Title

A root nodule microbiome sequencing data set from red alder (Alnus rubra Bong.).

Journal

Scientific Data, 11(1)

Authors

Bell, Callum

Sena, Johnny

Fajardo, Diego

et al.

Publication Date

2024-12-18

DOI

10.1038/s41597-024-04131-0

Peer reviewed

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A root nodule microbiome sequencing data set from red alder (*Alnus rubra* Bong.)

Callum J. Bell[1✉], Johnny A. Sena[1], Diego A. Fajardo[1], Evan M. Lavelle[1], Michael A. Costa[2], Barrington Herman[3], Laurence B. Davin[2], Norman G. Lewis[2] & Alison M. Berry[4]

There have been frequent reports of more than one strain of the nitrogen-fixing symbiont, *Frankia*, in the same root nodule of plants in the genus *Alnus*, but quantitative assessments of their relative contributions have not been made to date. Neither has the diversity of other microbes, having potential functional roles in symbiosis, been systematically evaluated. *Alnus rubra* root nodule microbiota were studied using Illumina short read sequencing and kmer-based read classification. Single end 76 bp sequencing was done to a median depth of 96 million reads per sample. Reads were assigned to taxa using KrakenUniq, with taxon abundances being estimated using its companion program Bracken. This was the first high resolution study of *Alnus* root nodules using next generation sequencing (NGS), quantifying multiple Cluster 1 A *Frankia* strains in single nodules, and in some cases, a Cluster 4 strain. Root nodules were found to contain diverse bacteria, including several genera containing species known to have growth-promoting effects. Evidence was found for partitioning of some bacterial strains in older versus younger lobes.

Background & Summary

Among the angiosperms, ten families of plants form mutually beneficial relationships with nitrogen-fixing bacteria, through the formation of root nodules, specialised plant organs that offer a beneficial environment for the microbes, and facilitate the exchange of metabolites between the host and the microsymbiont. The capability to form nitrogen-fixing root nodule symbioses in these families derives from a single common evolutionary origin[1–3]. In the Fabaceae, the bacterial partners are rhizobia, a diverse group within α- and β-Proteobacteria[4,5]. In contrast, the bacterial symbionts in the eight actinorhizal plant families are Actinobacteria in the genus *Frankia*[6].

Evidence for multiple strains of *Frankia* residing in the same root nodule has been found in numerous studies. Electrophoretic patterns of whole protein extracts of *Alnus incana* ssp. *rugosa* root nodules indicated the presence of two *Frankia* strains in the same nodule[7] while restriction patterns of total *Frankia* genomic DNA consistent with the presence of more than one strain were observed in *Elaeagnus angustifolia*[8] and *Myrica pensylvanica*[9] nodules. Sequence analyses of the nifD-nifK intergenic spacer[10] showed that Cluster 1 and Cluster 3 strains were found in nodules from many cultivars of *Myrica rubra*, and that two strains belonging to different Clusters of *Frankia* could be found in the same nodule. Using sequencing of partial 16 s rRNA genes, McEwan *et al.*[11] concluded that *Alnus glutinosa* nodules contained 2-3 *Frankia* strains, with one greatly outnumbering the others. High throughput sequencing of DNA extracts of root nodules of *Datisca glomerata* showed the presence of two closely related Cluster 2 strains along with another less abundant strain[12], while up to three Cluster 2 strains were found in root nodule microbiota originating in three different actinorhizal species[13]. Finally, Welsh *et al.*[14] demonstrated the presence of multiple *Frankia* strains in root nodules of *Alnus oblongifolia* by comparing nifH gene fragments. There is also evidence of non-*Frankia* bacteria inhabiting actinorhizal root nodules. Actinobacteria most closely related to Thermomonosporaceae and Micromonosporaceae were isolated from *Casuarina equisetifolia*[15], and *Micromonospora* was cultured from root nodules of actinorhizal plants in seven genera[16]. Recently, a study of *Casuarina glauca* root nodule metagenomes reported the presence of bacteria belonging to the genera *Micromonospora*, *Bacillus*, *Afipia*, *Phyllobacterium* and *Paenibacillus*, in addition to *Frankia*[17].

[1]National Center for Genome Resources, Santa Fe, NM, USA. [2]Institute of Biological Chemistry, Washington State University, Pullman, WA, USA. [3]Research and Extension Center, Washington State University, Puyallup, WA, USA. [4]Department of Plant Sciences, University of California, Davis, CA, USA. ✉e-mail: cjb@ncgr.org

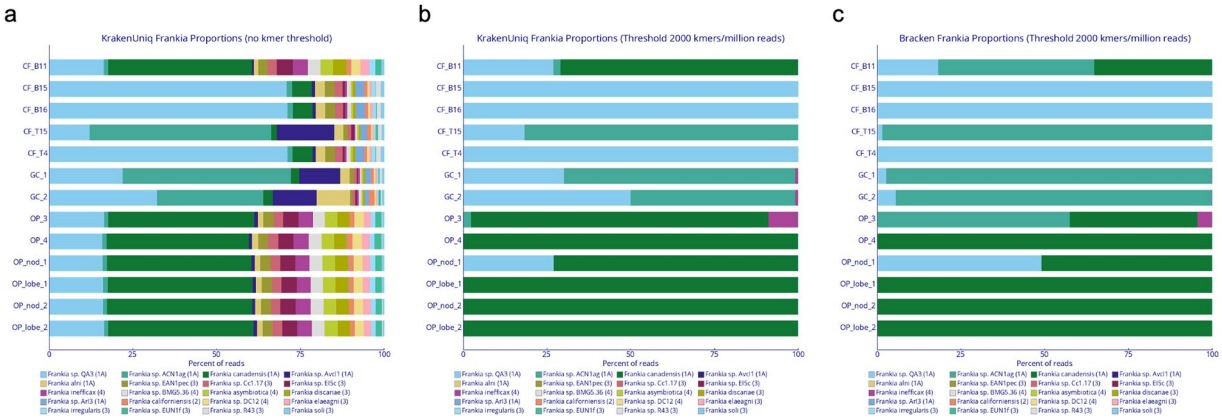

Fig. 1 The 20 most abundant *Frankia* strains across all samples. (**a**) Fractional read counts computed by KrakenUniq, with no kmer threshold. (**b**) Fractional read counts computed by KrakenUniq, applying the kmer threshold of 2000 kmers per million reads. (**c**) Fractional read counts computed by Bracken. Root nodules are described in Table 1. Each horizontal bar represents a nodule metagenome. The reads contributing to each strain within a metagenome are indicated by the different colors. The *Frankia* Cluster is indicated in parentheses after each strain name.

Sample	Inoculum	Host clone	Growth conditions	Specimen	Collection date	Reads
GC_1	Unknown	Clone 639	growth chamber, Pullman, WA	Entire nodule	7-24-17	172,992,227
GC_2	Unknown	Clone 639	growth chamber, Pullman, WA	Entire nodule	7-24-17	169,160,642
CF_B11	B11	Clone 10	Cold frames, Puyallup, WA	Entire nodule	9-26-17	87,528,475
CF_T15	T15	Clone 10	Cold frames, Puyallup, WA	Entire nodule	9-27-17	56,014,379
CF_B16	B16	Clone 10	Cold frames, Puyallup, WA	Entire nodule	9-28-17	81,660,549
CF_B15	B15	Clone 639	Cold frames, Puyallup, WA	Entire nodule	9-29-17	52,761,481
CF_T4	T4	Clone 639	Cold frames, Puyallup, WA	Entire nodule	9-30-17	80,354,146
OP_lobe_1	None/unknown	Field sample	Natural environment, Aberdeen, WA	Dissected lobes	10-24-17	122,596,774
OP_nod_1	None/unknown	Field sample	Natural environment, Aberdeen, WA	Entire nodule	10-24-17	67,128,101
OP_lobe_2	None/unknown	Field sample	Natural environment, Aberdeen, WA	Dissected lobes	10-24-17	104,672,037
OP_nod_2	None/unknown	Field sample	Natural environment, Aberdeen, WA	Entire nodule	10-24-17	64,344,350
OP_3	None/unknown	Field sample	Natural environment, Aberdeen, WA	Entire nodule	10-24-17	135,095,687
OP_4	None/unknown	Field sample	Natural environment, Aberdeen, WA	Entire nodule	10-24-17	167,108,699

Table 1. Root nodule characteristics and their sequence read counts.

Whereas culture and single gene sequencing approaches provide valuable insights into the composition of nodule microbiota, whole metagenome sequencing gives higher sensitivity, offers the opportunity to identify specific genes and pathways of interest, and can provide strain-level resolution. In this study, we applied Illumina whole metagenome sequencing and sensitive kmer-based read classification to DNA extracted from root nodules of *Alnus rubra* Bong. from the state of Washington, USA, grown in either cold frames, a growth chamber, or sampled from a natural environment. These data will be of interest to a broad range of researchers studying plant-microbe interactions, particularly in the quest to understand the roles of nodule-associated bacteria in the establishment and function of nitrogen-fixing root nodules, but also in the field of plant growth promotion by microbes more generally. This study contributes the first such data set (https://identifiers.org/ncbi/insdc.sra:SRP417554) from an ecologically and economically important tree that is a cornerstone species in Pacific Northwestern forest ecosystems.

Throughout our report, it should be noted that the read classifications depend on kmer matches to the bacteria represented in the database. Accordingly, it cannot be said with certainty that, for example, sample CF_B11 contains mostly *Frankia canadensis*. It is more accurate to say that the dominant strain shares more kmers with *Frankia canadensis* than with any other strain represented in the database. When we refer to the abundance of a particular strain, the above is the intended meaning.

Based upon reads assigned by KrakenUniq[18], the 20 most abundant *Frankia* strains are shown in Fig. 1a. The same data, after elimination of taxa having fewer than 2000 kmers per million assigned reads, are shown in Fig. 1b. The corresponding strain abundances, estimated by the Bracken algorithm[19] are shown in Fig. 1c. The corresponding read and kmer counts for all taxa are shown in Tables S1 through S5. *Frankia* read and kmer counts based on reads assigned by KrakenUniq[18] are shown in Table S1 and Table S2. The same data, after applying the 2000 kmers per million assigned reads threshold are shown in Table S3 and Table S4. The taxon abundances estimated by Bracken are shown in Table S5.

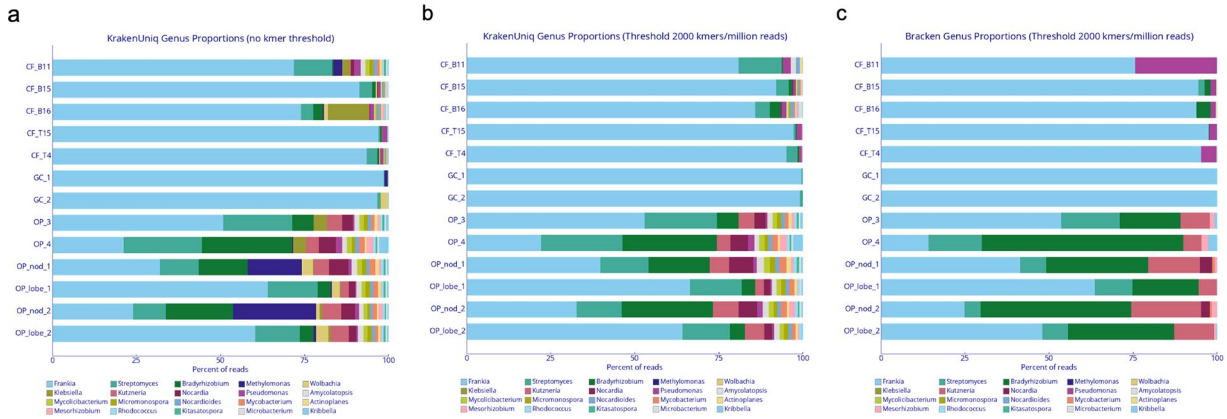

Fig. 2 The 20 most abundant genera across all samples. (**a**) Fractional read counts computed by KrakenUniq, without applying a kmer threshold. (**b**) Fractional read counts computed by KrakenUniq, applying the kmer threshold of 2000 kmers per million reads. (**c**) Fractional read counts computed by Bracken. Root nodules are described in Table 1. Each horizontal bar represents a nodule metagenome. The reads contributing to each genus within a metagenome are indicated by the different colors.

Our results provide strong evidence for the coexistence of multiple *Frankia* strains within individual root nodules. Pivotal to our observations is the practical guidance provided in KrakenUniq[18]. In our study, we applied their threshold recommendations for the elimination of false positive identification of bacterial taxa. Without applying any kmer threshold the results indicate the presence of *Frankia* strains belonging to all Clusters (Tables S1 and S2), with appreciable numbers of kmers being assigned to Cluster 1B strains e.g. *Frankia casuarinae* and *Frankia* sp. BMG5.23. Such observations would be difficult to reconcile with the general absence of Cluster 1B strains in North America. By applying the recommended kmer threshold, which ignores taxa having fewer than 2000 kmers per million reads, the diversity of *Frankia* strains was greatly simplified, consisting in the main of Cluster 1 A strains, with different strains being more abundant in different nodules (Fig. 1b,1c). The data from the lobes-only sample taken from a mature field-grown tree differ from samples taken from entire nodules (the bottom four rows in Fig. 1c), possibly indicating compartmentalization of *Frankia* strains between older and more recently developed parts of the same nodule. Equivalent analyses of the data, quantifying kmers and reads assigned to bacteria at the levels of genus and species are shown in Figs. 2 and 3 and in Tables S6–S15.

Methods

Root nodule collection and DNA extraction. Details of the root nodules and their sequencing statistics are shown in Table 1. The samples prefixed with CF_ came from saplings growing in cold frames at the Washington State University Research and Extension Center, Puyallup, WA. These inocula (B11, B15, B16, T15 and T4) each represent one original nodule that was taken from one tree each. These trees were founders in the clonal selection program, originally selected for traits of interest. Each inoculum was maintained separately in order to study their interactions with a range of tree clones. Accordingly, different totes contained different inocula. The CF_ samples were each derived from a single nodule, each taken from a different tree, in different totes. The cuttings were rooted in new perlite using rooting powder consisting of 4 or 6 g of indole butyric acid per kilogram of talc. Once rooted, the cuttings were kept in the perlite until they were at least 10 cm tall and then moved to a 50/50 perlite vermiculite mix and grown to around 30 cm. Irrigation was occasionally supplemented with half-strength Murashige and Skoog (MS) medium when the plants showed nutrient stress. At approx 30 cm the plants were transferred to acid-washed sand in plastic totes with five trees per tote. The sand was prepared by soaking in 0.5 M HCl, then rinsed 400 times with deionized water. Once in the totes, the cuttings were grown in half-strength Murashige and Skoog medium for 2-3 weeks, inoculated, and subsequently treated with deionized water. Inoculum slurries were prepared from crushed root nodules. After inoculation the nodulated trees were maintained for several years in cold frames that were open to the environment. Fallen leaves were allowed to remain in the totes. The nodules for sequencing were taken from the totes on 09/20/2017 at repotting time at which point they were 9 years old and sapling diameter at the root collar was 1–5 cm. Each nodule was collected from a different plant. The nodules were all in the range 1–1.5 cm. The plants that provided the GC_1 and GC_2 nodules were grown from cuttings prepared similarly to above, but instead of planting in acid washed sand, they were placed (at the 30 cm stage) into potting soil (Sungro Professional Growing Mix) before inoculation. They were transferred to WSU, Pullman, and maintained in a growth chamber with a 16 hour/8 hour light/dark cycle and a temperature of 22 C/20 C. The GC_1 and GC_2 nodules were collected from the young red alder saplings in the growth chamber described above. These young saplings originated at WSU Puyallup, and were inoculated with one of the inocula in rows 3–7 (Table 1). Samples prefixed by OP_ were harvested from a single mature tree near Aberdeen, WA, growing in a stand of red alder thought to be at least 50 years old. OP_lobe_1 and OP_nod_1 originate from the same root nodule. In the case of OP_nod_1, half of an entire nodule was used for DNA extraction. In the case of OP_lobe_1 DNA was extracted from the most recent lobes (the lobe tips) separated from the remainder of the nodule. OP_lobe_2 and OP_nod_2 represent another nodule treated similarly. DNA extractions were performed using the ZymoBiomics™ DNA Miniprep kit (Zymo Research).

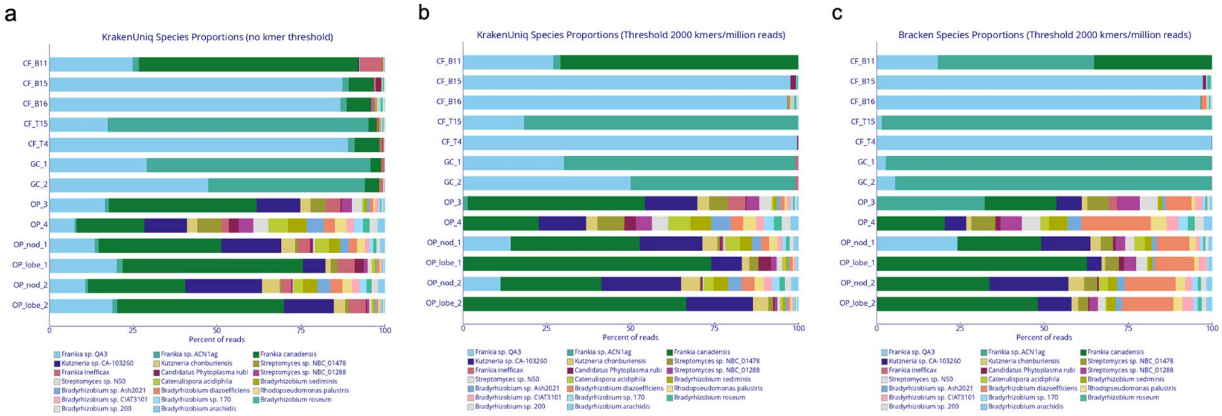

Fig. 3 The 20 most abundant species across all samples. (**a**) Fractional read counts computed by KrakenUniq, without applying a kmer threshold. (**b**) Fractional read counts computed by KrakenUniq, applying the kmer threshold of 2000 kmers per million reads. (**c**) Fractional read counts computed by Bracken. Root nodules are described in Table 1. Each horizontal bar represents a nodule metagenome. The reads contributing to each species within a metagenome are indicated by the different colors.

DNA sequencing. Single-ended Illumina 76 base pair sequencing was done to a median depth of 96 M reads on Illumina HiSeq3000 and 4000 instruments by GENEWIZ (now Azenta), South Plainfield, New Jersey. The data were provided to us demultiplexed and in FASTQ format, having undergone QC and trimming by the sequencing laboratory.

Bioinformatics analysis. Sequence reads aligning to the host plastid and mitochondrial genomes were identified and filtered out as follows: an organelle sequence database was created combining the assembled red alder chloroplast genome (GenBank accession: MG356709.1) and a set of genomic contigs containing putative mitochondrial sequences (contigs 000436F, 000470F, 000507F, 000509F, 000512F, 000550F, 000567F, 000626F, 000671F, 000675F, 000742F, 000783F in GenBank accession GCA_028654335.1)[20]. The metagenome sequence reads were aligned to these organelle sequences using HISAT2[21], using default parameters. The putative organelle reads, consisting of between 0.12% and 0.56% of the total reads per sample, were removed. The remaining sequence reads were assigned to bacterial genomes with KrakenUniq[18] version 0.6 using its standard bacterial database and the NCBI taxonomy, downloaded on 06/21/2024. The standard database was modified to include the *Frankia* genomes described in Table 2. KrakenUniq was chosen as the read classifier as it was shown to have the best performance when compared to a comprehensive benchmarking study[22]. Furthermore, the paper describing KrakenUniq offered objective criteria for avoiding false positive results, which is a recognized problem in metagenome studies[18]. Using synthetic metagenomes these authors demonstrated that ignoring taxa containing fewer than 2000 kmers per million sequence reads discriminates well between true and false positives. We applied this threshold throughout our study.

The *Frankia* genomes used to augment the standard KrakenUniq database were: *Frankia torreyi*[23], *Frankia* sp. ArI3[24], *Frankia* ACN1^{ag}[25], *Frankia* sp. EUN1f[26], *Frankia casuarinae*[27], *Frankia elaeagni*[28], *Frankia* sp. CeD[29], *Frankia* sp. R43[30], *Frankia* sp. EAN1pec[27], *Frankia inefficax*[31], *Frankia saprophytica*[32], *Frankia alni*[27], *Frankia discariae*[33], *Frankia* sp. AvcI1[34], *Frankia* sp. Cc1.17[35], *Frankia* sp. Ea1.12[36],*Frankia* sp. BMG5.23[37], *Frankia* sp. DC12[38], *Frankia* sp. EI5c[39], *Frankia* sp. KB5[40], *Frankia* sp. QA3[41], *Frankia irregularis*[42], *Frankia* sp. Allo2[43], *Frankia* sp. Iso899 (NCBI BioProject ID: 186458, NCBI Tax ID: 1283283), *Frankia* sp. CcI6[44], *Frankia* sp. CpI1-P[23], *Frankia coriariae*[45], *Frankia* sp. CgMI4[46], *Frankia* sp. CcI156[46], *Frankia* sp. CgS1[46], *Frankia* sp. CcI49[47], *Frankia* sp. BMG5.36[48], *Frankia* sp. BMG5.30[49], *Frankia* sp. EUN1h[48], *Frankia asymbiotica*[48], *Frankia canadensis*[50], *Candidatus Frankia californiensis*[12], *Frankia soli*[51], *Candidatus Frankia datiscae*[52].

KrakenUniq default parameters were used. Read counts were converted to taxon abundance estimates using the Bracken algorithm[19]. For each *Frankia* genome (Table 2) we obtained a set of predicted protein sequences by applying PPanGGOLiN[53] with default parameters. Metagenome assemblies were built using the SqueezeMeta pipeline version 1.3.1[54]. To maximize contig size and predicted peptides, SqueezeMeta was executed in coassembly mode to obtain a metagenome assembly of pooled nodule sequence data.

Glutamine synthetase 1 (glnA1) and nifH genes in the set of *Frankia* genomes (Table 2) were identified using BLASTP. The query sequences were the multispecies *Frankia* protein sequences (NCBI RefSeq WP_044887530.1 and WP_011438842.1, respectively), and the BLAST database consisted of all *Frankia* protein sequences predicted from the genomes from NCBI in Table 2 using PPanGGOLiN[53] with default parameters. This collection of *Frankia* glnA1and nifH proteins were then used to discover related sequences in the nodule metagenome data. Each *Frankia* glnA1and nifH protein was used to search a BLASTP database consisting of the predicted proteins in the metagenome assembly contigs. The top 50 BLASTP hits among the metagenome proteins were extracted, and their reciprocal best hits were determined using BLASTN with their cognate coding sequences against the database of 93,000 bacterial genomes used by KrakenUniq, supplemented with the *Frankia* strains in Table 2. Metagenome sequences having best hits to taxa other than Cluster 1 A Frankiae were analyzed further

NCBI taxon ID	NCBI taxonomy name	Assembly NCBI Accession	Other ID	*Frankia* cluster
1562887	*Frankia coriariae*	GCA_001017755.1	BMG5.1	2
1834514	*Frankia* sp. BMG5.30	GCA_001983005.1		2
1839754	*Candidatus Frankia californiensis*	GCA_900067225.1	Dg2	2
2716812	*Candidatus Frankia datiscae*	GCA_000177615.2	Dg1	2
102897	*Frankia*sp. EUN1f	GCA_000177675.1		3
222534	*Frankia elaeagni*	GCA_000374165.1	BMG5.12	3
269536	*Frankia*sp. R43	GCA_001306465.1		3
298653	*Frankia*sp. EAN1pec	GCA_000018005.1		3
365528	*Frankia discariae*	GCA_000373365.1	BCU110501	3
573497	*Frankia*sp. Cc1.17	GCA_001854655.1		3
573499	*Frankia*sp. Ea1.12	GCA_900465275.1	Framoi1121	3
683316	*Frankia* sp. EI5c	GCA_001636565.1		3
795642	*Frankia irregularis*	GCA_001536285.1	DSM 45899	3
1745382	*Frankia* sp. CcI49	GCA_001983215.1		3
2599596	*Frankia soli*	GCA_001854695.1	NRRL B-16219	3
298654	*Frankia inefficax*	GCA_000166135.1	EuI1c	4
298655	*Frankia saprophytica*	GCA_000235425.3	CN3	4
683315	*Frankia* sp. DC12	GCA_000966285.1		4
1834512	*Frankia* sp. BMG5.36	GCA_001854805.1		4
1834515	*Frankia* sp. EUN1h	GCA_001854645.1		4
1834516	*Frankia asymbiotica*	GCA_001983105.1	NRRL B-16386	4
1283283	*Frankia* sp. Iso899			-
1856	*Frankia torreyi*	GCA_000948395.1	Cpi1-S	1 A
1858	*Frankia*sp. ArI3	GCA_019581175.1		1 A
102891	*Frankia* sp. ACN1^{ag}	GCA_001414035.1		1 A
326424	*Frankia alni*	GCA_000058485.1	ACN14a	1 A
573496	*Frankia* sp. AvcI1	GCA_001420875.1		1 A
710111	*Frankia* sp. QA3	GCA_000262465.1		1 A
1502734	*Frankia* sp. CpI1-P	GCA_001421075.1		1 A
1836972	*Frankia canadensis*	GCA_900241035.1	FRACA1	1 A
106370	*Frankia casuarinae*	GCA_000013345.1	CcI3	1B
258230	*Frankia* sp. CeD	GCA_000732115.1		1B
683305	*Frankia* sp. BMG5.23	GCA_000685765.2		1B
683318	*Frankia* sp. KB5	GCA_002099325.1		1B
981405	*Frankia* sp. Allo2	GCA_000733325.1		1B
1352929	*Frankia*sp. CcI6	GCA_000503735.2		1B
1742262	*Frankia* sp. CgMI4	GCA_001756285.1		1B
1745380	*Frankia* sp. CcI156	GCA_001983015.1		1B
1745381	*Frankia* sp. CgS1	GCA_001854725.1		1B

Table 2. *Frankia* strains and their genome accession numbers used to supplement the KrakenUniq database.

using BLASTP against the non-redundant protein database at NCBI. Ribosomal 16 s RNA gene segments were identified in the metagenome assembled contigs using BLASTN. The query sequence was the V3-V4 segment of the *F. casuarinae* 16 s rRNA gene (nucleotides 351 to 772 of NCBI accession NR_153675.1), and the database consisted of the assembled contigs of the combined sample metagenome sequences. All contigs reporting alignments were used in a reciprocal BLASTN search against the non-redundant nucleotide database at NCBI. PPanGGOLiN[53] was also used to make gene family assignments across 38 strains of *Frankia* (Table 2). This process allowed identification of Cluster-specific *Frankia* genes. (During this process *Frankia* sp. *Iso899* was discovered not to belong to the genus *Frankia* based on glnA1 sequence identity and other measures of genome similarity). We then took those genes (911 genes in four Cluster 2 strains, 685 genes in eleven Cluster 3 strains, 900 genes in six Cluster 4 strains; there were no Cluster 1A- or 1B-specific genes) (Table 3) and identified their best BLASTN hits in the predicted genes from the assembled combined nodule metagenome. Those metagenome hits were aligned with BLASTP to a database of 87,516,077 bacterial proteins distributed with the metagenome classifier Kaiju version 1.8.0[55]. Proteins having reciprocal best hits below an E-value threshold of 10e-25 were identified. Functional annotations of the protein translations of genes of interest were made using eggNOG-Mapper version 2.0.1[56].

Data Records. The sequences used in this study were deposited in the NCBI Sequence Read Archive (https://identifiers.org/ncbi/insdc.sra:SRP417554)[57] and the assembled metagenomes were deposited in

Frankia cluster	Number of cluster-specific genes	Number of strains	Total number of genes	Number of reciprocal best hits
2	911	4	3644	13
3	685	11	7535	12
4	900	6	5400	52

Table 3. The numbers of *Frankia* cluster-specific genes identified in the combined metagenome assembly.

GenBank (accessions GCA_043110165.1, GCA_043110465.1, GCA_043109925.1, GCA_043110405.1, GCA_043110005.1, GCA_043110505.1, GCA_043110065.1, GCA_043110605.1, GCA_043110245.1, GCA_043110625.1, GCA_043110325.1, GCA_043110305.1, GCA_043110225.1). These records are all indexed at NCBI under BioProject PRJNA924029[58]. A Figshare repository (https://doi.org/10.6084/m9.figshare.24615723)[59] contains the data necessary to reconstruct the KrakenUniq index we used for read classification, and the supporting tables referenced as S1-S22 in the text.

Technical Validation

Because the KrakenUniq data indicated the presence of multiple *Frankia* strains in all root nodules, this diversity should also be reflected in specific genes. We chose to examine glutamine synthetase 1 (glnA1), which has been used for phylogenetic studies in *Frankia* and in the Actinobacteria more broadly[60,61]. The proteins predicted in the metagenome assembly were searched with BLASTP using a set of *Frankia* glnA1 sequences as described in Methods. The 50 best hits were used in a reciprocal BLASTN search of the KrakenUniq genome database, supplemented with the genomes of the *Frankia* strains in Table 2. Table S16 shows the metagenome contigs matching glnA1 and their reciprocal best BLASTN hits. All are from *Frankia* Cluster 1 A. Analysis of the metagenome representatives of the nifH gene and the V3-V4 segment of the 16 s rRNA gene yielded results supportive of the glnA1 observations. When analyzed by reciprocal BLAST, all such alignments were to sequences annotated as Cluster 1 A *Frankia* strains.

Analysis of genes belonging exclusively to each *Frankia* Cluster, and found in all members of that Cluster examined, yielded: 911 predicted gene families belonging exclusively to all four Cluster 2 strains; 685 predicted gene families belonging exclusively to all eleven Cluster 3 strains; 900 predicted gene families belonging exclusively to all six Cluster 4 strains. The eggNOG-Mapper annotations of these Cluster-specific gene sets are shown in Table S17, Table S18, and Table S19 for Clusters 2, 3, and 4, respectively. The reciprocal best hits numbered: Cluster 2, 13; Cluster 3, 12; Cluster 4, 52 (Table S16). The eggNOG-Mapper annotations of these reciprocal best hits are shown in Table S20, Table S21, and Table S22 for *Frankia* Clusters 2, 3, and 4, respectively.

Code availability

Perl scripts used to tabulate the data found in the Figshare repository tables and to draw the figures may be found here: https://github.com/phlatphish/metagenome_scripts.

Received: 24 January 2024; Accepted: 14 November 2024;
Published online: 18 December 2024

References

1. Soltis, D. E. *et al*. Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. *Proc. Natl. Acad. Sci. USA* **92**, 2647–2651 (1995).
2. Griesmann, M. *et al*. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science* **361**, eaat1743 (2018).
3. van Velzen, R., Doyle, J. J. & Geurts, R. A resurrected scenario: single gain and massive loss of nitrogen-fixing nodulation. *Trends Plant Sci.* **24**, 49–57 (2019).
4. Peter, J., Young, W. & Haukka, K. E. Diversity and phylogeny of rhizobia. *New Phytol.* **133**, 87–94 (1996).
5. Rahimlou, S., Bahram, M. & Tedersoo, L. Phylogenomics reveals the evolution of root nodulating alpha- and beta-Proteobacteria (rhizobia). *Microbiol. Res.* **250**, 126788 (2021).
6. Pawlowski, K. & Demchenko, K. N. The diversity of actinorhizal symbiosis. *Protoplasma* **249**, 967–979 (2012).
7. Benson, D. R. & Hanna, D. *Frankia* diversity in an alder stand as estimated by sodium dodecyl sulfate – polyacrylamide gel electrophoresis of whole-cell proteins. *Can. J. Bot.* **61**, 2919–2923 (1983).
8. Dobritsa, S. V. & Stupar, O. S. Genetic heterogeneity among *Frankia* isolates from root nodules of individual actinorhizal plants. *FEMS Microbiol. Lett.* **58**, 287–292 (1989).
9. Bloom, R. A., Mullin, B. C. & Tate, R. L. DNA restriction patterns and DNA-DNA solution hybridization studies of *Frankia* isolates from *Myrica pensylvanica* (Bayberry). *Appl. Environ. Microbiol.* **55**, 2155–2160 (1989).
10. He, X. H., Chen, L. G., Hu, X. Q. & Asghar, S. Natural diversity of nodular microsymbionts of *Myrica rubra*. *Plant Soil* **262**, 229–239 (2004).
11. McEwan, N. R. *et al*. Lobes on *Alnus glutinosa* nodules contain a single major ribotype of *Frankia*. *J. Endocytobiosis Cell Res.* **26**, 83–86 (2015).
12. Nguyen, T. V. *et al*. An assemblage of *Frankia* Cluster II strains from California contains the canonical *nod* genes and also the sulfotransferase gene *nodH*. *BMC Genom.* **17**, 796 (2016).
13. Nguyen, T. V. *et al*. Frankia-enriched metagenomes from the earliest diverging symbiotic *Frankia* cluster: they come in teams. *Genome Biol. Evol.* **11**, 2273–2291 (2019).
14. Welsh, A. K., Dawson, J. O., Gottfried, G. J. & Hahn, D. Diversity of *Frankia* populations in root nodules of geographically isolated Arizona alder trees in central Arizona (United States). *Appl. Environ. Microbiol.* **75**, 6913–6918 (2009).
15. Valdés, M. *et al*. Non-*Frankia* actinomycetes isolated from surface-sterilized roots of *Casuarina equisetifolia* fix nitrogen. *Appl. Environ. Microbiol.* **71**, 460–466 (2005).
16. Carro, L., Pujic, P., Trujillo, M. E. & Normand, P. *Micromonospora* is a normal occupant of actinorhizal nodules. *J. Biosci.* **38**, 685–693 (2013).

17. Ghodhbane-Gtari, F. *et al*. Alone yet not alone: *Frankia* lives under the same roof with other bacteria in actinorhizal nodules. *Front. Microbiol.* **12**, 749760 (2021).
18. Breitwieser, F. P., Baker, D. N. & Salzberg, S. L. KrakenUniq: confident and fast metagenomics classification using unique *k*-mer counts. *Genome Biol*. **19**, 198 (2018).
19. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *Peer J. Comput. Sci* **3**, e104 (2017).
20. Hixson, K. K. *et al*. Annotated genome sequence of a fast-growing diploid clone of red alder (*Alnus rubra* Bong). *G3* **13**, jkad060 (2023).
21. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
22. McIntyre, A. B. R. *et al*. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol*. **18**, 182 (2017).
23. Oshone, R. *et al*. Permanent draft genome sequences for two variants of *Frankia* sp. strain CpI1, the first Frankia strain isolated from root nodules of *Comptonia peregrina*. *Genome Announc*. **4**, e01588–15 (2016).
24. Bell, C. J., Sena, J. A., Gifford, I. S. & Berry, A. M. Contiguous genome sequence of *Frankia* sp. strain ArI3, isolated from root nodules of *Alnus rubra* Bong. *Microbiol. Resour. Announc*. **10**, e00800-21 (2021).
25. Swanson, E. *et al*. Permanent draft genome sequence of *Frankia* sp. strain ACN1^{ag}, a nitrogen-fixing actinobacterium isolated from the root nodules of *Alnus glutinosa*. *Genome Announc.* **3**, e01483–15 (2015).
26. Normand, P. *et al*. Plasmids in *Frankia* sp. *J. Bacteriol*. **155**, 32–35 (1983).
27. Normand, P. *et al*. Genome characteristics of facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography. *Genome Res*. **17**, 7–15 (2007).
28. Nouioui, I. *et al*. Draft genome sequence of *Frankia* sp. strain BMG5.12, a nitrogen-fixing actinobacterium isolated from *Tunisian soils*. *Genome Announc*. **1**, 00468-13 (2013).
29. Ngom, M. *et al*. Permanent draft genome sequence for *Frankia* sp. strain CeD, a nitrogen-fixing actinobacterium isolated from the root nodules of *Casuarina equistifolia* grown in Senegal. *Genome Announc*. **4**, e00265-16 (2016).
30. Pujic, P. *et al*. Genome sequence of the atypical symbiotic *Frankia* R43 strain, a nitrogen-fixing and hydrogen-producing actinobacterium. *Genome Announc*. **3**, e01387–15 (2015).
31. Nouioui, I. *et al*. *Frankia inefficax* sp. nov., an actinobacterial endophyte inducing ineffective, non nitrogen-fixing, root nodules on its actinorhizal host plants. *Antonie Leeuwenhoek* **110**, 313–320 (2017).
32. Ghodhbane-Gtari, F. *et al*. Draft genome sequence of *Frankia* sp. strain CN3, an atypical, noninfective (Nod–) ineffective (Fix–) isolate from *Coriaria nepalensis*. *Genome Announc*. **1**, e00085-13 (2013).
33. Wall, L. G. *et al*. Draft genome sequence of *Frankia* sp. strain BCU110501, a nitrogen-fixing actinobacterium isolated from nodules of *Discaria trinevis*. *Genome Announc*. **1**, e00503-13 (2013).
34. Swanson, E. *et al*. Permanent draft genome sequence of *Frankia* sp. strain AvcI1, a nitrogen-fixing actinobacterium isolated from the root nodules of *Alnus viridis* subsp. *crispa* grown in Canada. *Genome Announc*. **3**, e01511-15 (2015).
35. Swanson, E. *et al*. Permanent draft genome sequence for *Frankia* sp. strain Cc1.17, a nitrogen-fixing actinobacterium isolated from root nodules of Colletia cruciata. *Genome Announc*. **5**, e00530–17 (2017).
36. Navarro, E., Nalin, R., Gauthier, D. & Normand, P. The nodular microsymbionts of *Gymnostoma* spp. are *Elaeagnus*-infective *Frankia* strains. *Appl. Environ. Microbiol*. **63**, 1610–1616 (1997).
37. Ghodhbane-Gtari, F. *et al*. Draft genome sequence of *Frankia* sp. strain BMG5.23, a salt-tolerant nitrogen-fixing actinobacterium isolated from the root nodules of *Casuarina glauca* grown in Tunisia. *Genome Announc*. **2**, e00520-14 (2014).
38. Tisa, L. S. *et al*. Draft genome sequence of *Frankia* sp. strain DC12, an atypical, noninfective, ineffective isolate from *Datisca cannabina*. *Genome Announc*. **3**, e00889-15 (2015).
39. D'Angelo, T. *et al*. Permanent draft genome sequence for *Frankia* sp. strain EI5c, a single-spore isolate of a nitrogen-fixing actinobacterium, isolated from the root nodules of *Elaeagnus angustifolia*. *Genome Announc*. **4**, e00660-16 (2016).
40. Pesce, C. *et al*. Draft genome sequence of the symbiotic *Frankia* sp. strain KB5 isolated from root nodules of *Casuarina equisetifolia*. *J. Genomics* **5**, 64–67 (2017).
41. Sen, A. *et al*. Draft genome sequence of *Frankia* sp. strain QA3, a nitrogen-fixing actinobacterium isolated from the root nodule of *Alnus nitida*. *Genome Announc* **1**, e00103-13 (2013).
42. Nouioui, I. *et al*. *Frankia irregularis* sp. nov., an actinobacterium unable to nodulate its original host, *Casuarina equisetifolia*, but effectively nodulates members of the actinorhizal Rhamnales. *Int. J. Syst. Evol. Microbiol*. **68**, 2883–2890 (2018).
43. Oshone, R. *et al*. Permanent draft genome sequence of *Frankia* sp. strain Allo2, a salt-tolerant nitrogen-fixing actinobacterium isolated from the root nodules of *Allocasuarina*. *Genome Announc*. **4**, e00388-16 (2016).
44. Mansour, S. R. *et al*. Draft genome sequence of *Frankia* sp. strain CcI6, a salt-tolerant nitrogen-fixing actinobacterium isolated from the root nodule of *Casuarina cunninghamiana*. *Genome Announc*. **2**, e01205-13 (2014).
45. Gtari, M. *et al*. Cultivating the uncultured: growing the recalcitrant cluster-2 *Frankia* strains. *Sci. Rep.* **5**, 13112 (2015).
46. Mansour, S. *et al*. Draft genome sequences for the *Frankia* sp. strains CgS1, CcI156 and CgMI4, nitrogen-fixing bacteria isolated from Casuarina sp. in Egypt. *J. Genomics* **8**, 84–88 (2020).
47. Mansour, S. *et al*. Permanent draft genome sequence for *Frankia* sp. strain CcI49, a nitrogen-fixing bacterium isolated from *Casuarina cunninghamiana* that infects Elaeagnaceae. *J. Genomics* **5**, 119–123 (2017).
48. Gueddou, A. *et al*. Permanent draft genome sequences of three *Frankia* sp. strains that are atypical, noninfective, ineffective isolates. *Genome Announc*. **5**, e00174–17 (2017).
49. Gueddou, A. *et al*. Draft genome sequence of the symbiotic *Frankia* sp. strain BMG5.30 isolated from root nodules of Coriaria myrtifolia in Tunisia. *Antonie Leeuwenhoek* **112**, 67–74 (2019).
50. Normand, P. *et al*. *Frankia canadensis* sp. nov., isolated from root nodules of *Alnus incana* subspecies *rugosa*. *Int. J. Syst. Evol. Microbiol*. **68**, 3001–3011 (2018).
51. Ktari, A. *et al*. Permanent draft genome sequence of *Frankia* sp. NRRL B-16219 reveals the presence of canonical *nod* genes, which are highly homologous to those detected in Candidatus *Frankia* Dg1 genome. *Stand. Genomic Sci.* **12**, 51 (2017).
52. Persson, T. *et al*. Genome sequence of '*Candidatus Frankia datiscae*' Dg1, the uncultured microsymbiont from nitrogen-fixing root nodules of the dicot *Datisca glomerata*. *J. Bacteriol.* **193**, 7017–7018 (2011).
53. Gautreau, G. *et al*. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLOS Comput. Biol.* **16**, e1007732 (2020).
54. Tamames, J. & Puente-Sánchez, F. SqueezeMeta, A highly portable, fully automatic metagenomic analysis pipeline. *Front. Microbiol.* **9**, 3349 (2019).
55. Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **7**, 11257 (2016).
56. Huerta-Cepas, J. *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
57. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP417554 (2023).
58. *NCBI GenBank* https://identifiers.org/ncbi/bioproject:PRJNA924029 (2024).
59. Bell, C. *et al*. *Alnus rubra root* nodule metagenome data analysis. *figshare* https://doi.org/10.6084/m9.figshare.24615723 (2024).

60. Clawson, M. L., Bourret, A. & Benson, D. R. Assessing the phylogeny of *Frankia*-actinorhizal plant nitrogen-fixing root nodule symbioses with *Frankia* 16S rRNA and glutamine synthetase gene sequences. *Mol. Phylogenetics Evol.* **31**, 131–138 (2004).
61. Hayward, D., van Helden, P. D. & Wiid, I. J. F. Glutamine synthetase sequence evolution in the mycobacteria and their use as molecular markers for Actinobacteria speciation. *BMC Evol. Biol.* **9**, 48 (2009).

Acknowledgements

This work was supported by the National Science Foundation Plant Genome Research Program (1547842) and, in part, by the USDA National Institute of Food and Agriculture grants (2011-68005-30416) and Hatch umbrella project #1015621, as well as the Arthur M. and Katie Eisig-Tode Foundation. The authors gratefully acknowledge the contribution of red alder plants and root nodules by Weyerhaeuser, critical reading of the manuscript by Dr. Andrew Binns, and infrastructure support by the Department of Plant Sciences, University of California, Davis, CA 95616 (AMB). Grateful thanks also to Dr. Kent Bassett for permission to collect nodules from an approximately 50 year old *Alnus rubra* tree.

Author contributions

C.J.B., L.B.D., N.G.L., A.M.B., B.H. conceived and designed the experiments. J.A.S., E.M.L., D.A.F., M.A.C., B.H. performed the experiments. C.J.B., J.A.S., E.M.L., D.A.F. analyzed the data. C.J.B. prepared the figures and tables. C.J.B. drafted the manuscript. All authors revised the work critically.

Competing interests

N.G.L. is also President of Ealasid, Inc. which propagated red alder Clone 639 through a licensing agreement with Washington State University. The remaining authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.J.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.