SCIENTIFIC REPORTS

OPEN

Discovery of Novel Plant Interaction Determinants from the Genomes of 163 Root Nodule Bacteria

Received: 07 September 2015
Accepted: 20 October 2015
Published: 20 November 2015

Rekha Seshadri[1], Wayne G. Reeve[2], Julie K. Ardley[2], Kristin Tennessen[1], Tanja Woyke[1], Nikos C. Kyrpides[1,3] & Natalia N. Ivanova[1]

Root nodule bacteria (RNB) or "rhizobia" are a type of plant growth promoting bacteria, typified by their ability to fix nitrogen for their plant host, fixing nearly 65% of the nitrogen currently utilized in sustainable agricultural production of legume crops and pastures. In this study, we sequenced the genomes of 110 RNB from diverse hosts and biogeographical regions, and undertook a global exploration of all available RNB genera with the aim of identifying novel genetic determinants of symbiotic association and plant growth promotion. Specifically, we performed a subtractive comparative analysis with non-RNB genomes, employed relevant transcriptomic data, and leveraged phylogenetic distribution patterns and sequence signatures based on known precepts of symbiotic- and host-microbe interactions. A total of 184 protein families were delineated, including known factors for nodulation and nitrogen fixation, and candidates with previously unexplored functions, for which a role in host-interaction, -regulation, biocontrol, and more, could be posited. These analyses expand our knowledge of the RNB purview and provide novel targets for strain improvement in the ultimate quest to enhance plant productivity and agricultural sustainability.

The use of plant growth promoting bacteria to enhance crop yield and control disease is gaining worldwide acceptance as a sustainable agricultural practice, while reducing costs by supplanting the use of expensive (and polluting) agrochemicals. These bacteria can facilitate plant growth either directly, by providing essential nutrients (nitrogen, phosphorus and essential minerals), modulating plant hormones and development, or indirectly, by suppressing inhibitory effects of various plant pathogens, improving soil structure and bioremediating polluted soils[1,2]. In particular, root nodule bacteria (RNB) are free-living soil bacteria that have the ability to form nitrogen-fixing symbioses with legumes, and have been exploited for centuries to improve soil fertility and agricultural productivity[3]. The symbiosis is typically host-specific (although more promiscuous strains exist) and mediated by signaling molecules produced by both plant host and the bacterium[4]. RNB convert inert atmospheric nitrogen gas into bioavailable ammonia for their host in exchange for carbon (and shelter) within specialized root or stem nodules, resulting in improved plant growth and productivity[5].

The legume-RNB symbiosis is one of the best-studied associations between bacteria and eukarya due to both ecological and economic importance. It is estimated that increasing the efficiency of symbiotic nitrogen fixation (SNF) may have an annual benefit of $1,067 million in the U.S alone, while total elimination of nitrogen fertilization of major crops would have an annual benefit of $4,484 million[6]. Additionally, SNF reduces greenhouse gas emissions by displacing 873 m^3 of natural gas and the ultimate

[1]Department of Energy Joint Genome Institute, Walnut Creek, USA. [2]Centre for Rhizobium Studies, School of Veterinary and Life Sciences, Murdoch University, Murdoch 6150, Australia. [3]Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudia Arabia. Correspondence and requests for materials should be addressed to R.S. (email: rseshadri@lbl.gov)

release of ~2 tons of CO_2[7] in the manufacture every ton of conventional nitrogenous fertilizer, as well as reducing annual nitrous oxide emissions and NO_3^- in surface runoff. Other benefits to the environment include reducing dryland salinity, increasing soil fertility, promoting carbon sequestration and preventing eutrophication of water bodies. Furthermore, RNB play a role in the production of biofuel crops–*Millettia pinnata*, for example, is a leguminous tree nodulated by *Bradyrhizobium* and *Rhizobium* spp. that produces biodiesel, starch, ethanol and biogas[8]. With a burgeoning world population and increasing food demands, harnessing the innate potential of RNB to improve sustainable agricultural productivity is of paramount importance.

However, despite these significant environmental and economic incentives, only a few genomes of a phylogenetically restricted group of model RNB strains had been sequenced at the inception of this study. These strains were mostly laboratory "work horses", whereas sequencing of commercial inoculants that have the highly prized attributes of survival and persistence in soil, competitiveness, and high rates of N_2-fixation, had not been a priority. In addition, preliminary analyses focused on questions pertaining to genome evolution and structure, intra-genus conservation and physiological diversity[9–11]. A more recent paper surveyed the occurrence of **known** plant growth promotion genes in all available proteobacterial genomes[12], and clearly many RNB were found to possess plant growth promotion traits beyond nitrogen fixation, but little had been done to explore novel effectors of plant growth or even the accessory factors mediating RNB-plant interactions (including symbiosis). Thus, the primary objective of our study was to (i) increase the repertoire of available RNB genomes in terms of their phylogenetic, biogeographic and host legume diversity, and (ii) identify novel microbial effectors of symbiosis, and plant growth and productivity, beyond what is currently known about nodulation and nitrogen fixation. We sequenced the genomes of 110 RNB isolates sourced from a variety of leguminous hosts from diverse biogeographical locations, performed a comprehensive analysis of all RNB genera, and identified novel determinants of plant interaction and growth. These data not only provide a resource and conceptual framework for studying RNB-legume interactions, but our results highlight many new potential plant beneficial genes that could be targeted to improve legume productivity around the globe.

Results and Discussion

Overview of the Project. This study falls under the auspices of the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project, which was conceived to maximize the phylogenetic coverage of publicly available prokaryotic genomes[13–15]. Correspondingly, the GEBA-RNB sub-project was designed to capture RNB phylogenetic and symbiotic diversity, with the participation of an international consortium consisting of more than 30 experts in the field, from 15 different countries, and major culture collection centers in Australia, Belgium and the USA[14]. RNB strains were selected on the basis of (i) phylogenetic diversity, (ii) host legume diversity (spanning all the Vavilov centers of origin[16]) (iii) economic or commercial significance and, (iv) biogeographic origin (Fig. 1). Strains were also required to have comprehensive experimental and metadata records and well-characterized phenotypes, in particular, relating to symbiotic efficiency and host specificity. Biogeographic considerations were relevant as RNB survival and persistence as soil saprophytes is governed by environmental and edaphic constraints such as pH, temperature, salinity, soil moisture- and clay content. The RNB were therefore collected from sites that spanned a broad range of soils (varying pH, salinity) and climates (e.g. tropical, arid, temperate). Chosen RNB also varied in their physiological traits (e.g., ability to recycle hydrogen, methylotrophy, salt or acid tolerance, rhizobitoxine production, heavy metal resistance, etc.) and host specificities (ranging from strictly specific to highly promiscuous). Moreover, each sequenced RNB strain has been cryopreserved in a dedicated long term storage culture collection and is available to the global research community by request through the Centre for Rhizobium Studies (CRS).

To summarize, 110 RNB isolates from 70 diverse legume hosts from various biomes in over 30 countries were sequenced by us, and an additional 50 genomes were released to Genbank during the course of this study, resulting in a total of 163 RNB genomes analyzed here (Fig. 1). All major RNB lineages were represented with the overwhelming majority (145 genomes) belonging to seven genera within the Order Rhizobiales of Class α-proteobacteria, and 18 genomes belong to two genera from Class β-proteobacteria. The complete list of RNB genomes, metrics and metadata is presented in Supplementary Table 1. General assembly and annotation metrics are presented in Supplementary Table 2.

Predicting Novel Effectors of RNB-plant interaction. The RNB-legume symbiosis has been long-heralded as an excellent model for investigating plant-microbe associations; however, few studies have attempted to venture beyond describing the mechanisms and underpinnings of nodulation and N_2-fixation, the hallmark ability of RNB. With few exceptions, auxiliary functions that are undoubtedly necessary to colonize, communicate or interact with their plant host, and possibly regulate plant development, are largely anonymous. It is also evident that many RNB possess capacities well beyond bio-fertilization through N_2-fixation - for example, 1-aminocyclopropane-1-carboxylate (ACC) deaminase (TIGR01274), known to modulate plant development by reducing levels of the plant stress hormone, ethylene[17], is almost ubiquitously present in all the sequenced RNB genera with the exception of several strains of *Ensifer* spp. and *Rhizobium* spp. Furthermore, the introduction of an exogenous ACC

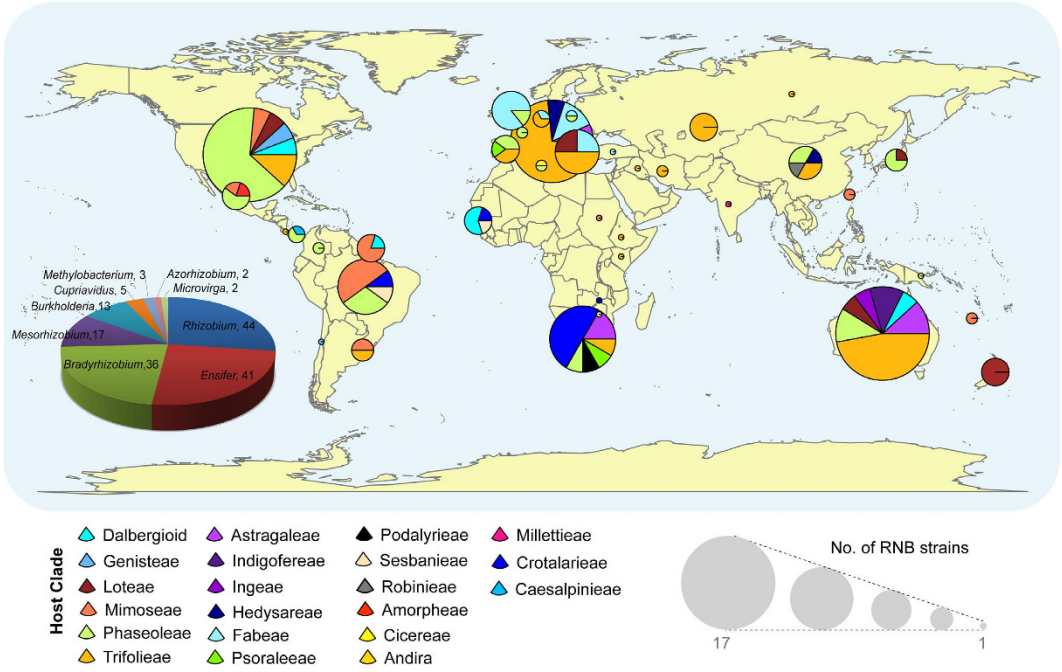

Figure 1. Summary of biogeography and taxonomy of 163 RNB strains analyzed in this study. Biogeographic information is depicted as world map overlaid with pie charts showing the legume clade of the plant host for strains originating from that geographic location; size of the pie scales to the total number of RNB strains from that location (ranging from 1 to 17 strains). Taxonomic composition of 163 RNB strains is shown as a separate pie chart in the bottom left side of the figure. See Supplementary Tables 1 and 2 for genome statistics and metadata details. Figure was generated using the R package "maps" (Brownrigg, R., Minka, T. P., Becker, R. A. & Wilks, A. R. maps: Draw Geographical Maps. R package version 2.1-5. http://CRAN.R-project.org/package=maps (2010), and edited further using Adobe Illustrator.

deaminase gene into the laboratory strain *Ensifer meliloti* Rm 1021, which lacks this gene, increased biomass and nodulation of host *Medicago sativa* (alfalfa or Lucerne) by over 40%[18].

A multi-step strategy was therefore devised to help identify novel plant-beneficial determinants from the 163 RNB genomes (Fig. 2), involving a subtractive comparative analysis with non-RNB genomes, leveraging relevant transcriptomic data for substantiation, and employing additional filters such as phylogenetic occurrence and sequence signatures based on known precepts of symbiotic and plant-microbe interactions. The first step identified functions that are over-represented or enriched in the RNB genomes set compared to a "negative control" (NC) genome set. To minimize the identification of false positives, the NC members were carefully selected from available genomes of phylogenetically-related organisms that are not known to be associated with the phytosphere (e.g., rhizosphere, phyllosphere) environment (based on available metadata for genomes from the GOLD database[19]). This resulted in a NC set containing 69 genomes from 35 genera belonging to either Order Rhizobiales of Class α-proteobacteria, or Class β-proteobacteria (Supplementary Table 3). To the best of our knowledge, these 69 isolates originated from a variety of aquatic, terrestrial and few host-associated habitats, and are not typically associated with the phytosphere, based on available GOLD metadata entries and published literature.

Next, gene counts for each protein family (Pfam) domain were retrieved and contrasted between the RNB and NC genomes. Pfam was chosen for this analysis because it is the largest and most widely used collection of manually-curated protein families[20], with >80% coverage (on average) of total CDS in these microbial genomes.

A primary approach to identify Pfams that were "over-represented" in the RNB involved contrasting median or upper and lower quartile gene counts between the RNB and NC genomes. A total of 437 Pfams (out of 4896 Pfam domains recruited by 163 RNB genomes) could be delineated based on a total RNB median (2^{nd} quartile) gene count of ≥ 1 and a total NC median gene count of 0 (Supplementary Table 4). Encouragingly, many core nodulation and nitrogen fixation-related Pfams (e.g., NodA (PF02474), NifD (PF00148), NifK (PF11844)) were identified and served to validate our approach (core components not represented could be attributed to the absence of a suitable or specific Pfam domain for that particular function (e.g., NifH)). For many others, a role in plant host interaction, biocontrol, stress tolerance, or more could be posited (discussed below), however, this approach also yielded 123 Pfams containing domains of unknown function (DUFs), which may also be important in legume-RNB interactions.

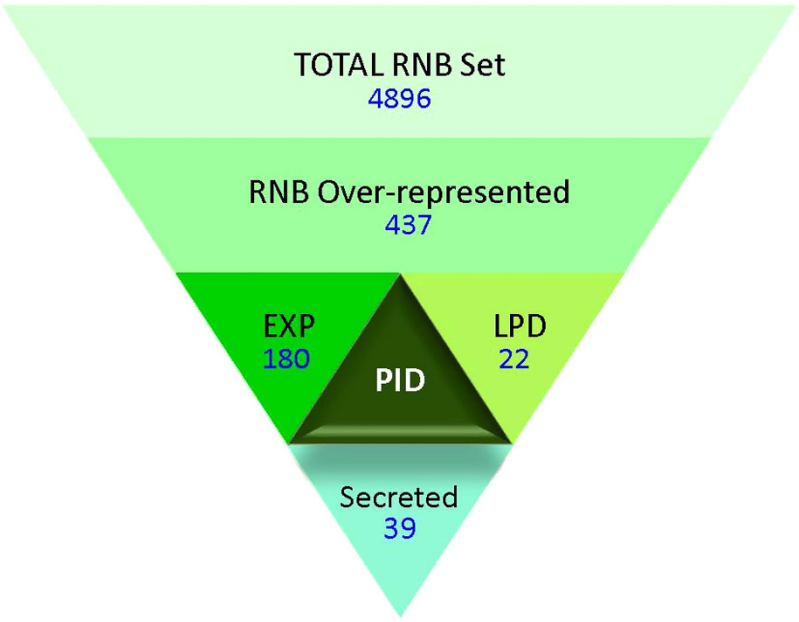

Figure 2. Delineating novel plant interaction determinants from 163 RNB genomes. Overall strategy devised to filter and identify protein family domains associated with plant interaction or growth promotion from 163 RNB genomes. LPD – limited phylogenetic distribution, EXP - upregulation or induction in published transcriptomic studies[21–24], PID – plant interaction determinants, "Secreted" refers to Pfam candidates that bear a signal peptide for possible secretion into the external milieu. See Supplementary Table 4 for list with details.

To garner further support for a proposed role for these over-represented Pfams in phytosphere interactions, corroboration by transcript expression (designated "EXP") under relevant experimental conditions was used as a key filtering step (Fig. 2). Out of 437 over-represented Pfams, 180 had a single candidate gene that showed upregulation or induction in one or more of four published RNB transcriptome studies of symbiotic nitrogen fixation[21–24]. For example, PhoD-like phosphatase family (PF09423) candidates were induced >64X in the symbiosome, in two independent RNA-seq-based transcriptome studies (*Ensifer fredii* NGR_c31990 and *E. meliloti* SM11_chr3272)[21,23]. It has previously been shown that plants obtaining nitrogen from symbiosis require higher levels of phosphorus for optimal growth than do plants grown with nitrogen fertilizers[25], and it is tempting to speculate a role in inorganic phosphate-solubilization to enable nodule bioavailability. Another example is CopC domain (PF04234) candidates that showed at least 16X induction in both RNA-seq experiments (*E. fredii* NGR_b06130 and *E. meliloti* SM11_pC0976). In the phytopathogen *Pseudomonas syringae*, CopC has been implicated in mediating copper resistance by binding and sequestering copper in the periplasm[26] and is believed to function in copper trafficking into cells[27,28]. Copper is a cofactor of the high-affinity *cbb3*-type (heme-copper sub-family) cytochrome oxidase, encoded by the *fixNOQP* operon, that terminates the symbiosis-specific respiratory chain of rhizobia[29,30] and CopC may play a role in trafficking copper to *cbb3*-type cytochrome oxidases. In this regard, it is interesting to note that in *E. meliloti* genomes, the locus encoding the CopC domain protein is found downstream of a pSym cluster of *fix* genes including *fixNOQP* and *fixGHISK*. A caveat of employing the EXP criterion is that false negatives are likely because of possible limitations of the method, such as inability to detect significant changes in genes with typically low levels of expression such as regulators.

Within the over-represented subset, we also observed Pfams with primarily eukaryal origin, i. e., the majority of all known sequences that are assigned to these Pfams originated from eukaryal genomes. This led to the hypothesis that RNB may produce eukarya-like factors in order to better interact with, or modulate plant host responses. Indeed, the notion of horizontal gene transfer from a plant host to its bacterial resident has been previously explored[31,32], although merely speculative here. We systematically looked for Pfams of primarily eukaryal origin within the RNB-over-represented set and identified 5 Pfam domains. However, to factor in a previously described observation of shared mechanisms of virulence between plant and animal pathogens, or shared strategies for infection and adaptation to growth within the eukaryal host between pathogens and symbionts[33,34], we extended the search beyond simply "eukaryal origin" to those recruiting sequences from a limited group of prokaryotic lineages (such as primarily Family Rhizobiaceae, or including alpha-proteobacterial pathogens (e.g., *Brucella* spp. and *Bartonella* spp.). This additional data filtering criterion was designated "LPD" for limited phylogenetic distribution (Fig. 2).

Intriguing examples of these LPD Pfams (Supplementary Table 4) that were over-represented in the RNB included the sterile alpha motif (SAM) domain (PF00536), PAN domain (PF00024), and many DUFs (e.g., PF06191). SAM domains are described as protein interaction modules involved in developmental processes in diverse eukarya. Almost 80% of the total number of sequences assigned to this Pfam within the IMG database belonged to Domain Eukarya. RNB SAM candidates are large multidomain proteins predominantly associated with adenylate guanylate cyclase (PF00211) and AAA ATPase (PF13191) domains, and a role in modulating host response seems likely. The PAN domain (PF00024) has a described role in mediating protein-protein or protein-carbohydrate interactions in eukarya. Again, many RNB PAN domain candidates are large multi-domain sequences associated with alpha-2-macroglobulin domains (PF11974, PF01835, PF07703, PF00207, PF10569) which were themselves not over-represented in the RNB set, however, we speculate the presence of the N-terminal PAN domain (perhaps due to a gene fusion event) may confer a new utility or specificity relevant to RNB activity within its eukaryal host. PAN candidates from eukaryal parasites (*Toxoplasma gondii, Sarcocystis muris*) are characterized as lectin adhesins mediating host binding and invasion by such parasites[35,36]. A similar role may be proposed for the RNB PAN domain candidates; this is supported by the presence of a classical signal peptidase I cleavage site (predicted by SignalP) for secretion in most instances.

The third obvious data-filtering criterion was therefore to determine if over-represented Pfam candidates were potentially secreted (see Methods), in order to identify products that may specifically interact with host cell components. Out of the list of 437 over-represented Pfams, 39 had a major proportion (≧50%) of assigned sequences bearing a SignalP motif. For Pfams not meeting this criterion, it is still possible that candidates may be translocated directly into the plant milieu by one of many protein secretion systems (e.g., Type III, Type IV) encoded by the RNB. Only 3 Pfam domains met all three criteria (EXP, LPD, Secretion), while 30 Pfams (out of the over-represented 437) satisfied two or more. Overall, we put forward 184 Pfam domains fulfilling one or more criteria to be designated as determinants of plant interaction or "PID" (Fig. 2, Supplementary Table 4). A proposed role for these can encompass phytostimulation, biocontrol, rhizosphere competence, stress tolerance, in addition to nodulation (including symbiotic signaling, triggering endocytosis and host cell differentiation) and nitrogen fixation. Out of 184 PID Pfams, at least 10 were clearly associated with nodulation and nitrogen fixation, about 8 Pfams may be assigned a regulatory role, and 74 Pfams were implicated in secondary metabolite degradation or synthesis (based on cross-referencing with Pfams used by AntiSMASH[37] for secondary metabolite operon identification). For example, a role in the hydrolysis of host secondary metabolites is proposed for Epoxide hydrolase N terminus (PF06441) candidates that are mostly predicted to be secreted, however, a role in synthesis may be possible in a few instances where co-localized polyketide biosynthesis genes are present (e.g., *Mesorhizobium loti* USDA 3471). A second example is the berberine-like domain (PF08031), with limited phylogenetic distribution, which typically is found in plant proteins that are involved in the synthesis of isoquinoline alkaloids as a pathogen defense response[38]. Plant berberine bridge enzymes are highly induced during various defense responses, when they may contribute to the oxidative burst leading to cell death, through H_2O_2 synthesis; intriguingly, they have been found in the secretome of the phytopathogen *Phytophthora infestans,* where they have been postulated to play a role in virulence[39]. In many RNB strains, the presence of an upstream decarboxylase gene with a putative role in alkaloid synthesis[40,41] favors a role in secondary metabolite biosynthesis, therefore a role in biocontrol of plant pathogens may be conjectured. In instances where domain candidates appear without a co-localized cognate function, a role in the degradation of plant alkaloids may be hypothesized, as seen in *Arthrobacter* spp.[42] In general, the biological activity of secondary metabolites produced by plant growth promoting bacteria ranges from antimicrobial to phytostimulatory[43,44].

A large number of the 184 PID Pfams were DUFs - DUF2950 (PF11453) and DUF3300 (PF11737) were particularly compelling examples that showed significant induction in both RNA-seq transcriptome studies[21,23], and possessed signal peptides for putative secretion (Supplementary Table 4). Furthermore, the CDS encoding these domains were universally co-localized (DUF3300 is upstream of DUF2950), suggesting cognate function. No other hints regarding their function could be gleaned based on gene neighborhood, and no characterized members were found in public databases.

Many PID candidates were also clearly arranged in discrete operons (Fig. 3) – predictably, the nitrogen fixation Pfams were co-localized in a large operon (Fig. 3a). Another notable PID operon was involved in phenyl acetic acid synthesis (Fig. 3b), which is a potential phytohormone (auxinomimetic) with a demonstrated role in nodulation and regulation of nitrogen fixation by *Frankia* spp.[45], or may function as an antimicrobial agent as suggested for *Azospirillum brasilense*[46]. Other PID operons have more cryptic functions, for example, the operon depicted in Fig. 3c is comprised of a serine protein kinase, a DUF and an unknown gene tenuously associated with sporulation (SpoVR). Certainly bacterial serine kinases have been previously implicated in mediating host-pathogen interactions[47], and a similar role in mediating plant interaction may be suggested for this operon. Constituents of the operon depicted in Fig. 3d appear to be mostly involved in carbohydrate metabolism and a role in secondary metabolite synthesis is suspected.

In addition to the strategy presented above, we also employed a statistical analysis using Fisher's Exact Test on pairwise comparisons of gene counts for each Pfam from 163 RNB genomes versus 69 non-phytosphere-associated control or NC genomes described above (see Methods for details). A total of 19 Pfam domains were significantly different in at least 25% of the pairings with a Benjamin-Hochberg

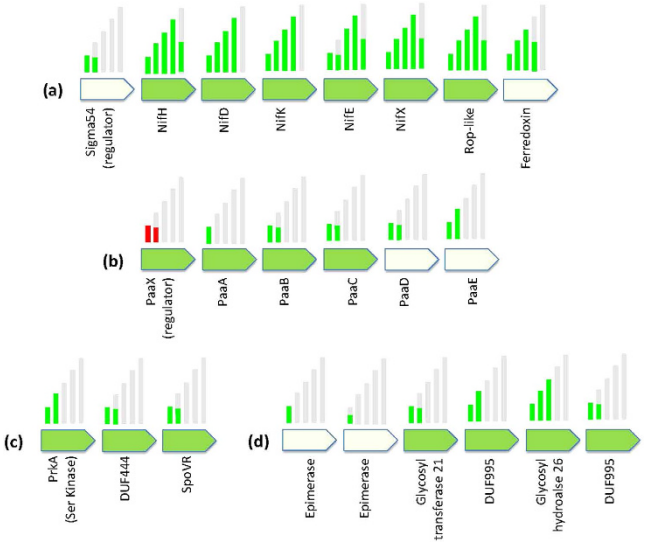

Figure 3. Operons encoding putative determinants of plant interaction. Examples of PID Pfam candidates (colored green) that are co-localized in a putative operon in *E. meliloti* strain 2011. Log_2 fold increase of transcript in the *E. meliloti* transcriptome experiment[21] is shown as green bars above the CDS, ranging from 2 to 10. Red bar denotes decrease in transcript abundance in tested condition. *Depicted lengths of coding sequences are not to scale.*

P-Value correction cutoff of <0.05 (Supplementary Table 5). Almost all of these significantly "enriched" domains were non-overlapping with the above discussed "over-represented" set because this method teases out domains that contain increased functional potential within the RNB, and likely results from lifestyle-specific expansions of gene families in the RNB compared to the NC genomes. The majority of the domains captured in the enriched set were associated with either transporter or regulatory functions. A role in microbe-host interactions may be postulated for many - for example, a PF13407 (periplasmic binding protein domain) candidate in *E. meliloti* (MocB) is involved in the uptake and degradation of a nodule-specific compound, rhizopine, which plays an important role in symbiosis[48]. Also, many of the candidate genes for this Pfam - 26 out of 38 PF13407 candidates show >2X induction in *Ensifer meliloti* transcriptome experiments[21]. Similarly, PF00211 (adenylate guanylate cyclase domain) candidates show significant induction in transcriptomic studies[21,23] – these genes are involved in the formation of the secondary messengers, cAMP and cGMP[49], which are triggered by an unknown plant host signal, and involved in establishing infection and symbiosis[50–53]. Secondary messenger signaling plays a major role in coordinating virulence gene expression in animal pathogens, as well as suppressing eukaryal host immune responses[54]. A list of these relatively-enriched candidate Pfams is presented in Supplementary Table 5.

In conclusion, we have performed an all-inclusive sequencing and analysis of RNB genomes across taxonomic genera, plant host types and biogeographical origins, providing an important scientific resource, and identifying a novel repertoire of determinants of the RNB lifestyle, including those with potential plant beneficial effects. We anticipate that a more comprehensive understanding of these mechanisms will aid the quest to extend N_2-fixation to non-legume crops, a goal described as essential for future sustainable food production[55]. An additional outcome is a furthering of our appreciation of the role of the RNB in plant growth promotion beyond its known biofertilization effects. Indeed, experimental validation including quantifying the production and relative contribution of these putative effectors of plant growth, relative to the effects of nitrogen fixation, will be very enlightening.

Methods

Sequencing, assembly, annotation. The draft genomes of RNB strains were generated at the DOE Joint Genome Institute (JGI) using Illumina technology[56]. For all genomes, we constructed and sequenced an Illumina short-insert paired-end library with an average insert size of 270 bp. All general aspects of library construction and sequencing performed at the JGI can be found at the JGI website (http://www.jgi.doe.gov). The details of sequencing and assembly of individual genomes are reported in the respective genome publications in the Standards in Genome Sciences (http://www.standardsingenomics.org/index.php/sigen/index).

Genomes were annotated by the DOE-JGI genome annotation pipeline[57]. Briefly, protein-coding genes (CDSs) were identified using Prodigal[58], followed by a round of automated and manual curation using the JGI GenePrimp pipeline[59]. Non-coding genes and miscellaneous features were predicted using tRNAscan-SE[60], RNAMMer[61], Rfam[62]. The predicted CDSs were translated and transmembrane

regions and signal peptides were predicted using TMHMM[63], and SignalP[64]. Functional annotation and additional analyses were performed within the Integrated Microbial Genomes (IMG-ER) platform (https://img.jgi.doe.gov/cgi-bin/mer/main.cgi)[65] including searches against IMG non-redundant database, UniProt, TIGRFam, Pfam, PRIAM, KEGG, COG, and InterPro databases. These data sources were combined to assert a product description for each predicted protein. General genome assembly and annotation statistics are presented in Supplementary Table 2. All available genomic data and annotations may be accessed through the IMG portal (https://img.jgi.doe.gov/cgi-bin/mer/main.cgi).

Statistical analysis of significantly different or "enriched" Pfams in RNB versus control sets. Pairwise comparisons of gene counts for each Pfam for 163 RNB genomes versus 69 control genomes ($163 * 69 = 11{,}247$ comparisons) were performed. Using Fisher's Exact Test, Pfams that were significantly different in at least 25% of the pairings with a Benjamin-Hochberg P-Value correction cut-off of <0.05 were identified. Normalization of gene counts was determined to be unnecessary since no consistent correlation between number of Pfam hits per genome and genome size were found. Results of this pairwise comparison are presented in Supplementary Table 5.

References

1. Glick, B. R. Plant growth-promoting bacteria: mechanisms and applications. *Scientifica* **2012,** 963401, doi: 10.6064/2012/963401 (2012).
2. Lugtenberg, B. & Kamilova, F. Plant-growth-promoting rhizobacteria. *Annual review of microbiology* **63,** 541–556, doi: 10.1146/annurev.micro.62.081307.162918 (2009).
3. Howieson, J. G., O'Hara, G. W. & Carr, S. J. Changing roles for legumes in Mediterranean agriculture: developments from an Australian perspective. *Field Crop Res* **65,** 107–122, doi: 10.1016/S0378-4290(99)00081-7 (2000).
4. Oldroyd, G. E. D., Murray, J. D., Poole, P. S. & Downie, J. A. The Rules of Engagement in the Legume-Rhizobial Symbiosis. *Annu Rev Genet* **45,** 119–144, doi: 10.1146/annurev-genet-110410-132549 (2011).
5. Lindstrom, K., Murwira, M., Willems, A. & Altier, N. The biodiversity of beneficial microbe-host mutualism: the case of rhizobia. *Research in microbiology* **161,** 453–463, doi: 10.1016/j.resmic.2010.05.005 (2010).
6. Tauer, L. W. Economic-Impact of Future Biological Nitrogen-Fixation Technologies on United-States Agriculture. *Plant Soil* **119,** 261–270, doi: 10.1007/Bf02370418 (1989).
7. Vance, C. P. Symbiotic nitrogen fixation and phosphorus acquisition. Plant nutrition in a world of declining renewable resources. *Plant physiology* **127,** 390–397 (2001).
8. Rasul, A., Amalraj, E. L., Praveen Kumar, G., Grover, M. & Venkateswarlu, B. Characterization of rhizobial isolates nodulating Millettia pinnata in India. *FEMS microbiology letters* **336,** 148–158, doi: 10.1111/1574-6968.12001 (2012).
9. Tian, C. F. *et al.* Comparative genomics of rhizobia nodulating soybean suggests extensive recruitment of lineage-specific genes in adaptations. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 8629–8634, doi: 10.1073/pnas.1120436109 (2012).
10. Sugawara, M. *et al.* Comparative genomics of the core and accessory genomes of 48 Sinorhizobium strains comprising five genospecies. *Genome biology* **14,** R17, doi: 10.1186/gb-2013-14-2-r17 (2013).
11. Kumar, N. *et al.* Bacterial genospecies that are not ecologically coherent: population genomics of Rhizobium leguminosarum. *Open Biol* **5,** doi: Unsp 140133 doi: 10.1098/Rsob.140133 (2015).
12. Bruto, M., Prigent-Combaret, C., Muller, D. & Moenne-Loccoz, Y. Analysis of genes contributing to plant-beneficial functions in Plant Growth-Promoting Rhizobacteria and related Proteobacteria. *Scientific reports* **4,** 6261, doi: 10.1038/srep06261 (2014).
13. Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462,** 1056–1060, doi: 10.1038/nature08656 (2009).
14. Reeve, W. *et al.* A Genomic Encyclopedia of the Root Nodule Bacteria: assessing genetic diversity through a systematic biogeographic survey. *Standards in genomic sciences* **10,** 14, doi: 10.1186/1944-3277-10-14 (2015).
15. Kyrpides, N. C. *et al.* Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLoS biology* **12,** e1001920, doi: 10.1371/journal.pbio.1001920 (2014).
16. Vavilov, N. I. Centers of origin of cultivated plants. *Trends Pract Bot Gener Sel* **16,** 3–24 (1926).
17. Glick, B. R. Modulation of plant ethylene levels by the bacterial enzyme ACC deaminase. *FEMS microbiology letters* **251,** 1–7, doi: 10.1016/j.femsle.2005.07.030 (2005).
18. Ma, W., Charles, T. C. & Glick, B. R. Expression of an exogenous 1-aminocyclopropane-1-carboxylate deaminase gene in Sinorhizobium meliloti increases its ability to nodulate alfalfa. *Appl Environ Microbiol* **70,** 5891–5897, doi: 10.1128/AEM.70.10.5891-5897.2004 (2004).
19. Reddy, T. B. K. *et al.* The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta) genome project classification. *Nucleic acids research* **43,** D1099–D1106, doi: 10.1093/Nar/Gku950 (2015).
20. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic acids research* **42,** D222–D230 (2013).
21. Roux, B. *et al.* An integrated analysis of plant and bacterial gene expression in symbiotic root nodules using laser-capture microdissection coupled to RNA sequencing. *Plant J* **77,** 817–837, doi: 10.1111/tpj.12442 (2014).
22. Tsukada, S. *et al.* Comparative genome-wide transcriptional profiling of Azorhizobium caulinodans ORS571 grown under free-living and symbiotic conditions. *Appl Environ Microbiol* **75,** 5037–5046, doi: AEM.00398-09 (2009).
23. Li, Y. *et al.* High-resolution transcriptomic analyses of Sinorhizobium sp. NGR234 bacteroids in determinate nodules of Vigna unguiculata and indeterminate nodules of Leucaena leucocephala. *PLoS One* **8,** e70531, doi: 10.1371/journal.pone.0070531 (2013).
24. Chang, W. S. *et al.* An oligonucleotide microarray resource for transcriptional profiling of Bradyrhizobium japonicum. *Mol Plant Microbe Interact* **20,** 1298–1307, doi: 10.1094/MPMI-20-10-1298 (2007).
25. Israel, D. W. Investigation of the role of phosphorus in symbiotic dinitrogen fixation. *Plant physiology* **84,** 835–840 (1987).
26. Cha, J. S. & Cooksey, D. A. Copper resistance in Pseudomonas syringae mediated by periplasmic and outer membrane proteins. *Proceedings of the National Academy of Sciences of the United States of America* **88,** 8915–8919 (1991).
27. Puig, S., Rees, E. M. & Thiele, D. J. The ABCDs of periplasmic copper trafficking. *Structure* **10,** 1292–1295 (2002).
28. Cooksey, D. A. Copper uptake and resistance in bacteria. *Molecular microbiology* **7,** 1–5 (1993).
29. Preisig, O., Zufferey, R., Thony-Meyer, L., Appleby, C. A. & Hennecke, H. A high-affinity cbb3-type cytochrome oxidase terminates the symbiosis-specific respiratory chain of Bradyrhizobium japonicum. *Journal of bacteriology* **178,** 1532–1538 (1996).
30. Delgado, M. J., Bedmar, E. J. & Downie, J. A. Genes involved in the formation and assembly of rhizobial cytochromes and their role in symbiotic nitrogen fixation. *Advances in microbial physiology* **40,** 191–231 (1998).

31. Nielsen, K. M., Bones, A. M., Smalla, K. & van Elsas, J. D. Horizontal gene transfer from transgenic plants to terrestrial bacteria–a rare event? *FEMS microbiology reviews* **22,** 79–103 (1998).
32. Pontiroli, A. *et al.* Visual evidence of horizontal gene transfer between plants and bacteria in the phytosphere of transplastomic tobacco. *Appl Environ Microbiol* **75,** 3314–3322, doi: 10.1128/AEM.02632-08 (2009).
33. Buttner, D. & Bonas, U. Common infection strategies of plant and animal pathogenic bacteria. *Current opinion in plant biology* **6,** 312–319 (2003).
34. Hardt, W. D. & Galan, J. E. A secreted Salmonella protein with homology to an avirulence determinant of plant pathogenic bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **94,** 9887–9892 (1997).
35. Lourenco, E. V. *et al.* Toxoplasma gondii micronemal protein MIC1 is a lactose-binding lectin. *Glycobiology* **11,** 541–547 (2001).
36. Muller, J. J., Weiss, M. S. & Heinemann, U. PAN-modular structure of microneme protein SML-2 from the parasite Sarcocystis muris at 1.95 A resolution and its complex with 1-thio-beta-D-galactose. *Acta crystallographica. Section D, Biological crystallography* **67,** 936–944, doi: 10.1107/S0907444911037796 (2011).
37. Blin, K. *et al.* antiSMASH 2.0–a versatile platform for genome mining of secondary metabolite producers. *Nucleic acids research* **41,** W204–212, doi: 10.1093/nar/gkt449 (2013).
38. Dittrich, H. & Kutchan, T. M. Molecular cloning, expression, and induction of berberine bridge enzyme, an enzyme essential to the formation of benzophenanthridine alkaloids in the response of plants to pathogenic attack. *Proceedings of the National Academy of Sciences of the United States of America* **88,** 9969–9973 (1991).
39. Raffaele, S., Win, J., Cano, L. M. & Kamoun, S. Analyses of genome architecture and gene expression reveal novel candidate virulence factors in the secretome of Phytophthora infestans. *BMC genomics* **11,** 637, doi: 10.1186/1471-2164-11-637 (2010).
40. Herbert, R. B. The biosynthesis of plant alkaloids and nitrogenous microbial metabolites. *Natural product reports* **20,** 494–508 (2003).
41. Facchini, P. J., Huber-Allanach, K. L. & Tari, L. W. Plant aromatic L-amino acid decarboxylases: evolution, biochemistry, regulation, and metabolic engineering applications. *Phytochemistry* **54,** 121–138 (2000).
42. Decker, K. & Bleeg, H. Induction and purification of stereospecific nicotine oxidizing enzymes from Arthrobacter oxidans. *Biochimica et biophysica acta* **105,** 313–324 (1965).
43. Bloemberg, G. V. & Lugtenberg, B. J. Molecular basis of plant growth promotion and biocontrol by rhizobacteria. *Current opinion in plant biology* **4,** 343–350 (2001).
44. Compant, S., Duffy, B., Nowak, J., Clement, C. & Barka, E. A. Use of plant growth-promoting bacteria for biocontrol of plant diseases: Principles, mechanisms of action, and future prospects. *Appl Environ Microb* **71,** 4951–4959, doi: 10.1128/Aem.71.9.4951-4959.2005 (2005).
45. Hammad, Y. *et al.* A possible role for phenyl acetic acid (PAA) on Alnus glutinosa nodulation by Frankia. *Plant Soil* **254,** 193–205, doi: 10.1023/A:1024971417777 (2003).
46. Somers, E., Ptacek, D., Gysegom, P., Srinivasan, M. & Vanderleyden, J. Azospirillum brasilense produces the auxin-like phenylacetic acid by using the key enzyme for indole-3-acetic acid biosynthesis. *Appl Environ Microbiol* **71,** 1803–1810, doi: 10.1128/AEM.71.4.1803-1810.2005 (2005).
47. Canova, M. J. & Molle, V. Bacterial serine/threonine protein kinases in host-pathogen interactions. *The Journal of biological chemistry* **289,** 9473–9479, doi: 10.1074/jbc.R113.529917 (2014).
48. Rossbach, S., Kulpa, D. A., Rossbach, U. & de Bruijn, F. J. Molecular and genetic characterization of the rhizopine catabolism (mocABRC) genes of Rhizobium meliloti L5-30. *Molecular & general genetics: MGG* **245,** 11–24 (1994).
49. Beuve, A., Boesten, B., Crasnier, M., Danchin, A. & O'Gara, F. Rhizobium meliloti adenylate cyclase is related to eucaryotic adenylate and guanylate cyclases. *Journal of bacteriology* **172,** 2614–2621 (1990).
50. Tian, C. F., Garnerone, A. M., Mathieu-Demaziere, C., Masson-Boivin, C. & Batut, J. Plant-activated bacterial receptor adenylate cyclases modulate epidermal infection in the Sinorhizobium meliloti-Medicago symbiosis. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 6751–6756, doi: 10.1073/pnas.1120260109 (2012).
51. Catanese, C. A., Emerich, D. W. & Zahler, W. L. Adenylate cyclase and cyclic AMP phosphodiesterase in Bradyrhizobium japonicum bacteroids. *Journal of bacteriology* **171,** 4531–4536 (1989).
52. Tellez-Sosa, J., Soberon, N., Vega-Segura, A., Torres-Marquez, M. E. & Cevallos, M. A. The Rhizobium etli cyaC product: characterization of a novel adenylate cyclase class. *Journal of bacteriology* **184,** 3560–3568 (2002).
53. An, S. Q. *et al.* A cyclic GMP-dependent signalling pathway regulates bacterial phytopathogenesis. *The EMBO journal* **32,** 2430–2438, doi: 10.1038/emboj.2013.165 (2013).
54. McDonough, K. A. & Rodriguez, A. The myriad roles of cyclic AMP in microbial pathogens: from signal to sword. *Nature reviews. Microbiology* **10,** 27–38, doi: 10.1038/nrmicro2688 (2012).
55. Charpentier, M. & Oldroyd, G. How close are we to nitrogen-fixing cereals? *Current opinion in plant biology* **13,** 556–564, doi: 10.1016/j.pbi.2010.08.003 (2010).
56. Bennett, S. Solexa Ltd. *Pharmacogenomics* **5,** 433–438, doi: 10.1517/14622416.5.4.433 (2004).
57. Huntemann, M. *et al.* The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4). *Standards in Genomic Sciences* **10,** 86, 10.1186/s40793-015-0077-y (2015).
58. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* **11,** 119, doi: 10.1186/1471-2105-11-119 (2010).
59. Pati, A. *et al.* GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nature methods* **7,** 455–457, doi: 10.1038/nmeth.1457 (2010).
60. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* **25,** 955–964 (1997).
61. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research* **35,** 3100–3108, doi: 10.1093/nar/gkm160 (2007).
62. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: an RNA family database. *Nucleic acids research* **31,** 439–441 (2003).
63. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology* **305,** 567–580, doi: 10.1006/jmbi.2000.4315 (2001).
64. Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *Journal of molecular biology* **340,** 783–795, doi: 10.1016/j.jmb.2004.05.028 (2004).
65. Markowitz, V. M. *et al.* IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* **25,** 2271–2278, doi: 10.1093/bioinformatics/btp393 (2009).

Acknowledgements

We thank Amrita Pati at Roche Molecular Systems for help with gathering Pfams for secondary metabolite detection from AntiSMASH and Emiley Eloe-Fadrosh at Joint Genome Institute for assistance with

illustrations. We also thank the DOE JGI production sequencing, IMG, and Genomes OnLine Database teams for their support. This work was conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, under Contract No. DE-AC02-05CH11231.

Author Contributions

N.C.K., N.N.I, W.G.R. and T.W. conceived the project. W.G.R., J.K.A., N.C.K., N.N.I. and T.W. collected and sequenced the microbial strains. R.S. conceived and executed the analysis plan, N.N.I. and K.T. analyzed the data. R.S., N.N.I. and N.C.K. wrote the manuscript and produced tables and figures. All authors edited and approved the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Seshadri, R. *et al.* Discovery of Novel Plant Interaction Determinants from the Genomes of 163 Root Nodule Bacteria. *Sci. Rep.* **5**, 16825; doi: 10.1038/srep16825 (2015).