydrate transport and metabolism; column H = coenzyme transport and metabolism; column I = lipid transport and metabolism; column J = translation, ribosomal structure, and biogenesis; column K = transcription; column L = replication, recombination, and repair; column M = cell wall/membrane/envelope biogenesis; column N = cell motility; column O = posttranslational modification, protein turnover, and chaperones; column P = inorganic ion transport and metabolism; column Q = secondary metabolites biosynthesis, transport, and catabolism; column R = general function prediction only; column S = function unknown; and column T = signal transduction mechanisms.

Partial proteomics coverage of the '*Ca. B. kirkii*' genome prevented us from evaluating the FDR of presumed functional genes (i.e., true pseudogenes annotated as functional). However, we concluded that our new annotation work flow performed well because i) the FDR (estimated by the number of false TPP per total number of proteins detected) was very low (0.26%) and of little consequence for downstream functional genomics analyses, because the misannotated CDS coded for proteins without functional annotation; and ii) the number of presumed functional genes was significantly lower than our previous predictions (9.1% improvement on the previous annotation). Combined, these two observations indicate that our new pipeline has a higher discriminatory power compared with the original annotation.

Functional analysis of '*Ca. B. kirkii*' proteins expressed in the leaf nodule.

We reasoned that functional categories with an expression bias in the leaf nodule might be of particular significance for the symbiosis. We assigned COG (clusters of orthologous groups) functional category values to each CDS in the '*Ca. B. kirkii*' genome and counted instances for each functional category in the nodule proteome (Fig. 1A). We found significantly fewer gene products falling in the categories N (cell motility and secretion) and S (function unknown) in the proteome than we had predicted from the genome. Importantly, proteins of these categories were not significantly enriched in short and basic proteins or membrane proteins that were difficult to detect with our protocols (data not shown). Our previous genome analysis showed that most genes coding for the biosynthesis of flagella or pili have been eroded in '*Ca. B. kirkii*' (Carlier and Eberl 2012). Therefore, under-representation of COG category

N indicates that the remnant category N genes of the '*Ca. B. kirkii*' genome are not expressed in the nodule.

Genes in COG category P (inorganic ion transport and metabolism), encoding many transporters, also appear under-represented in our samples. By definition, these transporters are located in the membrane and are significantly enriched in insoluble proteins intractable in our analysis. After controlling for membrane localization, we found that the under-representation of COG category P proteins in our observed data was not significant and may be attributed to sampling bias (data not shown).

In contrast, products of genes belonging to COG categories J (translation, ribosomal structure, and biogenesis) and F (nucleotide transport and metabolism) are significantly over-represented. These categories encompass many housekeeping genes required for the normal function of the cell and are expected to be expressed regardless of the physiological state of the cell. Interestingly, detailed analysis of category F genes showed that the products detected in our proteomics study belonged mostly to nucleotide biosynthesis (discussed below). Therefore, nucleotide metabolism may play an important role for the lifestyle of '*Ca. B. kirkii*' in the nodule.

Semiquantitative analysis showed a wide variation in the relative abundance of proteins of different COG categories (Fig. 1B). Proteins belonging to the functional categories J, O, and M account for a large share (over 46%) of the detected spectra. These categories contain many housekeeping functions which are often found in abundance in *Burkholderia* proteomics studies (Riedel et al. 2006; Wongtrakoongate et al. 2011).

Our data also show that proteins involved in secondary metabolites biosynthesis (category Q) are highly abundant. This is

Fig. 2. Expression of genes of the putative C₇N aminocyclitol metabolism. **A**, Map of plasmid pKIR01. The outer circle represents genomic features. Gray boxes depict pseudogenes, blue boxes depict functional genes, and green boxes depict genes thought to be a part of the secondary metabolism. Histograms in the inner circle represent the abundance of the respective gene product (logarithmic scale). Bars are colored according to the relative abundance of the gene product, with red denoting the most abundant proteins (above the 75th percentile of total detected proteins) and yellow the least abundant (below the 25th percentile). **B**, Statistical analysis of protein abundance. The box plot to the left represents the abundance distribution of all detected proteins in our experiments. The plot to the right shows the distribution of proteins of the C₇N aminocyclitol secondary metabolism. Proteins of the secondary metabolism are significantly overrepresented in our experiments (Mann-Whitney U test: *P* value < 0.001).

due to the high number of spectral counts (9.1% of all PSM) of 23 proteins (or 1.1% of the annotated '*Ca. B. kirkii*' genome) encoded on pKIR01, which appear to be involved in the biosynthesis of an unknown aminocyclitol metabolite. These proteins are among the most abundant proteins detected in our study (Fig. 2), suggesting that the synthesis of the secondary metabolite plays a key role in the symbiosis.

Metabolism of '*Ca. B. kirkii*' in the *Psychotria* leaf nodule.

Previous analysis of the '*Ca. B. kirkii*' genome led us to hypothesize that '*Ca. B. kirkii*' had extensive metabolic capabilities and relied on simple compounds from the host plant. We created a metabolic pathways database using the updated '*Ca. B. kirkii*' genome annotation and information about proteins detected in our shotgun proteomics experiments (Supplementary Table S4).

Energy metabolism.

The '*Ca. B. kirkii*' genome encodes all the components necessary for oxidative phosphorylation using an *aa3*-type cytochrome c oxidase (Carlier and Eberl 2012). We found proteomics evidence for all the components of the aerobic respiratory chain. In addition, all the enzymes of the TCA cycle are well represented in our dataset, indicating that degradation of substrates provided by the plant proceeds through the TCA cycle. '*Ca. B. kirkii*' lacks a functional class II fructose biphosphate aldolase, preventing energy production through the glycolysis pathway. Instead, dicarboxylic acids such as malate, fumarate, or maleate might serve as substrates for the TCA cycle. This is further supported by two lines of evidence: i) a putative C4-dicarboxylic acid transporter encoded by the *dctA* homologue BKIR_c2_4659 is among the most abundant transporters in our samples and ii) '*Ca. B. kirkii*' lacks a functional phosphoenol pyruvate carboxylase (pseudogene BKIR_c46_0473), and, thus, is incapable of replenishing TCA intermediates from sugars. Enzymes of the pentose phosphate pathway (PPP) are also well represented in our data. It is possible that '*Ca. B. kirkii*' derives some of its energy from sugars; however, the role of the PPP may also be to provide precursors for other metabolic pathways in the cell.

Carbohydrate transport and metabolism.

We searched our proteomics data for evidence of carbohydrate transporters and found only a few candidates expressed in *Psychotria* nodules. We found evidence for the presence of a putative maltose or glucose ABC transporter (BKIR_c1_2901–2904) and a putative trehalose MFS family transporter (BKIR_c8_3457). The gene BKIR_c16_1932 encodes a putative gluconate permease of the GntP family but we could not find matching peptides in our proteomics dataset. Moreover, proteomics evidence for the presence of a glucose-6-phosphate dehydrogenase (BKIR_81_1223) catalyzing the first step of the oxidative branch of the PPP suggests that glucose-6-phosphate is the entry point into the PPP. The gluconeogenesis pathway, all enzymes of which are expressed in the nodule according to our data, could provide the substrate. Alternatively, some carbohydrates could be metabolized via the Entner-Doudoroff pathway (BKIR_c81_1223, BKIR_c1_2896, BKIR_c16_1930, and BKIR_c16_1931).

Nitrogen metabolism.

Ammonium is most likely taken up through a putative ammonium transporter, encoded by gene BKIR_c2_4671, although we could not find peptide evidence for this membrane protein in our dataset. However, we found strong evidence for an active GS/GOGAT ammonium assimilation pathway (genes BKIR_c77_5228, BKIR_c31_1437, and BKIR_c31_1438).

Comparatively low levels of glutamate dehydrogenase (BKIR_c37_3363) indicate that available ammonium is scarce in the nodule. We found that biosynthetic pathways for all amino acids were active, except for leucine and methionine. Enzymes of the leucine biosynthetic pathway, while encoded by the '*Ca. B. kirkii*' genome, are generally not represented in our samples, and it is possible that Leu or metabolic precursors might be imported from the extracellular milieu by a putative LIV-I branched-chain amino acid transport system encoded by genes BKIR_c26_5785 to BKIR_c26_5788. Met is likely synthesized by methylation of homocysteine via the action of a 5-methyltetrahydrofolate-homocysteine methyltransferase (BKIR_c2_4615–BKIR_c2_4616). We also found peptide evidence for a putative histidine transporter (encoded by genes BKIR_109_2718 to BKIR_109_2721) and a putative glutamate/aspartate ABC transporter (BKIR_c37_3359–BKIR_c37_3362). Expression of these transporters, contrasted with the expression of most amino acid de novo synthesis pathways, suggests that the plant supplies some amino acids to the bacteria but perhaps not in amounts sufficient to meet the needs of the symbiont.

Nucleoside, nucleotide, cofactor, and vitamin metabolism.

Similarly, we found peptide evidence for most enzymes of all nucleoside and nucleotide biosynthesis pathways. We could not find evidence that any nucleotide or nucleoside salvage pathways were expressed in the nodule bacteria. As previously noted, '*Ca. B. kirkii*' is likely a biotin and ascorbate auxotroph (Carlier and Eberl 2012) but all other enzymes for main cofactors and vitamins biosynthesis pathways are well represented in our nodule proteomics dataset.

Secondary metabolite biosynthesis.

We previously found that plasmid pKIR01 harbored genes thus far unique to '*Ca. B. kirkii*' among the Burkholderiaceae that were likely involved in the biosynthesis of C₇N aminocyclitols (Carlier and Eberl 2012). Several other genes not found in other *Burkholderia* spp. (BKIR_c114_0657–0664), also clustered on pKIR01, encode putative enzymes tied to secondary metabolism (Supplementary Table S5). The biosynthesis of secondary metabolites of the aminocyclitol family of compounds involves cyclization of sedoheptulose 7-phosphate, an intermediate of the PPP, to 2-epi-5-epi-valiolone (Bai et al. 2006; Yu et al. 2005). A putative 2-epi-5-epi-valiolone synthase is encoded by gene BKIR_c149_4879 and is among the top 2% most abundant proteins identified, accounting for approximately 0.85% of all spectra (Fig. 2). The intermediate 2-epi-5-epi-valiolone is then usually phosphorylated by a C7-cyclitol-7-kinase (Wehmeier and Piepersberg 2004). This phosphorylation step is probably carried out in '*Ca. B. kirkii*' by the putative sugar kinase encoded by BKIR_c149_4878, which is also among the 2% most abundant proteins in the nodule. Although the exact enzymatic functions of many putative genes of the secondary metabolism are difficult to predict, all are found in abundance in our proteomics data (Fig. 2A). Overall, we found that the quantity of the 23 putative enzymes and transporters of the '*Ca. B. kirkii*' putative secondary metabolism make up a cumulated 5.5% of the proteins in our samples. The average relative abundance of these proteins is also significantly higher than the rest of the detected proteome (Fig. 2B). This observation indicates that a large proportion of '*Ca. B. kirkii*' metabolism is dedicated to secondary metabolite biosynthesis in the *Psychotria* nodule. At the moment, little is known about the role of secondary metabolites in the *Psychotria*–*Burkholderia* symbiosis but preliminary evidence suggests that they may protect the plants against herbivores or fungal pathogens (*our unpublished data*).

Regulation of secondary metabolite biosynthesis.

Contrary to previous speculations (Miller 1990; Miller et al. 1983), the high proportion of central metabolism enzymes in the nodule proteome provides compelling evidence that nodule bacteria are metabolically active and that expression of enzymes directing the biosynthesis of a secondary metabolite is important for the symbiosis. We hypothesized that, if the role of mature nodule bacteria was to supply the host with a secondary metabolite, the bacterial colony at the shoot apex may only serve for the propagation of the symbionts. Therefore, the apical colony should exhibit markedly lowered secondary metabolism.

In order to test this hypothesis, we measured the levels of 2-epi-5-epi-valiolone synthase, catalyzing the first committed step of aminocyclitol biosynthesis, in the nodules and in the bacteria of the apical colony. Because the bacteria in the shoot apex were too few to be amenable to proteomics analysis, we chose to measure the levels of BKIR_c149_4879 mRNA by quantitative RT-PCR. The expression of BKIR_c149_4879 was induced by an average of 11.4-fold (standard deviation of 4.6-fold) in mature nodules compared with apical bacteria.

These results indicate that the synthesis of secondary metabolites is tightly regulated and varies within the life cycle of the symbiotic bacteria. The fact that secondary metabolite biosynthesis is upregulated in the mature nodule prompted us to speculate that the bacteria fulfill at least part of their symbiotic role at that site. Moreover, quorum sensing might play a role in regulating expression of genes of the putative aminocyclitol pathway, because we have found proteomics evidence for expression of a *luxIR*-type regulatory pair of genes on pKIR01 (BKIR_c48_6370 and BKIR_c48_6371). The switch from a low-density lifestyle in the shoot apical chamber to the tightly packed leaf nodule might induce the expression of the putative aminocyclitol biosynthetic pathway.

Conclusions.

We previously sequenced the genome of the uncultured '*Ca. B. kirkii*' symbiont with the aim to formulate testable hypotheses about the nature of the obligate plant–bacterium symbiosis affecting some Rubiaceae species (Carlier and Eberl 2012). One of the main challenges when attempting to understand '*Ca. B. kirkii*'s physiology was to exclude the many pseudogenes from our model without accidentally rejecting functional genes. In this study, we used a proteogenomic approach to validate our original gene models and improve their accuracy. Our proteomics experiment was very successful, considering that we collected in planta expression data for over 54% of the CDS features of the '*Ca. B. kirkii*' genome (excluding predicted pseudogenes). This is in contrast to typical figures in the range of 20 to 30% proteome coverage for root nodule bacteria in planta, for example (Delmotte et al. 2010).

The vast majority of the pseudogenes we had predicted were absent from our proteomics dataset. We found only 22 translated pseudogenes and 32 functional genes wrongly annotated as pseudogenes. After careful review of the data, we were able to develop an algorithm that could detect with high accuracy whether a CDS feature was a protein coding gene or a pseudogene. We used this program to update the genome annotation of '*Ca. B. kirkii*' but the methodology and tools described here are also applicable to other bacterial genomes with similarly high proportions of pseudogenes; for example, Rubiaceae leaf nodule symbionts (currently being sequenced by us) and other obligate endosymbionts (Moran 2002).

Previous microscopic investigations suggested that at least two physiologically distinct '*Ca. B. kirkii*' populations co-exist in the adult *P. kirkii* host plant (Lemaire et al. 2012; Miller et al. 1983). Cells of these two populations have markedly dis-

tinct morphologies: nodule bacteria are filamentous and misshapen, whereas bacteria at the shoot tip are short rods more typical of free-living *Burkholderia* spp. Moreover, histological staining techniques showed that bacterial RNA metabolism peaks at the early stages of nodule development (Whitmoyer and Horner 1970). These observations led to the speculation that bacteria contribute to the symbiosis at the early stage of nodule colonization and development whereas they are dormant in the mature nodule (Miller 1990; Miller et al. 1983). In contrast, functional analysis of our proteomics dataset showed that bacteria were metabolically active within the nodules. Moreover, our data suggest that a large part of the bacterial activity is devoted to the production of secondary metabolites, likely of the C₇N aminocyclitol family. Because compounds of the C₇N aminocyclitol family tend to exhibit strong biological activities (fungicidal, antibiotic, or insecticidal), it is tempting to speculate that these compounds provide a strong fitness benefit to the plant that may explain the evolutionary maintenance of a strict, vertically transmitted symbiosis (Carlier and Eberl 2012). Heightened resistance against pathogens or herbivores mediated by secondary metabolites has been documented to provide such fitness benefits to some endophyte–host symbioses (Lane et al. 2000). The fact that the key enzyme in C₇N aminocyclitol biosynthesis is downregulated in the apical bud colony further substantiates our hypothesis that the bacteria fulfill their symbiotic role in the mature nodule and not at the shoot apex. Work underway aims at determining the structures of the secondary metabolites and at elucidating their role in the *Psychotria–Burkholderia* leaf nodule symbiosis. We do not have sufficient data at this time to speculate whether secondary metabolism alone can explain the obligate nature of the symbiosis. One could envision, for instance, that the host plant has evolved a dependence on secondary metabolites as a means to keep underperforming bacterial symbionts out of the breeding population, thereby enforcing partner collaboration.

MATERIALS AND METHODS

Isolation of symbiotic bacteria from *P. kirkii* leaves.

Source leaves were collected from approximately 1-year-old *P. kirkii* plants kept in the greenhouse of the botanical garden of the University of Zürich. Leaves were collected in the middle of the light cycle and washed in distilled water. Symbiotic bacteria were isolated from dissected nodules by density gradient centrifugation, as described previously (Carlier and Eberl 2012), with all solutions supplemented with complete protease inhibitor cocktail (Roche Diagnostics AG, Rotkreuz, Switzerland) throughout the procedure. The experiment was performed independently on three *P. kirkii* specimens grown in our greenhouse.

Protein sample preparation.

Whole '*Ca. B. kirkii*' cells purified by density gradient centrifugation were lysed by boiling in SDS loading buffer. Proteins were separated according to size by SDS-PAGE on 12% polyacrylamide gels. The gel was stained with Coomassie blue reagent and individual lanes were cut into 16 slices containing approximately similar amounts of protein. Each gel slice was subjected to Trypsin treatment (Promega Corp., Madison, WI, U.S.A.), as described previously (Toyofuku et al. 2012).

Protein identification.

Chromatographic separation and m/z spectra acquisition was done according to Toyofuku and associates (2012). A first run was used to generate a mass spectrometry exclusion list which, in turn, was used to maximize identification of low-abundant proteins during a second analytical run (Kristensen et

al. 2004). The mass spectra were searched against a protein sequence database containing products of both predicted protein coding genes and pseudogenes from the previously published *B. kirkii* genome (Carlier and Eberl 2012), predicted *P. kirkii* chloroplastic proteins (*unpublished data*), and products of ORF predicted in intergenic regions as well as protein sequences of 256 common contaminants. We relied on the Mascot search engine (version 2.3.0; Matrix Science, Boston) with the following search parameters. Carbamidomethylation was set as a fixed modification on all cysteine residues. Oxidation of methionine residues, deamidation of asparagine and glutamine residues, as well as cyclization of N-terminal glutamine residues were all considered as optional modifications. Precursor ion mass tolerance was set to 10 ppm, fragment ion mass tolerance was set to 0.8 Da, and the automatic decoy search option was enabled. Spectra were searched for a match to fully tryptic and semitryptic peptides with up to two missed cleavage sites. The Mascot Percolator program (Brosch et al. 2009) was used to improve the PSM scores based on the target-decoy search strategy (Elias and Gygi 2007). Score cutoffs were set that resulted in a 0.1% FDR at the PSM level. All PSM above this cutoff were assessed with the PeptideClassifier software (Qeli and Ahrens 2010): peptides that unambiguously matched one protein sequence were considered, while peptides which matched multiple proteins were not included. For unambiguous protein identification, we required at least three independent PSM or, alternatively, two PSM for two different peptides. This resulted in an FDR of 1.2% at the protein level. Protein abundance was estimated based on the observed spectral counts using the method described by Schrimpf and colleagues (2009). Proteomics data can be downloaded from the ProteomeXchange repository under accession PXD000147.

Pseudogene assignment and updated annotation of the '*Ca. B. kirkii*' genome.

The assignment of pseudogenes and putative functional genes was done using a computer script written in the Python 2.7 language (script available upon request). Predicted protein sequences of all CDS features from the '*Ca. B. kirkii*' genome were searched against a database comprising translated CDS of nine publicly available genomes of free-living *Burkholderia* spp. related to *B. kirkii* (BKH9 database). These species are *B. phyatum* STM815, *B. phytofirmans* PsJN, *B. graminis* C4D1M, *B. xenovorans* LB400, *Burkholderia* sp. CCGE1001, *Burkholderia* sp. CCGE1002, *Burkholderia* sp. CCGE1003, *Burkholderia* sp. SJ98, and *Burkholderia* sp. YI23. Reciprocal best blastp hit strategy (cutoff: 50% identity over 50% length and e value $< 10^{-6}$) identified putative orthologs for each '*Ca. B. kirkii*' feature, as described previously (Carlier and Eberl 2012). The nucleotide sequence of the '*Ca. B. kirkii*' CDS features, including 200 bp of surrounding upstream and downstream sequence, were extracted and aligned against the corresponding protein sequence of the highest-scoring putative ortholog using the TFASTY program, v3.6 (Pearson 2000). '*Ca. B. kirkii*'-predicted CDS were considered pseudogenes if they met the following criteria: i) TFASTY alignments were larger than 50% of the query protein sequence and ii) the '*Ca. B. kirkii*' subject CDS sequence contained a frame shift or a null mutation that resulted in an uninterrupted alignment shorter than 80% of the query BKH9 sequence. In order to identify spurious '*Ca. B. kirkii*' CDS features that overlapped true pseudogenes, the next step in our work flow involved searching the predicted CDS nucleotide sequence against the BKH9 protein database using the Blastx program. If the search returned a hit with more than 50% identity, the CDS was tagged as a pseudogene. Finally, CDS features over 100 aa in length that did not generate high-scoring blastx hits in the

BKH9 database were searched against the NCBI RefSeq database (excluding sequences from the '*Ca. B. kirkii*' genome project CAFE01). Query sequences returning hits with an e value higher than 1.0×10^{-3} were tagged as pseudogenes. The genome annotation was edited using the Artemis software (Carver et al. 2008). Subcellular localization of proteins was predicted using the computer program PSORTb, v3.0 (Yu et al. 2010). The revised genome annotation has been submitted to the European Nucleotide Archive (accession number CAFE01000000).

Metabolic reconstruction.

CDS features of the newly updated '*Ca. B. kirkii*' genome were assigned Enzyme Code numbers and Gene Ontology annotations using the computer software Blast2Go (Conesa et al. 2005) against the NCBI RefSeq protein database with stringent cut-offs (e value $< 10^{-5}$). Annotations were curated using the Artemis software package (Carver et al. 2008) and used to create a database for upload into the Pathway Tools software suite (Karp et al. 2010). Automated reaction assignment and metabolic modeling was done using the pathologic module of the Pathway Tools package, using the MetaCyc database (accessed September 2012) as template. The '*Ca. B. kirkii*' pathway tools database was subsequently manually curated for better accuracy.

RNA extraction and RT-PCR analysis.

Plant samples were washed for 1 min in 70% ethanol, followed by rinsing several times in sterile water. Nodules were dissected from mature source leaves using a Harris Uni-core 2.0-mm core-punching tool (Whatman Inc., Kent, U.K.). Shoot tips were dissected with a scalpel and young emerging leaves were removed so that the leaf primordia and the shoot apical meristematic zone remained. Dissected tissue samples were stored immediately in RNAlater solution (Life Technologies Europe, Zug, Switzerland) and kept at 4°C for 24 h. Samples were washed in 10% RNAlater solution and ground with a pestle in sterile 1.5-ml microcentrifuge tubes.

RNA was then extracted according to standard hot phenol extraction protocol (Ares 2012). Contaminating DNA was removed from RNA preparations using the DNA-free kit (Life Technologies Europe). Approximately 100 ng was used for first-strand cDNA synthesis using the RevertAid H Minus First Strand cDNA Synthesis Kit (Thermo Scientific, Waltham, MA, U.S.A.) at 50°C with specific oligonucleotide primers c149-rev 5'-ATCACCGTTCACTTCTC, rpoD-rev 5'-ACCATCGCTGACATTCTC, and recA-rev 5'-ATGAAGATGACGAGGCAG. RT control reactions were prepared by omitting the oligonucleotides in the reaction mix. The cDNA were analyzed by quantitative (q)PCR using the GoTaq qPCR SYBR Green MasterMix kit (Promega Corp.) in an Mx3000P QPCR system (Agilent Technologies, Inc., Santa Clara, CA, U.S.A.). The PCR primers used for specific cDNA detection were, in addition to the oligonucleotides cited above, c149-fwd 5'-CGCCGTCTTGTTTTT GTC, rpoD-fwd 5'-CCTCTTCTTCTCTTCTTCT, and recA-fwd 5'-GAAGGCGGAAATCGAAGG. Data were analyzed using the Agilent MxPro software, and PCR amplification cycle threshold values were fitted on a standard curve prepared with cDNA standards. Relative quantification values were calculated using the efficiency corrected comparative quantitation module of the Agilent MxPro software using BKIR_c149_4879 values using *recA* and *rpoD* as reference genes.

ACKNOWLEDGMENTS

We thank A. Grunau and B. Roschitzki (Functional Genomics Center Zürich) for their help with protein samples preparation and chroma-

tography and G. Pessi for her help with RNA extraction and RT-PCR techniques. This work has been supported by the URPP Evolution in Action of the University of Zürich. C. H. Ahrens was supported by the Swiss National Science Foundation (SNF) under grant 31003A_130723.

LITERATURE CITED

- Ansong, C., Tolić, N., Purvine, S. O., Porwollik, S., Jones, M., Yoon, H., Payne, S. H., Martin, J. L., Burnet, M. C., Monroe, M. E., Venepally, P., Smith, R. D., Peterson, S. N., Heffron, F., McClelland, M., and Adkins, J. N. 2011. Experimental annotation of post-translational features and translated coding regions in the pathogen *Salmonella typhimurium*. *BMC Genomics* 12:433.
- Ares, M. 2012. Bacterial RNA isolation. *Cold Spring Harb. Protocols* 2012:1024-1027.
- Armengaud, J. 2009. A perfect genome annotation is within reach with the proteomics and genomics alliance. *Curr. Opin. Microbiol.* 12:292-300.
- Bai, L., Li, L., Xu, H., Minagawa, K., Yu, Y., Zhang, Y., Zhou, X., Floss, H. G., Mahmud, T., and Deng, Z. C.-P. 2006. Functional analysis of the validamycin biosynthetic gene cluster and engineered production of validoxylamine A. *Chem. Biol.* 13:387-397.
- Brosch, M., Yu, L., Hubbard, T., and Choudhary, J. 2009. Accurate and sensitive peptide identification with Mascot Percolator. *J. Proteome. Res.* 8:3176-181.
- Brunner, E., Ahrens, C. H., Mohanty, S., Baetschmann, H., Loevenich, S., Potthast, F., Deutsch, E. W., Panse, C., de Lichtenberg, U., Rinner, O., Lee, H., Pedrioli, P. G. A., Malmström, J., Koehler, K., Schrimpf, S., Krijgsvelde, J., Kregenow, F., Heck, A. J. R., Hafen, E., Schlapbach, R., and Aebersold, R. 2007. A high-quality catalog of the *Drosophila melanogaster* proteome. *Nat. Biotech.* 25:576-583.
- Carlier, A. L., and Eberl, L. 2012. The eroded genome of a *Psychotria* leaf symbiont: Hypotheses about lifestyle and interactions with its plant host. *Environ. Microbiol.* 14:2757-2769.
- Carver, T., Berriman, M., Tivey, A., Patel, C., Böhme, U., Barrell, B. G., Parkhill, J., and Rajandream, M. A. C.-P. 2008. Artemis and ACT: Viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* 24:2672-2676.
- Castellana, N., and Bafna, V. 2010. Proteogenomics to discover the full coding content of genomes: A computational perspective. *J. Proteomics* 73:2124-2135.
- Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M., Bafna, V., and Briggs, S. P. 2008. Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proc. Natl. Acad. Sci. U.S.A.* 105:21034-1038.
- Centifanto, Y. M., and Silver, W. S. 1964. Leaf-nodule symbiosis. I. Endophyte of *Psychotria bacteriophila*. *J. Bacteriol.* 88:776-781.
- Christie-Oleza, J. A., Miotello, G., and Armengaud, J. 2012. High-throughput proteogenomics of *Ruegeria pomeroyi*: Seeding a better genomic annotation for the whole marine Roseobacter clade. *BMC Genomics* 13:73.
- Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674-3676.
- Delmotte, N., Ahrens, C. H., Knief, C., Qeli, E., Koch, M., Fischer, H.-M., Vorholt, J. A., Hennecke, H., and Pessi, G. 2010. An integrated proteomics and transcriptomics reference data set provides new insights into the *Bradyrhizobium japonicum* bacteroid metabolism in soybean root nodules. *Proteomics* 10:1391-1400.
- De Souza, G. A., Søfteland, T., Koehler, C. J., Thiede, B., and Wiker, H. G. 2009. Validating divergent ORF annotation of the *Mycobacterium leprae* genome through a full translation data set and peptide identification by tandem mass spectrometry. *Proteomics* 9:3233-243.
- Edwards, W. J., and Lamotte, C. E. C.-P. 1975. Evidence for cytokinin in bacterial leaf nodules of *Psychotria punctata* (Rubiaceae). *Plant Physiol.* 56:425-428.
- Elias, J. E., and Gygi, S. P. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 4:207-214.
- Gupta, N., Tanner, S., Jaitly, N., Adkins, J. N., Lipton, M., Edwards, R., Romine, M., Osterman, A., Bafna, V., Smith, R. D., and Pevzner, P. A. 2007. Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. *Genome Res.* 17:1362-1377.
- Karp, P. D., Paley, S. M., Krummenacker, M., Latendresse, M., Dale, J. M., Lee, T. J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I. M., and Caspi, R. 2010. Pathway Tools version 13.0: Integrated software for pathway/genome informatics and systems biology. *Brief. Bioinf.* 11:40-79.
- Kristensen, D. B., Brønd, J. C., Nielsen, P. A., Andersen, J. R., Sørensen, O. T., Jørgensen, V., Budin, K., Matthiesen, J., Venø, P., Jespersen, H. M., Ahrens, C. H., Schandorff, S., Ruhoff, P. T., Wisniewski, J. R., Bennett, K. L., and Podtelejnikov, A. V. 2004. Experimental peptide identification repository (EPIR): An integrated peptide-centric platform for validation and mining of tandem mass spectrometry data. *Mol. Cell. Proteomics* 3:1023-1038.
- Lane, G., Christensen, M., and Miles, C. 2000. Coevolution of fungal endophytes with grasses: The significance of secondary metabolites. Pages 341-388 in: *Microbial Endophytes*. C. Bacon and J. White, eds. Marcel Dekker, Inc., New York.
- Lemaire, B., Vandamme, P., Merckx, V., Smets, E., and Dessein, S. 2011. Bacterial leaf symbiosis in angiosperms: Host specificity without co-speciation. *PLoS One* 6:e24430. Published online.
- Lemaire, B., Janssens, S., Smets, E., and Dessein, S. 2012. Endosymbiont transmission mode in bacterial leaf nodulation as revealed by a population genetic study of *Psychotria leptophylla*. *Appl. Environ. Microbiol.* 78:284-287.
- Lemaire, B., Lachenaud, O., Persson, C., Smets, E., and Dessein, S. 2012. Screening for leaf-associated endophytes in the genus *Psychotria* (Rubiaceae). *FEMS (Fed. Eur. Microbiol. Soc.) Microbiol. Ecol.* 81:364-372.
- Mahmud, T. 2003. The C7N aminocyclitol family of natural products. *Nat. Prod. Rep.* 20:137-166.
- Miller, I. M. 1990. Bacterial leaf nodule symbiosis. *Adv. Bot. Res. Adv. Plant Pathol.* 17:163-234.
- Miller, I. M., Scott, A., and Gardner, I. C. 1983. Leaf nodule development in *Psychotria kirkii* Hiern. (Rubiaceae). *Ann. Bot.* 52:791-802.
- Moran, N. A. 2002. Microbial minimalism: Genome reduction in bacterial pathogens. *Cell* 108:583-586.
- Pearson, W. R. 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* 132:185-219.
- Qeli, E., and Ahrens, C. H. 2010. PeptideClassifier for protein inference and targeted quantitative proteomics. *Nat. Biotech.* 28:647-650.
- Ran, L., Larsson, J., Vigil-Stenman, T., Nylander, J. A., Ininbergs, K., Zheng, W. W., Lapidus, A., Lowry, S., Haselkorn, R., and Bergman, B. C.-P. 2010. Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. *PLoS One* 5:e11486. Published online.
- Riedel, K., Carranza, P., Gehrig, P., Potthast, F., and Eberl, L. 2006. Towards the proteome of *Burkholderia cenocepacia* H111: Setting up a 2-DE reference map. *Proteomics* 6:207-216.
- Sachs, J. L., Skophammer, R. G., and Regus, J. U. 2011. Evolutionary transitions in bacterial symbiosis. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl.):10800-10807.
- Saridaki, A., and Bourtzis, K. 2010. Wolbachia: More than just a bug in insects genitals. *Curr. Opin. Microbiol.* 13:67-72.
- Schrimpe-Rutledge, A. C., Jones, M. B., Chauhan, S., Purvine, S. O., Sanford, J. A., Monroe, M. E., Brewer, H. M., Payne, S. H., Ansong, C., Frank, B. C., Smith, R. D., Peterson, S. N., Motin, V. L., and Adkins, J. N. 2012. Comparative omics-driven genome annotation refinement: Application across Yersiniaceae. *PLoS One* 7:e33903. Published online.
- Schrimpf, S. P., Weiss, M., Reiter, L., Ahrens, C. H., Jovanovic, M., Malmström, J., Brunner, E., Mohanty, S., Lercher, M. J., Hunziker, P. E., Aebersold, R., von Mering, C., and Hengartner, M. O. 2009. Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol.* 7:e48. Published online.
- Toyofuku, M., Roschitzki, B., Riedel, K., and Eberl, L. 2012. Identification of proteins associated with the *Pseudomonas aeruginosa* biofilm extracellular matrix. *J. Proteome. Res.* 11:4906-4915.
- Ussery, D. W., Kiil, K., Lagesen, K., Sicheritz-Pontén, T., Bohlin, J., and Wassenaar, T. M. 2009. The genus *Burkholderia*: Analysis of 56 genomic sequences. *Genome Dynamics* 6:140-157.
- Van Oevelen, S., De Wachter, R., Vandamme, P., Robbrecht, E., and Prinsen, E. 2002. Identification of the bacterial endosymbionts in leaf galls of *Psychotria* (Rubiaceae, angiosperms) and proposal of '*Candidatus Burkholderia kirkii*' sp. nov. *Int. J. Syst. Evol. Microbiol.* 52:2023-2027.
- Van Oevelen, S., De Wachter, R., Robbrecht, E., and Prinsen, E. 2003. Induction of a crippled phenotype in *Psychotria* (Rubiaceae) upon loss of the bacterial endophyte. *Bulg. J. Plant Physiol. Spec. Issue* 242-247.
- Van Oevelen, S., De Wachter, R., Vandamme, P., Robbrecht, E., and Prinsen, E. 2004. '*Candidatus Burkholderia calva*' and '*Candidatus Burkholderia nigropunctata*' as leaf gall endosymbionts of African *Psychotria*. *Int. J. Syst. Evol. Microbiol.* 54:2237-2239.
- von Faber, F. C. 1912. Das erbliche Zusammenleben von Bakterien und tropischen Pflanzen. Pages 285-375 in: *Jahrbuch für wissenschaftliche Botanik*. Borntraeger, Berlin.
- Wehmeier, U. F., and Piepersberg, W. 2004. Biotechnology and molecular biology of the alpha-glucosidase inhibitor acarbose. *Appl. Microbiol. Biotechnol.* 63:613-625.

- Whitmoyer, R. E., and Horner, H. T. 1970. Developmental aspects of bacterial leaf nodules in *Psychotria bacteriophila* Val. (Rubiaceae). Bot. Gaz. 3:193-200.
- Wongtrakoonkate, P., Roytrakul, S., Yasothornsrikul, S., and Tungpradabkul, S. 2011. A proteome reference map of the causative agent of melioidosis *Burkholderia pseudomallei*. J. Biomed. Biotechnol. 2011: 530926.
- Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J., and Brinkman, F. S. L. 2010. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 26:1608-1615.
- Yu, Y., Bai, L., Minagawa, K., Jian, X., Li, L., Li, J., Chen, S., Cao, E., Mahmud, T., Floss, H. G., Zhou, X., and Deng, Z. C.-P. 2005. Gene cluster responsible for validamycin biosynthesis in *Streptomyces hygroscopicus* subsp. *jinggangensis* 5008. Appl. Environ. Microbiol. 71:5066-5076.
- Zhang, Z., Carriero, N., Zheng, D., Karro, J., Harrison, P. M., and Gerstein, M. 2006. PseudoPipe: An automated pseudogene identification pipeline. Bioinformatics 22:1437-1439.
- Zhao, L., Liu, L., Leng, W., Wei, C., and Jin, Q. 2011. A proteogenomic analysis of *Shigella flexneri* using 2D LC-MALDI TOF/TOF. BMC Genomics 12:528.
- Zimmermann, A. 1902. Über Bakterienknoten in den Blättern einiger Rubiaceen. Pages 1-11 in: Jahrbuch für wissenschaftliche Botanik. Borntraeger, Berlin.

AUTHOR-RECOMMENDED INTERNET RESOURCE

EMBL-EBI and the Wellcome Trust Sanger Institute joint Ensembl database: www.ensembl.org