LBNL-59824
Preprint

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Title Genome Characteristics of facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography

Author(s), Philippe Normand, Pascal Lapierre, et al

Division Genomics

January 2007

Genome Research Vol. 17

Running Title: *Frankia* genome evolution

Key Words: Actinobacteria, genome evolution, nodulation, nitrogen fixation

Genome characteristics of facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography

Philippe Normand[1], Pascal Lapierre[2], Louis S. Tisa[3], J. Peter Gogarten[2], Nicole Alloisio[1], Emilie Bagnarol[1], Carla A. Bassi[2], Alison M. Berry[4], Derek M. Bickhart[2], Nathalie Choisne[5,6], Arnaud Couloux[6], Benoit Cournoyer[1], Stephane Cruveiller[7], Vincent Daubin[8], Nadia Demange[6], M. Pilar Francino[9], Eugene Goltsman[9], Ying Huang[2], Olga R. Kopp[10], Laurent Labarre[7], Alla Lapidus[9], Celine Lavire[1], Joelle Marechal[1], Michele Martinez[9], Juliana E. Mastronunzio[2], Beth C. Mullin[10], James Niemann[3], Pierre Pujic[1], Tania Rawnsley[3], Zoe Rouy[7], Chantal Schenowitz[6], Anita Sellstedt[11], Fernando Tavares[12], Jeffrey P. Tomkins[13], David Vallenet[7], Claudio Valverde[14], Luis G. Wall[14], Ying Wang[10], Claudine Medigue[7], & David R. Benson[2]

[1]UMR CNRS 5557 Ecologie Microbienne, IFR41 Bio Environnement et Santé Université Lyon I, Villeurbanne 69622 cedex, France;

[2]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06279

[3]Department of Microbiology, University of New Hampshire, Durham, NH, 03824.

[4]Department of Plant Sciences, University of California, Davis, CA 95616

[5]INRA-URGV, 2 rue Gaston Crémieux BP5708 91057 Evry cedex, France

[6]Genoscope, Centre National de Séquençage, 2 rue Gaston Crémieux BP5706 91057 Evry cedex, France.

[7]Genoscope,CNRS-UMR 8030, Atelier de Génomique Comparative, 2 rue Gaston Crémieux BP5706 91006 Evry cedex, France

[8]Bioinformatics and Evolutionary Genomics Laboratory, UMR CNRS 5558, Université Lyon I, Villeurbanne 69622 cedex, France

[9]DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598

[10]Departments of Botany and Biochemistry, Cellular & Molecular Biology and The Genome Science & Technology Program, The University of Tennessee, Knoxville, TN 37996

[11]Dept of Plant Physiology, Umeå university, Sweden, UPSC, Dept of Plant Physiology, Umeå University, S-90187 Umeå, Sweden

[12]Instituto de Biologia Molecular e Celular, Microbiologia Celular e Aplicada, Rua do Campo Alegre, 823, 4150-180 Porto, Portugal.

[13]Clemson University Genomics Institute, Room 304 Biosystems Research Complex, Clemson, SC 29634

[14]Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Saenz Peña 180 Bernal B1876BXD Argentina

Corresponding Author: David R. Benson, Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06279

Telephone: 860-486-4258; Fax: 860-486-4331

e-mail: david.benson@uconn.edu

Soil bacteria that also form mutualistic symbioses in plants encounter two major levels of selection. One occurs during adaptation to and survival in soil, and the other occurs in concert with host plant speciation and adaptation. Actinobacteria from the genus *Frankia* are facultative symbionts that form N_2-fixing root nodules on diverse and globally distributed angiosperms in the "actinorhizal" symbioses. Three closely related clades of *Frankia* sp. strains are recognized; members of each clade infect a subset of plants from among eight angiosperm families. We sequenced the genomes from three strains; their sizes varied from 5.43 Mbp for a narrow host range strain (*Frankia* sp. strain HFPCcI3) to 7.50 Mbp for a medium host range strain (*Frankia alni* strain ACN14a) to 9.04 Mbp for a broad host range strain (*Frankia* sp. strain EAN1pec.) This size divergence is the largest yet reported for such closely related soil bacteria (97.8-98.9% identity of 16S rRNA genes). The extent of gene deletion, duplication and acquisition is in concert with the biogeographic history of the symbioses and host plant speciation. Host plant isolation favored genome contraction, whereas host plant diversification favored genome expansion. The results support the idea that major genome expansions as well as reductions can occur in facultative symbiotic soil bacteria as they respond to new environments in the context of their symbioses.

Two very different groups of bacteria can form nitrogen-fixing root nodules on angiosperms - Gram negative proteobacteria from several families, and high Mol% G+C Gram positive actinobacteria in the family *Frankiaceae*. Nodulating proteobacteria have symbiotic genes (*nod* genes) subject to horizontal transfer among α- and some β-Proteobacteria (Chen et al. 1991; Moulin et al. 2001; Young and Haukka 1996). By contrast, all *Frankia* sp. strains are closely related with no evidence of dissemination of nodulating ability to related actinobacteria (Fig. 1) (Clawson et al. 2004; Normand et al. 1996).

In plants, the capacity to form N_2-fixing root nodules occupied by bacteria is retained in a single lineage of angiosperms known as the "N_2-fixing clade" (Soltis et al. 1995). Ten families within the Eurosid I clade have members that are nodulated (Clawson et al. 2004; Soltis et al. 1995; Swensen 1996). Only two of the families have members that associate with nodulating proteobacteria while eight associate with *Frankia* sp. strains to form the actinorhizal symbiosis (Table 1).

Frankia strains fall into three closely related clusters. Members of each cluster have distinct host ranges (Table 1; Fig. 1). Cluster 1 strains nodulate plants in the Fagales in the Betulaceae and Myricaceae and are often refered to as "Alnus strains" (Normand et al. 1996). A subclade within Cluster 1 is comprised of the narrow host range “Casuarina strains” that under natural conditions nodulate only *Casuarina* and *Allocasuarina* species in the Casuarinaceae (Benson et al. 2004). Conversely, Cluster 3 “Elaeagnus strains” are considered to have a broad host range since they nodulate plants from five families in the Fagales and Rosales (Benson et al. 2004). Finally, the “Rosaceous strains” form Cluster 2, which is sister to the others; representatives of this

cluster have not been isolated and grown in culture. Cluster 2 strains nodulate plants from four families in the Rosales and Cucurbitales (Benson et al. 2004; Vanden Heuvel et al. 2004).

To gain insight into the evolutionary trajectory followed by these closely related, yet host-range and geographically divergent, *Frankia* sp. strains, we sequenced and compared the genomes of three isolates, including a narrow host range Casuarina strain, a medium host range Alnus strain, and a broad host range Elaeagnus strain. The results suggest that gene deletion and duplication have occured to a differential extent in the genomes during adaptation to host plants and their environments. The concept of genome contraction echoes the changes known to occur in obligate bacterial pathogens and symbionts (Mira et al. 2001; Moran 2003; Ochman and Moran 2001), but the observation that both contraction and expansion can occur in closely related lineages of facultatively symbiotic soil bacteria in relation to host distribution has not previously been reported.

Results and Discussion

Actinorhizal plant families emerged in the late Cretaceous (about 100 Myr bp) and subsequently adapted to a wide variety of environments (Magallon et al. 1999). Currently, they are globally distributed in climate zones ranging from alpine and subarctic to tropical (Fig. 2) where they add nitrogen and organic material to nutrient poor soils (Silvester 1976). The native geographical distributions of hosts range from limited in the case of *Casuarina* sp. to broad in the case of *Morella* sp. (Fig. 2). The distribution of bacterial symbionts is obviously more difficult to assess but numerous

studies have shown some correlation with plant distribution (reviewed in (Benson et al. 2004).

Frankia sp. strain HFPCcI3 (CcI3) represents narrow host range Casuarina strains commonly detected in nodules collected from casuarinas in their native Australia (Fig. 2a) and in areas of the world where casuarina trees have been planted as windbreaks or for erosion control (Simonet et al. 1999). Similar strains have not been found in soils in the absence of a suitable host, indicating that the bacteria depend on the plant for their soil propagation (Simonet et al. 1999).

Frankia alni sp. strain ACN14a (ACN) represents Alnus strains that are globally distributed in soils regardless of the presence of a suitable host plant (Benson et al. 2004). This ubiquity parallels the distribution of host plants from the Betulaceae and Myricaceae that have a combined native range spanning all continents except Australia (Table 1; Fig. 2b).

Frankia sp. strain EAN1pec (EAN) represents broad host range Elaeagnus strains that are also globally distributed in soils with or without host plants (Benson et al. 2004). Cognate hosts are the most diverse and have the widest distribution with representatives on all continents including Australia (Table 1; Fig. 2c).

The strains used in this study have 16S rRNA gene sequences that are 97.8% identical between ACN or CcI3 versus EAN, and 98.9% identical between ACN and CcI3 (Fig. 1). This similarity level is frequently observed among bacteria from the same species (Gevers et al. 2005; Wayne et al. 1987), and is typical of the similarity levels found within the genus *Frankia* (Fig. 1; Clawson et al. 2004).

Genome characteristics. The genomes from ACN and CcI3 have been finished and that from EAN has been rendered in a single scaffold with some gaps corresponding to regions that have proven difficult to resolve due to sequence repeats and high GC content (Table 2). Nevertheless, unlike *Streptomyces* (Bentley et al. 2002), all three genomes are circular as demonstrated directly from their sequences (Fig. 3). None of the strains have yielded independently replicating plasmids. Unlike what is observed for obligate symbionts, the coding capacity of the genomes remains quite high at 89% for ACN, 84% for CcI3 and 86% for EAN.

The most striking difference between the three genomes is their sizes, ranging from 5.43 Mb for CcI3 (4499 protein coding sequences, coding sequences = CDS) to 7.50 Mb for ACN (6786 CDS) to 9.04 Mb for EAN (7976 CDS) (Table 2). On a total DNA basis, this range is the largest reported for any group of free-living prokaryotes related at the 98-99% 16S rRNA sequence level described to date. The smallest genome belongs to the narrow host range and geographically limited representative CcI3, and the largest is from strain EAN that belongs to the broadest host range group. This size correlation raises the hypothesis that genome size and content is driven by the host range and biogeography of the symbiosis. We addressed this hypothesis through comparative analysis of the genomes' contents and structures to determine how the disparate sizes have come about.

The genome maps shown in Figure 3 indicate that the patterns of synteny are quite similar with synteny decreasing as the terminus of replication is approached, corresponding to a high degree of gene rearrangement, duplication or deletion in this region. Indeed much of the size differences can be accounted for by expansion in this

area of the genomes of EAN and ACN. Genes related to symbiosis shown in Figure 3 include those encoding nitrogenase (*nif*), uptake hydrogenase (*hup*), squalene hopane cyclase (*shc*) involved in bacteriohopane biosynthesis. Only one gene similar to the common nodulation genes in rhizobia has been found in each strain but with a general function prediction and relatively low Blast scores (the product of FRAAL4911, annotated as a chitin deacetylase, resembles NodB from *Rhizobium sp.* N33 with a score of $3e^{-34}$ but also resembles similar proteins from many other organisms, mainly *Bacillus* sp.). Little is known about the genetic basis of the actinorhizal symbiosis but it is clearly very different from that known to exist among the nodulating proteobacteria.

IS elements and prophage. Insertion elements, transposases, integrated phage and plasmids tend to reflect the degree of plasticity of genomes (Frost et al. 2005). In the three *Frankia* strains, integrases plus transposases comprise about 0.6% (46) of the ACN CDS, 4.1% (187) of the CcI3 CDS, and 3.4% (269) of the EAN CDS (Table 3). Such elements tend to cluster in regions where there is loss of synteny between genomes (Fig. 3). CcI3 has a higher average density (34 per Mb) than EAN (30 per Mb) and both have a much higher density than ACN (6 per Mb). Indeed, of the 33 transposase genes identified in ACN all but four are found in the same context in CcI3 and all but six are in EAN suggesting that such genes in ACN have inactive for at least the 55 million years since the genomes diverged (Clawson et al. 2004; Normand et al. 1996). Examining regions surrounding phage integrases we estimate that prophage contribute 0.4, 11.7 and 7.1% of the CDS of ACN, CcI3 and EAN, respectively. Overall, EAN and CcI3 have had far more dynamic genomes than ACN in their recent history and this plasticity,

conferred by IS elements and phage moving into and out of the genomes, may in large part have driven the size differences observed.

Gene Deletions. To examine how the three genomes have evolved to have such different sizes, we used the order of divergence of the three strains (Fig. 1) to estimate how the genome size differences reflect gene deletion, duplication and acquisition. Using the approximation of 50 Myr per 1% divergence in the 16S rRNA genes (Ochman et al. 1999), the clade containing EAN diverged an estimated 115 MyrBP from the clade containing CcI3 and ACN which diverged from each other about 55 MyrBP (Fig. 1). Therefore, orthologous genes present in ACN and EAN but absent in CcI3 may be assumed to have been lost from CcI3 after it diverged from ACN. A similar approach can be used for identifying deleted genes from ACN. However, genes absent from EAN, but present in the other two strains, could either have been lost in EAN or acquired by horizontal gene transfer (HGT) in the progenitor of ACN and CcI3.

Defining the presence and absence of orthologs by using a reciprocal best BLAST hit criterion with an E-value threshold of 10^{-4} we find that 1054 genes present in both ACN and EAN are not found in CcI3 (Table 3). Similarly, 466 genes are missing from ACN and 555 are missing from EAN. The most logical explanation for the smaller number of genes missing in EAN and ACN versus CcI3 is an accelerated rate of gene loss in CcI3 over the past 55 Myr since diverging from a common ancestor with ACN. Since EAN has had longer to lose genes, its apparent rate of loss is lower than that of either ACN or CcI3. The number deleted is underestimated in all cases since genes

deleted from two of the three strains would appear as strain specific or horizontally transferred (below).

A BlastClust analysis was done to categorize deleted genes using 30% sequence identity over 52% of the length. This analysis indicated that some categories of genes were preferentially deleted over others. Genes annotated as hypothetical, conserved hypothetical or of unknown function account for about one-third of the total (28-34%) lost in each genome (Table 4). CcI3 is missing the most genes in all categories shown in Table 4 except for integrases and transposases; the latter have been especially reduced in ACN. In CcI3, genes involved in transport (including ABC transporters, solute binding dependent transport proteins), regulatory proteins (including genes in the categories TetR, IclR, LuxR, MarR, MerR), dehydrogenases, amidotransferase, oxygenases, and many hypothetical proteins, proteins of unknown function and conserved hypothetical proteins, have been lost. In short, genes encoding the capacity to scavenge and metabolize substrates from the environment have been reduced in CcI3.

Several genes lost by CcI3 are concerned with metabolic activities of potential importance to survival or symbiosis. These include genes encoding the DNA repair enzymes AP endonuclease, photolyase, DNA-formamidopyrimidine glycosylase, DNA alkylation repair and RadC, two cellulases that might be involved in survival or infection, gas vesicle proteins whose loss could signal adaptation to dry environments where *Casuarina* sp. grow, general metabolism enzymes (NAD-dependent glutamate dehydrogenase, PEP carboxylase), and a large number of regulatory and solute transport proteins. Among the latter there is only one iron siderophore gene cluster in CcI3 as compared to two in ACN and three in EAN. More directly related to symbiosis, CcI3 has

lost one of the two copies of the *shc* (squalene hopene cyclase) genes involved in synthesizing bacteriohopane lipids that comprise the envelope of *Frankia* vesicles and provide protection for nitrogenase against oxygen. Unlike ACN and EAN, oxygen protection is conferred by secondary plant cell walls when CcI3 is in symbiosis (Berg and McDowell 1988), perhaps making bacteriohopane synthesis less of a priority.

In general, the classes of genes lost by CcI3 (DNA repair, metabolic enzymes, regulatory proteins) resemble those known to be lost by bacterial endosymbionts of animals (Mira et al. 2001; Ochman and Moran 2001), and indicate that CcI3 is evolving towards a greater dependence on its host. However, CcI3 can still grow on minimal medium so such strains have not yet made a commitment to an obligate symbiotic existence.

Gene Duplication, Acquisition and ORFans. Gene duplication is a major means by which soil bacteria adapt to new niches, or to the availability of new substrates (Francino 2005; Konstantinidis and Tiedje 2005). Gene acquisition is known to be similarly involved in bacterial adaptation to new environments, particularly in the emergence of pathogens (Mira et al. 2001; Ochman and Moran 2001) and in the evolution of mutualistic bacteria in the legume symbiosis (Chen et al. 2001; Moulin et al. 2004; Young and Haukka 1996). *Frankia* symbionts have adapted both to living in diverse soils in most parts of the world, and to living in root nodules from phylogenetically diverse angiosperms.

We defined duplicates as having the lowest BLAST E-value with a gene from the same genome when compared with genomes from other *Frankia* strains, *Acidothermus*

and *Kineococcus*, both close relatives to *Frankia* in the *Frankineae*, *Streptomyces* spp. and the NR (non-redundant) database. Using this approach, about 7.5% (512) of the ORFs in ACN, 9.8% (444) in CcI3 and 18.5% (1355) in EAN could be considered duplicates of other genes in the same genomes (Table 3). Core metabolic genes are generally not duplicated, a differential amplification noted in other bacteria (Francino 2005; Konstantinidis and Tiedje 2005). Surprisingly, CcI3, that has sustained strong reducing evolutionary pressures, nevertheless had a slightly higher percentage of duplicates than ACN, an observation that is accounted for by the proliferation of transposase genes in CcI3 (Tables 3 & 5). Gene duplication has thus enlarged the EAN genome to a greater extent than the genomes of ACN or CcI3. Most of the duplicated genes seem to be located near the replication terminus in all strains. Localization of contingency genes to the terminus has been observed in the linear genomes of *Streptomyces* sp. and in other large genomes (Bentley et al. 2002; Ikeda et al. 2003).

To assess the types of genes duplicated, a BlastClust (NCBI) analysis was done to cluster proteins using a standard of 25% identity over at least 40% of the length of the amino acid sequence. A more stringent analysis using 30% identity over 52% of the sequence gave essentially the same results. In the top twenty duplicated gene families in CcI3, 116 out of 165 (70%) genes belong to several classes of transposases and genes associated with prophages (Table 5). In ACN, no transposases are found in the 151 genes in the top 20 families; instead genes annotated as serine-threonine protein kinases, short-chain dehydrogenases/reductases, endonucleases, SAM-dependent methyltransferases, transport proteins, and a variety of dehydrogenases are duplicated. In EAN, 132 out of 406 (32.5%) genes are associated with integrases, transposases or reverse transcriptases

in the top 20 families, with the remainder annotated as short chain dehydrogenase/reductases, cytochrome P450, transport proteins, regulatory proteins, and dioxygenases.

In sum, EAN has the most duplicated genes in all categories including those whose products are associated with metabolic processes as well as mobile genetic elements. ACN has the fewest duplicates, and those are of genes involved in general metabolism. Finally, a large portion of all duplicates in CcI3 (33% overall) are of transposases. In all strains, the majority of duplicates appeared as two copies of a single gene.

Strain-specific genes (SSGs) include genes lost by two of the three *Frankia* strains plus genes that have no hits in databases (ORFans). Such genes could also have been horizontally transferred from other bacteria. Using a permissive threshold (E-value $\leq 10^{-4}$) between the genomes, and allowing self-genome hits to eliminate duplication, we found that about 23% (1563) of the genes in ACN, 12.8% (578) in CcI3 and 17.7% (1289) in EAN have no clear homologs in the other two genomes. Of those, 854 (12.5%) in ACN, 158 (3.5%) in CcI3, and 355 (4.9%) in EAN were ORFans with no hits in NR, or the related *Acidothermus*, *Kineococcus* or *Streptomyces* spp. genomes. The higher number and percentage of ORFans in ACN may reflect a lower evolutionary pressure to eliminate non-essential genes, a characteristic also reflected in its having the fewest deleted genes overall (Table 3).

Conclusions. We have shown that the unusual size divergence displayed by the *Frankia* genomes has arisen by the processes of deletion, duplication and retention/acquisition

operating in all strains but to different extents (Table 3). These processes have driven the genomes in different directions, reducing that of CcI3, expanding that of EAN and keeping ACN relatively stable. The results of these broad comparisons lead us to propose a link between the biogeographic history of the actinorhizal plants and the genome evolution of the bacterial symbionts.

Evidence from ecological (Zimpfer et al. 1997), molecular ecological (Simonet et al. 1999), physiological (Sellstedt 1995) and now genomic studies indicates that Casuarina strains represented by CcI3 have evolved to become specialists with reduced genomes. Unlike Alnus and Elaeagnus strains, they have not been detected by trapping experiments in soils outside the native ranges of their host plants (Simonet et al. 1999; Zimpfer et al. 1997), and they infect a narrow spectrum of hosts (Fig. 2a; Table 1). Genome reduction is well documented in obligate pathogens and obligate symbionts in plants and animals (Batut et al. 2004; Mira et al. 2001; Moran 2003; Ochman and Moran 2001), and in some free-living cyanobacterial *Prochlorococcus* sp. (Dufresne et al. 2005). Genome reduction has not been described in bacterial facultative symbionts that also exist free-living in the soil; indeed this is a most unexpected finding.

We suggest that a likely explanation for genome reduction in CcI3 is its geographic and symbiotic isolation in Australia and the Pacific islands paralleling its host plants' isolation beginning about 100-65 Myr bp. Casuarinaceae species emerged as part of the flora of Gondwana as evidenced by fossils in New Zealand and South America that today are outside the native range (Campbell and Holden 1984). These plants, and their bacterial symbionts, coadapted to a hotter, drier climate as Australia split from Antarctica

and moved north towards the equator. Present day Casuarina strains live in locales where the soil biotic capacity is reduced and actinorhizal host diversity is limited.

In contrast, plants infected by Elaeagnus strains have a global distribution (Fig. 2c) with ancestral origins in both Gondwana (*Gymnostoma* in the Casuarinaceae in Western Oceania, actinorhizal Colletieae in the Rhamnaceae) and Laurasia (Elaeagnaceae, Myricaceae). Such plants occupy a wide range of soil types and climates. Genome expansion by gene duplication and divergence is a mechanism used by soil bacteria to exploit new niches and new substrates (Francino 2005; Konstantinidis and Tiedje 2005), and may be inferred to have occurred in the ancestors of EAN as they and their hosts coadapted to new and diverse soils. Indeed, the types of genes duplicated are largely involved in introducing substrates into central metabolic pathways. .

The genome of ACN appears more stable than CcI3 and EAN, in the sense that it has few transposases and integrases; it also has lost the fewest genes by deletion, has the fewest duplicated genes (when transposases in CcI3 are not counted), and retains the most strain specific genes, including ORFans. Its stability may reflect its host range focused on the ancient lineages in the Betulaceae and Myricaceae leading to high soil abundance and relatively strong genome homogenization. Its host plants have the longest fossil record of the N_2-fixing clade (Magallon et al. 1999), and have inhabited similar and milder environments in northern latitudes since appearing in Laurasia during the late Cretaceous (Crane 1989).

Taken together, the gene contents of the three *Frankia* strains appear to reflect the biogeographic history of the host plants they infect, and as such may provide the first example of differential genome contraction and expansion occurring in closely related

facultatively symbiotic soil bacteria that may be linked to the evolutionary history of their hosts on a global scale.

Materials and Methods

Strains. CcI3 was isolated from *Casuarina cunninghamiana* plants growing in a greenhouse at Harvard Forest in Petersham, MA (Zhang et al. 1984) on soils coming from its original provenance. ACN was isolated initially from *Alnus viridis* subsp. *crispa* plants in Tadoussac, Quebec (Benson et al. 2004; Normand and Lalonde 1982). Strain EAN was isolated from field nodules of *E. angustifolia* growing in Ohio (Lalonde et al. 1981).

Genome Sequencing, Assembly and Finishing – CcI3 and EAN1pec. We sequenced the three genomes of *Frankia* strains ACN14a, CcI3 and EAN1pec using a shotgun approach. The genomes of *Frankia* strains CcI3, and EAN1pec were sequenced at the Joint Genome Institute (JGI) using a combination of 3 kb, 8 kb and 40 kb (fosmid) DNA libraries for each strain. Draft assemblies were based on 82,561 total reads for CcI3 and 125,615 total reads for EAN1pec. The different libraries provided 4.6X (3 kb), 4.1X (8 kb) and 0.5X (fosmids) coverage of CcI3 and 4.0X (3 kb), 3.4X (8 kb) and 0.6X (fosmids) coverage of EAN1pec. End sequencing and fingerprinting of fosmid clones aided in assembly verification, determination of gap sizes and ordering and orientation of scaffolds beyond assembly gaps.

Sequencing gaps were closed mainly by primer walking on plasmid and fosmid subclone templates. In cases where no acceptable template was available, PCR products

were made and sequenced using customized primers. Gaps resulting from hard-to-sequence DNA structures had to be covered using special chemistries and in-house developed protocols. Mis-assemblies were identified and corrected by means of clone pairing; these primarily occured due to long repeats (rRNAs, IS elements). Over-collapsing of repeat copies often resulted in pseudo-gaps in the assembly, which could not be closed by routine primer walking. Each one of those had to be filled in using one of the following two methods. Small pseudo-gaps were closed using the editing features of CONSED (Gordon et al. 1998), by locating and placing appropriate reads individually into their proper repeat copy. Long pseudo-gaps and long misassembled repeats (over 2kb) had to be isolated and separately assembled. Only consistent, partially unique clone-mates would be allowed in those subassemblies. After verifying the subassembly's integrity and primer-walking over the poorly covered regions, the isolated contigs were re-introduced into the main assembly as "fake reads", that is single continuous long sequences reflecting the correctly assembled repeat copy.

All other general aspects of library construction, sequencing and automated annotation were carried out as previously described for bacterial genomes sequenced at the JGI (Chain et al. 2003). In addition, predicted coding sequences are subject to manual analysis using the Integrated Microbial Genomes (IMG) annotation pipeline. Detailed information about genome annotation and other genome properties can be obtained at http://img.jgi.doe.gov (Markowitz et al. 2006).

Genome sequencing, assembly and annotation – ACN14a. For ACN14a, four libraries were made: two plasmid libraries of 3 kb and 10 kb, obtained by mechanical shearing,

were constructed at Genoscope (Evry, France) into pcDNA2.1 (InVitrogen) and into pCNS home vector (pSU18 modified) (Bartolome et al. 1991), respectively. Two BAC libraries of average insert size of 104 kb were constructed at CUGI (Clemson University Genomics Institute) by enzymatic digestion (*Eco*RI and *Hin*dIII) into pCUGIBAC1 (Luo and Wing 2003). Plasmid and BAC DNAs were purified and end-sequenced using dye-terminator chemistry on ABI3730xl DNA Analyzer sequencers. We generated 150,890 sequences from both ends of genomic clones from the four libraries.

The Phred/Phrap/Consed software package (www.phrap.com) was used for sequence assembly and quality assessment (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998). A total of 8956 additional reactions were necessary to close gaps, generally with the transposing method (Entranceposon, Finnzymes, Espoo, Finland) and to raise the quality of the finished sequence. The validity of the final sequence was assessed by comparing the restriction enzyme pattern deduced from the sequence to the experimentally observed restriction pattern obtained by digestion of genomic DNA. After a first round of annotation, regions of lower quality as well as regions with putative frame-shifts were re-sequenced from PCR amplification of the dubious regions.

Using the AMIGene software (Bocs et al. 2003) a total of 5279 CDSs were predicted and submitted to automatic functional annotation (Vallenet et al. 2006). Each predicted gene was assigned a unique identifier prefixed with "FRAAL". Sequence data for comparative analyses were obtained from the NCBI databank (RefSeq section). Putative orthologs and synteny groups (conservation of the chromosomal co-localisation between pairs of orthologous genes from different genomes) were computed between ACN and the 265 other complete genomes using the procedure described in Vallenet *et*

al. (Vallenet et al. 2006). Manual validation of the automatic annotation was performed using the MaGe (Magnifying Genomes) interface which allows graphic visualization of the ACN annotations enhanced by a synchronized representation of synteny groups in other genomes chosen for comparisons. The ACN nucleotide sequence and annotation data have been deposited at EMBL databank under accession number CT573213. In addition, all the data (*i.e*, syntactic and functional annotations, and results of comparative analysis) was stored in a relational database, called FrankiaScope (Vallenet et al. 2006). This database is publicly available via the MaGe interface at http://www.genoscope.cns.fr/agc/mage/frankia/Login/log.php.

Methods used for determining deleted, duplicated, strain specific and ORFan genes. Genes deleted from one strain were identified by using reciprocal BLAST hits from each pair of genomes. That is, each pair of orthologs identified each other as the lowest BLAST score. Genes were scored as deleted if they did not have a reciprocal hit in another *Frankia* genome. The dataset included three *Frankia* strains.

Gene duplications were assessed as having the best BLAST hits within the same genome (duplicates) using an E-value cut-off of 10^{-4} and a dataset consisting of NR (minus *Frankia* sequences) + *Kineococcus radiodurans* + *Streptomyces coelicolor* + *S. avermitilis*+ ACN + CcI3 + EAN. To cluster duplicates, the program BlastClust (NCBI) was used with settings reported in the text.

Strain-specific genes (SSGs) include genes found in one but not another *Frankia* strain at an E-value cutoff of 10^{-4}, plus genes that have no hits in databases. The latter are referrred to as ORFans.

Acknowledgements

This work was supported by the National Science Foundation Microbial Genome sequencing program to DRB, LST and MPF. The work on CcI3 and EAN was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory and Los Alamos National Laboratory. The work on ACN14a was performed at Genoscope, Evry, France and supported by CNRS/ACI Microbiologie and MRT/ACI IMPBio2004. Funds to Anita Sellstedt are from FORMAS and NFR, Sweden.

Correspondence and requests for materials should be addressed to DRB (david.benson@uconn.edu) for CcI3, PN (normand@biomserv.univ-lyon1.fr) for ACN and LST (lst@hypatia.unh.edu) for EAN. Sequence datasets are available as outlined in the Supporting Information.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. DE-AC52-06NA25396.

References

APG, A.P.G. 1998. An ordinal classification for the families of flowering plants. *Ann. Missouri Bot. Gard.* **85:** 531-553.

Bartolome, B., Y. Jubete, E. Martinez, and F. de la Cruz. 1991. Construction and properties of a family of pACYC184-derived cloning vectors compatible with pBR322 and its derivatives. *Gene* **102:** 75-78.

Batut, J., S.G. Andersson, and D. O'Callaghan. 2004. The evolution of chronic infection strategies in the alpha-proteobacteria. *Nat Rev Microbiol* **2:** 933-945.

Benson, D.R., B.D. Vanden Heuvel, and D. Potter. 2004. Actinorhizal symbioses: diversity and biogeography. In *Plant Microbiology* (ed. M. Gillings). BIOS Scientific Publishers Ltd., Oxford.

Bentley, S.D., K.F. Chater, A.M. Cerdeno-Tarraga, G.L. Challis, N.R. Thomson, K.D. James, D.E. Harris, M.A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C.W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C.H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M.A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B.G. Barrell, J. Parkhill, and D.A. Hopwood. 2002. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* **417:** 141-147.

Berg, R.H. and L. McDowell. 1988. Cytochemistry of the wall of infected *Casuarina* actinorhizae. *Can. J. Bot.* **66:** 2038-2047.

Bocs, S., S. Cruveiller, D. Vallenet, G. Nuel, and C. Medigue. 2003. AMIGene: Annotation of MIcrobial Genes. *Nucleic Acids Res* **31:** 3723-3726.

Campbell, J.D. and A.M. Holden. 1984. Miocene casuarinacean fossils from Southland and Central Otago, New Zealand. *New Zealand J. Bot.* **22:** 159-167.

Chain, P., S. Kurtz, E. Ohlebusch, and T. Slezak. 2003. An applications-focused review of comparative genomics tools: capabilities, limitations and future challenges. *Brief Bioinform* **4:** 105-123.

Chen, L.-M., Y. Cui, M. Qin, Y. Wang, X. Bai, and Q. Ma. 1991. Identification of a *nod*D-like gene in *Frankia* by direct complementation of a *Rhizobium nod*D-mutant. *Mol. Gen. Genet.* **233:** 311-314.

Chen, W.M., S. Laevens, T.M. Lee, T. Coenye, P. De Vos, M. Mergeay, and P. Vandamme. 2001. *Ralstonia taiwanensis* sp. nov., isolated from root nodules of *Mimosa* species and sputum of a cystic fibrosis patient. *Int J Syst Evol Microbiol* **51:** 1729-1735.

Clawson, M.L., A. Bourret, and D.R. Benson. 2004. Assessing the phylogeny of *Frankia*-actinorhizal plant nitrogen-fixing root nodule symbioses with *Frankia* 16S rRNA and glutamine synthetase gene sequences. *Mol. Phyl. Evol.* **31:** 131-138.

Crane, P.R. 1989. 6. Early fossil history and evolution of the Betulaceae. In *Evolution, Systmatics, and Fossil History of the Hamamelidae* (eds. P.R. Crane and S. Blackmore), pp. 87-116. Clarendon Press, Oxford.

Dufresne, A., L. Garczarek, and F. Partensky. 2005. Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol* **6:** R14.

Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8:** 186-194.

Ewing, B., L. Hillier, M.C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8:** 175-185.

Francino, M.P. 2005. An adaptive radiation modelfor the origin of new gene functions. *Nature Genetics* **37:** 573-577.

Frost, L.S., R. Leplae, A.O. Summers, and A. Toussaint. 2005. Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* **3:** 722-732.

Gevers, D., F.M. Cohan, J.G. Lawrence, B.G. Spratt, T. Coenye, E.J. Feil, E. Stackebrandt, Y. Van de Peer, P. Vandamme, F.L. Thompson, and J. Swings. 2005. Opinion: Re-evaluating prokaryotic species. *Nat Rev Microbiol* **3:** 733-739.

Gordon, D., C. Abajian, and P. Green. 1998. Consed: a graphical tool for sequence finishing. *Genome Res* **8:** 195-202.

Ikeda, H., J. Ishikawa, A. Hanamoto, M. Shinose, H. Kikuchi, T. Shiba, Y. Sakaki, M. Hattori, and S. Omura. 2003. Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nat Biotechnol* **21:** 526-531.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16:** 111-120.

Konstantinidis, K.T. and J.M. Tiedje. 2005. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* **102:** 2567-2572.

Lalonde, M., H.E. Calvert, and S. Pine. 1981. Isolation and use of *Frankia* strains in actinorhizae formation. In *Current perspectives in nitrogen fixation* (eds. A.H. Gibson and W.E. Newton), pp. 296-299. Australian Academy of Sciences, Canberra.

Luo, M. and R.A. Wing. 2003. An improved method for plant BAC library construction. *Methods Mol Biol* **236:** 3-20.

Magallon, S., P.R. Crane, and P.S. Herendeen. 1999. Phylogenetic pattern, diversity, and diversification of eudicots. *Ann. Missouri Bot. Gard.* **86:** 297-372.

Markowitz, V., F. Korzeniewski, K. Palaniappan, E. Szeto, G. Werner, A. Padki, X. Zhao, I. Dubchak, P. Hugenholtz, I. Anderson, A. Lykidis, K. Mavrommatis, N. Ivanova, and N.C. Kyrpides. 2006. The Integrated Microbial Genomes (IMG) System. *Nucleic Acids Res (special database issue)* **34:** D344-D348.

Mira, A., H. Ochman, and N.A. Moran. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet* **17:** 589-596.

Moran, N.A. 2003. Tracing the evolution of gene loss in obligate bacterial symbionts. *Curr Opin Microbiol* **6:** 512-518.

Moulin, L., G. Bena, C. Boivin-Masson, and T. Stepkowski. 2004. Phylogenetic analyses of symbiotic nodulation genes support vertical and lateral gene co-transfer within the *Bradyrhizobium* genus. *Mol Phylogenet Evol* **30:** 720-732.

Moulin, L., A. Munive, B. Dreyfus, and C. Boivin-Masson. 2001. Nodulation of legumes by members of the beta-subclass of Proteobacteria. *Nature* **411:** 948-950.

Normand, P. and M.P. Fernandez. 2007. Evolution and Diversity of *Frankia*. In *Actinorhizal Symbioses* (ed. K. Pawlowski). Springer Verlag.

Normand, P. and M. Lalonde. 1982. Evaluation of *Frankia* strains isolated from provenances of two *Alnus* species. *Can. J. Microbiol.* **28:** 1133-1142.

Normand, P., S. Orso, B. Cournoyer, P. Jeannin, C. Chapelon, J. Dawson, L. Evtushenko, and A.K. Misra. 1996. Molecular phylogeny of the genus *Frankia* and related genera and emendation of family *Frankiaceae*. *Int. J. Syst. Bacteriol.* **46:** 1-9.

Ochman, H., S. Elwyn, and N.A. Moran. 1999. Calibrating bacterial evolution. *Proc Natl Acad Sci U S A* **96:** 12638-12643.

Ochman, H. and N.A. Moran. 2001. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* **292:** 1096-1099.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406-425.

Sellstedt, A. 1995. Specificity and effectivity in nodulation by Frankia on southern hemisphere actinorhiza. *FEMS Microbiol Lett* **125:** 231-236.

Silvester, W.B. 1976. Ecological and economic significance of the non-legume symbiosis. In *Proceedings 1st International Symposium on Nitrogen Fixation* (eds. W.E. Newton and C.J. Nyman), pp. 489-506. Washington University Press, Pullman.

Silvester, W.B. 1977. Dinitrogen fixation by plant associations excluding legumes. In *A Treatise on Dinitrogen Fixation* (eds. R.W.F. Hardy and A.H. Gibson), pp. 141-190. John Wiley and Sons, New York.

Simonet, P., E. Navarro, C. Rouvier, P. Reddell, J. Zimpfer, Y. Dommergues, R. Bardin, P. Combarro, J. Hamelin, A.-M. Domenach, F. Gourbiere, Y. Prin, J.O. Dawson, and P. Normand. 1999. Co-evolution between *Frankia* populations and host

plants in the family Casuarinaceae and consequent patterns of global dispersal. *Environ. Microbiol.* **1:** 525-533.

Soltis, D.E., P.S. Soltis, D.R. Morgan, S.M. Swensen, B.C. Mullin, J.M. Dowd, and P.G. Martin. 1995. 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.

Swensen, S.M. 1996. The evolution of actinorhizal symbioses: Evidence for multiple origins of the symbiotic association. *Am. J. Bot.* **83:** 1503-1512.

Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25:** 4876-4882.

Vallenet, D., L. Labarre, Z. Rouy, V. Barbe, S. Bocs, S. Cruveiller, A. Lajus, G. Pascal, C. Scarpelli, and C. Médigue. 2006. MaGe - a microbial genome annotation system supported by synteny results. *Nucl. Acids Res.* **34:** 53-65.

Vanden Heuvel, B.D., D.R. Benson, E. Bortiri, and D. Potter. 2004. Low genetic diversity among *Frankia* spp. strains nodulating sympatric populations of actinorhizal species of Rosaceae, *Ceanothus* (Rhamnaceae) and *Datisca glomerata* (Datiscaceae) west of the Sierra Nevada (California). *Can J Microbiol* **50:** 989-1000.

Wayne, L.G., D.J. Brenner, R.R. Colwell, P.A.D. Grimont, O. Kandler, M.I. Krichevsky, L.H. Moore, W.E.C. Moore, R.G.E. Murray, E. Stackebrandt, M.P. Starr, and H.G. Truper. 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int. J. Syst. Bacteriol.* **37:** 463-464.

Young, J.P. and K.E. Haukka. 1996. Diversity and phylogeny of rhizobia. *New Phytol.* **133:** 87-94.

Zhang, Z., M.F. Lopez, and J.G. Torrey. 1984. A comparison of cultural characteristics and infectivity of *Frankia* isolates from root nodules of *Casuarina* species. *Plant and Soil* **78:** 79-90.

Zimpfer, J.F., C.A. Smyth, and J.O. Dawson. 1997. The capacity of Jamaican mine spoils, agricultural and forest soils to nodulate *Myrica cerifera, Leucaena leucocephala* and *Casuarina cunninghamiana*. *Physiol. Plant.* **99:** 664-672.

Table 1. Relationship between actinorhizal plants and groups of *Frankia* strains [a]

Order[b]	**Family (#nod/#gen)[c]**	**Genus (No. species)**	**Geographical Distribution[d]**	**Frankia Cluster[e]**
Fagales	Betulaceae (1/6)	*Alnus* (30)	N. temperate, SA, N. Africa, Asia	1, (3)
	Myricaceae (3/4)	*Comptonia* (1)	Eastern NA	1, 3
		Morella (20)	Cosmopolitan, not Aust or Mediterranean	1, 3
		Myrica (2)	Circumpolar	1 (3)
	Casuarinaceae (4/4)	*Allocasuarina* (58)	Australia	1
		Casuarina (17)	Australia	1 (3)
		Ceuthostoma (1)	Malesia	ND
		Gymnostoma (10)	Malaysia to W. Pacific	3
Rosales	Elaeagnaceae (3/3)	*Elaeagnus* (10)	Europe, Asia, NA	3
		Hippophae (2)	Temperate Eurasia	3
		Shepherdia (2)	NA	3
	Rhamnaceae (6/55)	*Ceanothus* (55)	Western NA	2, (3)
		Colletia (17)	Southern SA	3

		Discaria (15)	Southern SA, Aust., New Zealand	3
		Kentrothamnus (1)	Southern SA	3
		Retanilla (4)	Southern SA	3
		Trevoa (1)	Southern SA	3
	Rosaceae (4/100)	*Cercocarpus* (6-10)	Western NA	2
		Chamaebatia (2)	Western NA	2
		Dryas (2-3)	Circumboreal, arctic-alpine	2
		Purshia (8)	Western NA	2
Cucurbitales	Coriariaceae (1/1)	*Coriaria* (5-20)	Mexico to SA, W Mediterranean, New Zealand, Papua New Guinea, SE Asia	2
	Datiscaceae (1/1)	*Datisca* (2)	Western NA, S Asia	2

[a] Compiled after Benson et al. (2004) and Swensen (1996)

[b] According to the classification of the Angiosperm Phylogeny Group (APG 1998); all of these orders fall in the "Eurosid I" group of eudicots.

[c]The number of genera within the family is listed along with the number of genera nodulated. Not all genera within a family are capable of nodulation.

[d]NA, North America; SA, South America

[e]The *Frankia* cluster refers to the clusters of *Frankia* strains in Figure 1 typically found in root nodules of each plant genus. 1, alder and casuarina strains; 3, elaeagnus strains; 2, rosaceous strains; ND, not determined; (3) indicates rare detection of an elaeagnus strain in a nodule or in surface layers of a nodule (Benson et al. 2004)

Table 2. Summary of Genome characteristics[1]			
	ACN (finished)[3]	CcI3 (finished)[2]	EAN (draft-1 scaffold)[2]
Size in bp	7,497,934	5,433,628	9,035,218
Predicted CDS*	6786	4499	7976
Genes in COGs	4502 (67%)	2564 (57%)	4815 (60%)
tRNA	46	46	47
rRNA (5S-16S-23S)	2	2	3
%G+C	72.83	70.07	70.94
Accession	NC_008278	CP000249	NZ_AAII00000000

[1]Numbers are derived from the accession numbers indicated as of July 30, 2006.

[2]Numbers were from accessions indicated at http://www.ncbi.nlm.nih.gov except for the genes in COGs for EAN listed at http://www.jgi.doe.gov

[3]Numbers are from http://www.genoscope.cns.fr

Table 3. Summary of genes involved in size differentiation of the three *Frankia* strains.[1]

Category	ACN	CcI3	EAN
Deleted genes	466	1054	555
Duplicated genes	512 (7.5%)	444 (9.8%)	1355 (18.5%)
Strain Specific Genes (SSG)	709 (10.4%)	420 (9.3%)	934 (11.7%)
ORFans	854 (12.5%)	158 (3.5%)	355 (4.9%)
Transposases	33 (0.4%)	155 (3.4%)	195 (2.4%)
Integrases	13 (0.2%)	32 (0.7%)	74 (0.9%)

[1]Deleted genes, duplicated genes, ORFans (no hits in any database) and strain specific genes (SSG; hits in databases but not in the other two *Frankia* strains) were detected as described in the text. The number of genes annotated as transposases and inactivated derivatives plus integrases were annotated as described.

Table 4. Categories of deleted genes assessed by BlastClust analysis.[1]

General Category	CcI3	ACN	EAN
Hypothetical, Conserved hypothetical, Unknown Function	289	158	176
Transport associated	113	33	28
Regulatory	95	21	35
Short chain dehydrogenase/reductase	32	4	4
Acyl-CoA dehydrogenase-like	19	3	1
Alkanesulfonate monooxygenase	17	0	0
AMP-dependent synthetase and ligase	14	3	2
Protein kinase	12	2	6
Amidohydrolase	13	1	1
Enoyl-CoA hydratase/isomerase	10	1	2
L-carnitine dehydratase/bile acid inducible protein F	9	2	2
Alcohol dehydrogenase GroES-like	7	1	2
Cytochrome P450	8	5	3
Transposases	7	33	11
Integrases	3	7	6
Others	406	192	276

[1]Deleted genes are defined as genes present in two of the three strains as assessed by Blast hits below a cutoff of 10^{-4} but absent in the third. The general categories correspond to the major groups identified by a BlastClust analysis of missing genes.

Table 5. Top twenty families of duplicated genes in each *Frankia* strain assessed by BlustClust analysis

ACN		CcI3		EAN	
Annotated function	**#**	**Annotated function**	**#**	**Annotated function**	**#**
Putative serine/threonine protein kinase	22	Transposase, IS4	31	Short-chain dehydrogenase/reductase SDR	49
Putative oxidoeductase, short chain dehydrogenase/reductase family	20	Transposase IS66	15	Integrase, catalytic region	48
Hypothetical protein putative HNH endonuclease domain	17	Transposase, IS4	14	ABC-type branched-chain amino acid transport, periplasmic component	32
Hypothetical protein; putative dehydrogenase	16	Transposase	12	Transposase, IS605 OrfB	27
Putative oligopeptide transport	14	Regulatory protein,	11	Acyl-CoA dehydrogenase-like	23

protein (ABC superfamily)		MerR:Recombinase			
Putative SAM-dependent methyltransfeases	13	Twin-arginine translocation pathway signal	7	Cytochrome P450	22
Putative Alpha-methylacyl-CoA racemase	10	Transposase, IS4	7	Luciferase-like	21
Branched-chain amino acid ABC transport, binding protein	10	Transposase and inactivated derivatives-like	7	Integrase, catalytic region	20
Hypothetical protein; putative signal peptide	9	Transposase, IS111A/IS1328/IS1533:Transposase IS116/IS110/IS902	7	Amidohydrolase 2	20
Putative non-ribosomal peptide synthetase	8	Transposase (probable), IS891/IS1136/IS1341:Transposase, IS605 OrfB	7	Extracellular solute-binding protein, family 5	19
Branched-chain amino acid transport potein (ABC	6	Putative IS630 family transposase	7	Membrane-bound lytic murein transglycosylase B-like	17

superfamily)					
Conserved hypothetical protein; putative amidohydrolase domain	6	Hydantoinase/oxoprolinase	7	GGDEF domain	14
Putative GntR-family transciptional regulator	6	ATP-binding region, ATPase-like	6	Putative transposase	13
Putative TetR-family transciptional regulator	5	Putative O-methyltransferase	5	Transposase, IS4	13
Hypothetical protein; putative dibenzothiophene desulfurization	5	Putative plasmid replication initiator protein	5	ABC transporter related	12
Cytochrome P450	5	Transposase, IS4	4	Regulatory protein, LuxR	12
Hypothetical protein	5	Hypothetical protein	4	RNA-directed DNA polymerase	11
Putative monooxygenase	4	Amino acid adenylation	3	Phenylpropionate dioxygenase/related ring-hydroxylating dioxygenases	11
Hypothetical protein	4	Putative DNA-binding protein	3	Taurine catabolism dioxygenase	11

				TauD/TfdA	
Putative aldehyde dehydrogenase	4	Hypothetical protein	3	Regulatory protein, TetR	11

[a]The top twenty categories of duplicated genes were defined by a BlastClust analysis of a dataset comprised of all duplicates. Duplicates were defined as having the best BLAST score of another gene within the same genome relative to genes within NR plus the other *Frankia* strains, *Acidothermus, Kineococcus*, and *Streptomyces* spp. with a maximum cutoff of 10^{-4}.

Figure Legends

Figure 1. Phylogenetic tree based on 16S rRNA gene analysis. Neighbor joining (Saitou and Nei 1987) tree calculated with Clustalx 1.83 (Thompson et al. 1997). from 16S rRNA gene sequences. Distances were corrected for multiple substitutions(Kimura 1980), otherwise default settings were used. Numbers give bootstrap support values from 1000 bootstrapped samples. The outgroup used is *Streptomyces coelicolor* (NC003888). Accession numbers for the organisms are given after the name and species number as given in Normand and Fernandez (2007). In the case of the unisolated cluster 2 frankiae, the host plant genus from which 16S rRNA gene sequences were amplified is given. Distances in the bar are in substitutions/site.

Fig. 2. Present day native distribution of actinorhizal plant hosts. a, Distribution of plant hosts for CcI3 including *Casuarina* and *Allocasuarina* of the Casuarinaceae (C in the legend). **b,** Distribution of plant hosts for ACN including *Alnus* sp. in the Betulaceae (B) and Myricaceae (M) and their overlap (M+B). **c,** Distribution of plant hosts for EAN including members of the Elaeagnaceae (E), Myricaceae (M) and the actinorhizal Tribe Colletieae of the Rhamnaceae in South America, Australia and New Zealand (R). Elaeagnaceae and Myriceae (E+M) overlap in some areas. Maps were drawn with information from Silvester (1977) and from the Missouri Botanical Garden website (www.mobot.org).

Fig. 3. Genome maps of the three *Frankia* strains. Circles, from the outside in, show (1) gene regions related to symbiosis including *shc*1, *hup2*, *hup1* and *nif*, (2) the coordinates in Mb beginning at 0=oriC, (3) regions of synteny (syntons) calculated as a

minimum of 5 contiguous genes present in all strains with an identity > 30% over 80% of the length of the shortest gene (syntons are tagged with a spectrum-based (red-yellow-green) color-code standardized on ACN to indicate regions where syntons have moved in the other strains), (4) IS elements and transposases. Circles were drawn using GenVision Software from DNAStar (Madison, Wisconsin).

Figure 1.

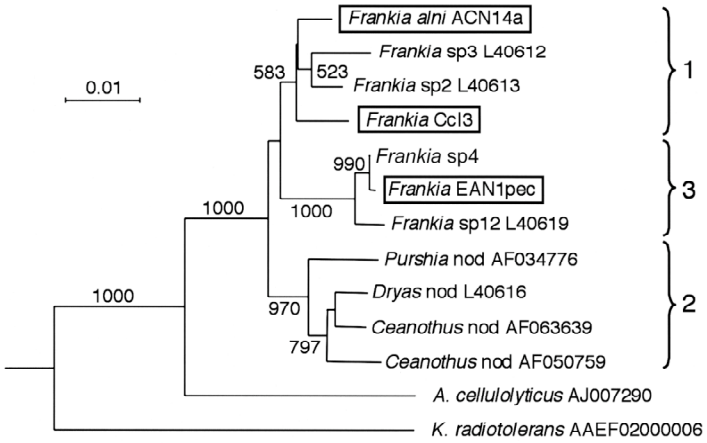

Frankia alni ACN14a
Frankia sp3 L40612
Frankia sp2 L40613
Frankia Ccl3
Frankia sp4
Frankia EAN1pec
Frankia sp12 L40619
Purshia nod AF034776
Dryas nod L40616
Ceanothus nod AF063639
Ceanothus nod AF050759
A. cellulolyticus AJ007290
K. radiotolerans AAEF02000006
0.01
583
523
990
1000
1000
1000
970
797
1
3
2

Figure 2.

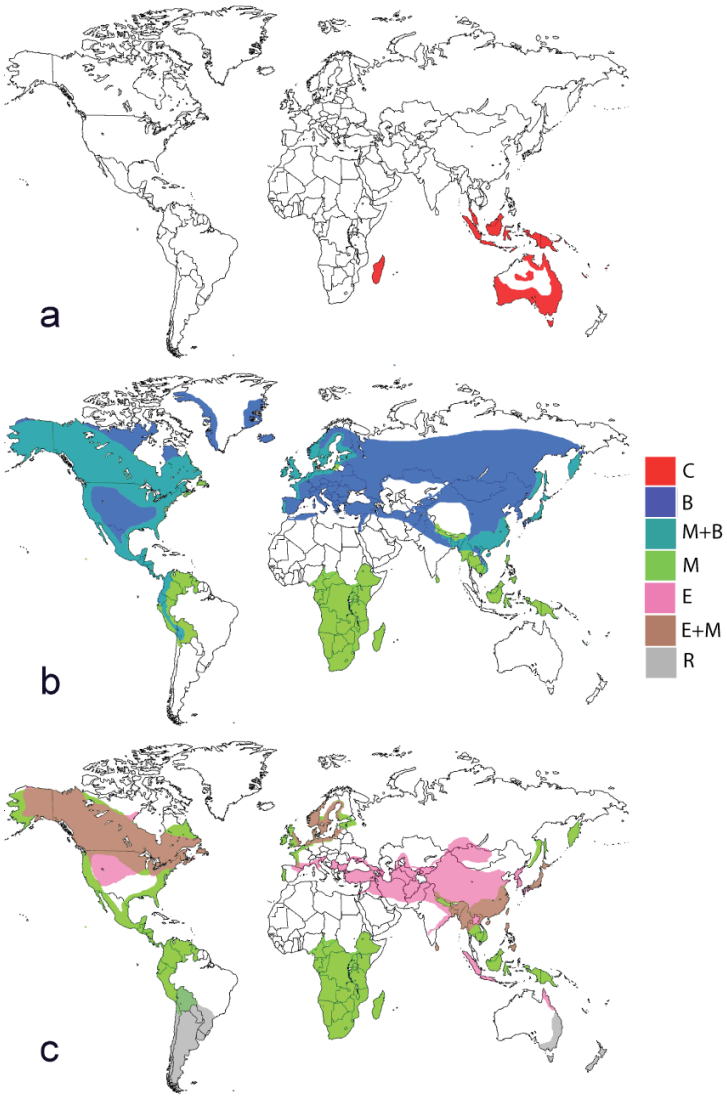
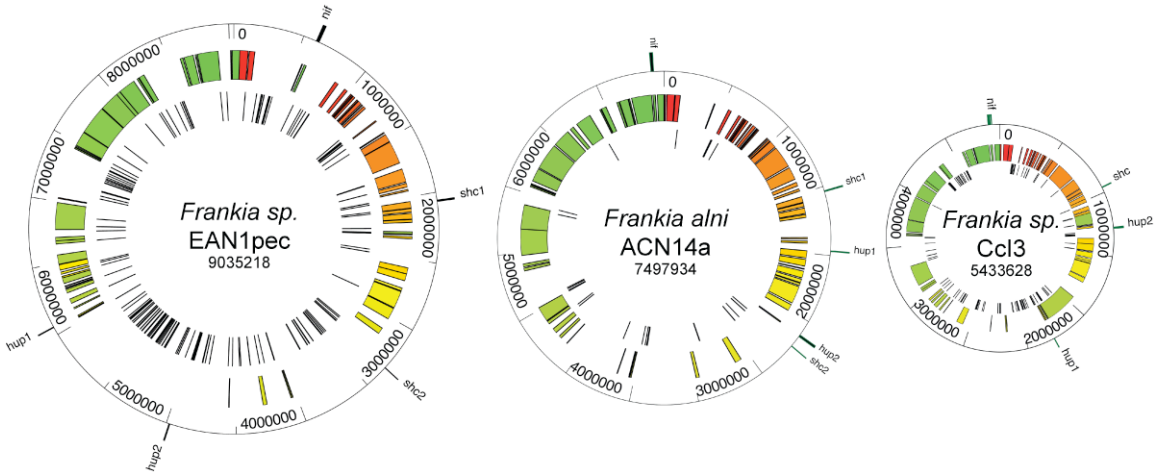


Figure 3.


Frankia sp.
EAN1pec
9035218
Frankia alni
ACN14a
7497934
Frankia sp.
Ccl3
5433628