nature biotechnology

Complete genome of the mutualistic, N_2-fixing grass endophyte *Azoarcus* sp. strain BH72

Andrea Krause[1], Adarsh Ramakumar[1], Daniela Bartels[2], Federico Battistoni[1], Thomas Bekel[2], Jens Boch[3], Melanie Böhm[1], Frauke Friedrich[1], Thomas Hurek[1], Lutz Krause[2], Burkhard Linke[2], Alice C McHardy[2,6], Abhijit Sarkar[1], Susanne Schneiker[2,4], Arshad Ali Syed[1], Rudolf Thauer[5], Frank-Jörg Vorhölter[2,4], Stefan Weidner[2], Alfred Pühler[2,4], Barbara Reinhold-Hurek[1], Olaf Kaiser[2,4,7] & Alexander Goesmann[2,7]

Azoarcus **sp. strain BH72, a mutualistic endophyte of rice and other grasses, is of agrobiotechnological interest because it supplies biologically fixed nitrogen to its host and colonizes plants in remarkably high numbers without eliciting disease symptoms. The complete genome sequence is 4,376,040-bp long and contains 3,992 predicted protein-coding sequences. Genome comparison with the *Azoarcus*-related soil bacterium strain EbN1 revealed a surprisingly low degree of synteny. Coding sequences involved in the synthesis of surface components potentially important for plant-microbe interactions were more closely related to those of plant-associated bacteria. Strain BH72 appears to be 'disarmed' compared to plant pathogens, having only a few enzymes that degrade plant cell walls; it lacks type III and IV secretion systems, related toxins and an N-acyl homoserine lactones–based communication system. The genome contains remarkably few mobile elements, indicating a low rate of recent gene transfer that is presumably due to adaptation to a stable, low-stress microenvironment.**

Endophytic bacteria reside within the living tissue of plants without substantively harming them. They are of high interest for agrobiotechnological applications, such as the improvement of plant growth and health, phytoremediation[1] or even as biofertilizer[2]. Supply of nitrogen derived from fixation of atmospheric N_2 by grass endophytes, such as *Gluconacetobacter diazotrophicus* and *Azoarcus* sp. strain BH72, which has been shown to occur in sugarcane[3] and Kallar grass[2], is a process of potential agronomical and ecological importance.

Although the lifestyle of these endophytes is relatively well documented, the molecular mechanisms by which they interact beneficially with plants have only been poorly elucidated. A combination of features makes *Azoarcus* sp. strain BH72 an excellent model grass-endophyte[4]. (i) It supplies nitrogen derived from N_2 fixation to its host, Kallar grass (*Leptochloa fusca* (L.) Kunth); *in planta* it is usually not culturable, but can be detected by culture-independent methods based on *nifH*-encoding nitrogenase reductase, the key enzyme for N_2 fixation[2]. (ii) It colonizes nondiseased plants in remarkably high numbers: estimates range from 10^8 cells (culturable cells per gram root dry weight (RDW) of field-grown Kallar grass[5]) to 10^{10} cells (estimated on the basis of abundance of bacterial *nifH*-mRNA in roots)[2]. (iii) It is the only cultured grass endophyte shown by molecular methods to be the most actively N_2-fixing bacterium of the natural population in roots[2]. (iv) It also colonizes the roots of rice, a cereal of global importance, in high numbers (10^9 cells per g RDW) in the laboratory, and spreads systemically into shoots[6]. Plant stress response is only very limited in a compatible, that is, well-colonized rice cultivar[7]. Notably, *Azoarcus* sp. strain BH72 is capable of endophytic N_2-fixation inside the roots of rice[8].

For a wider application in agriculture, more knowledge is required on mechanisms of interaction and host specificities. Although the genome of a related species, strain EbN1, belonging to a branch of *Azoarcus* species that typically occurs in soils and sediments but not in association with plants[9], is available[10], phenotypic differences and phylogenetic distances of 5–6% suggest they might deserve the rank of a separate genus in future[9]. The plant-associated strain BH72—like many N_2-fixing endophytes grass endophytes—has not been detected in root-free soil[11]. In this study, we present the complete genome sequence of a diazotrophic grass endophyte, *Azoarcus* sp. strain BH72, and highlight features that may contribute to knowledge of the endophytic lifestyle of these plant-beneficial bacteria, which may be instrumental in developing biotechnological applications.

RESULTS

General features of the genome and mobile elements

The *Azoarcus* sp. strain BH72 genome sequence was obtained with a whole genome shotgun approach, the assembly being validated by a complete fosmid (**Fig. 1b**) and a bacterial artificial chromosome

[1]Laboratory of General Microbiology, University of Bremen, PO Box 330440, D-28334 Bremen, Germany. [2]Center for Biotechnology (CeBiTec), Bielefeld University, PO Box 100131, D-33501 Bielefeld, Germany. [3]Institut für Genetik, Martin-Luther-Universität, Weinbergweg 10, D-06120 Halle/Saale, Germany. [4]Lehrstuhl für Genetik, Bielefeld University, PO Box 100131, D-33501 Bielefeld, Germany. [5]Max-Planck-Institute for Terrestrial Microbiology, Karl-von-Frisch-Strasse, D-35043 Marburg, Germany. [6]Present address: Bioinformatics & Pattern Discovery Group, IBM Thomas J Watson Research Center, Yorktown Heights, New York 10598, USA. [7]These authors contributed equally to the work. Correspondence should be addressed to Barbara Reinhold-Hurek (breinhold@uni-bremen.de).

Received 22 May; accepted 4 August; published online 15 October 2006; corrected after print 26 March 2007; doi:10.1038/nbt1243

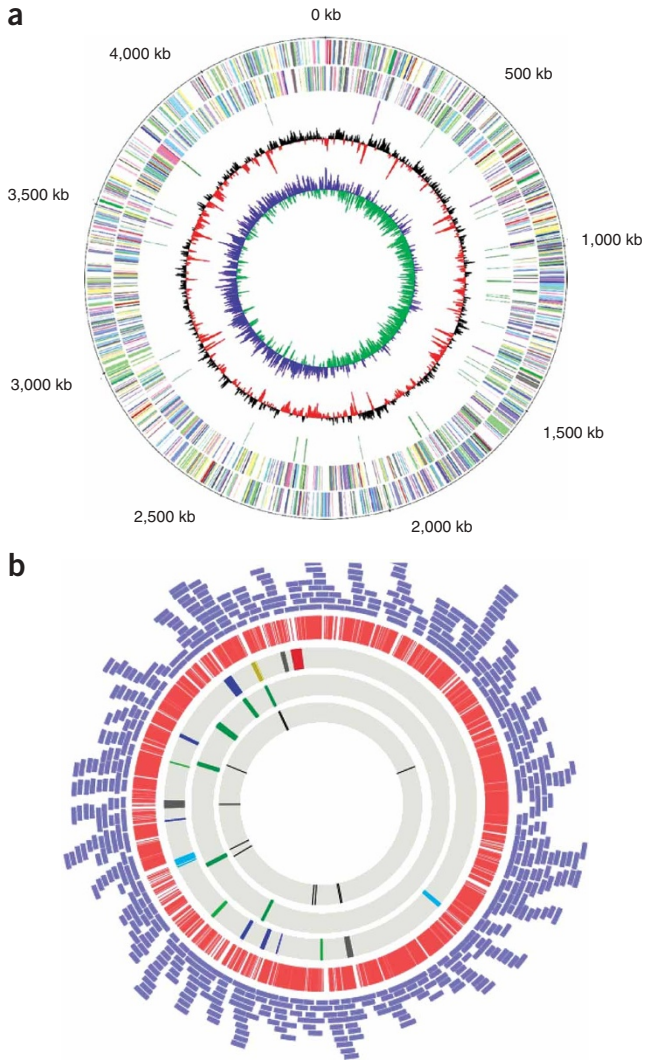

Figure 1 Circular representation of the *Azoarcus* sp. strain BH72 genome displaying relevant genome features and validation of the sequence assembly by a fosmid map. In the final consensus sequence each base matched at least phred40 quality. (**a**) From the outer to the inner concentric circle: circle 1, genomic position in kilobases; the origin of replication was clearly detectable by a bias of G toward the leading strand (GC skew); the start of the *dnaA* gene located in this region was defined as zero point of the chromosome; circles 2 and 3, predicted protein-coding sequences (CDS) on the forward (outer wheel) and the reverse (inner wheel) strands colored according to the assigned COG (clusters of orthologous groups) classes; leading strand, 2,074 CDS = 52.0%; lagging strand, 1,918 CDS = 48.0%; circle 4, tRNA genes (green) and the four rRNA operons (pink); circle 5, the G+C content showing deviations from the average (67.92%); circle 6, GC skew; a bi-directional replication mechanism suggested by a clear division into two equal replichores. (**b**) Fosmid map of the *Azoarcus* sp. strain BH72 chromosome. Each blue arc represents a single fosmid clone mapped onto the assembled sequence; circle 1, CDS with homologs in the chromosome of *Azoarcus* sp. *strain* EbN1 (e-value below e^{-30}); circle 2, gene clusters coding for surface-related proteins or other functions not related to proteins of *Azoarcus* sp. strain EbN1: exopolysaccharide/ lipopolysaccharide-related and pilus-related gene clusters (blue), flagella and chemotaxis related gene clusters (light blue), virulence-related gene clusters (red), proteins related to metabolism (gray), conserved hypothetical proteins or other proteins related to proteins of rhizobia or plant commensals, and various genes not present in *Azoarcus* sp. strain EbN1 (gold); circle 3, putative genomic islands predicted by the Pai-Ida program 1.1 (score > 3.8); circle 4, transposases and phage-related genes.

(BAC) map[12]. Characteristics of the single, circular chromosome and the predicted genes are shown in **Figure 1** and **Table 1**.

The genome contains remarkably few phage- or transposon-related genes, indicating a low degree of lateral transfer and genome rearrangements; just eight loci (**Fig. 1b**, circle 4) contain genes for integrases, recombinases, transposases or phage-related genes (**Supplementary Table 1** online). Only a few loci correspond to predicted anomalous gene clusters or putative pathogenicity islands (**Fig. 1b**, circle 3). In contrast, the genome of the *Azoarcus*-related soil isolate strain EbN1 contains 237 transposon-related genes[10]. Also rhizobial genomes harbor >100 transposases or phage-related genes (http://www.kazusa.or.jp/rhizobase/). Likewise, many plant-pathogenic proteobacteria contain high numbers of mobile elements[14]. High genomic plasticity might reflect the need for continuous adaptation to changing environments like soil or to host defense mechanisms. For nodule symbionts, soil is an alternative habitat in their life cycle; in contrast typical grass endophytes can not usually be isolated from root-free soil[11,13]. The comparatively low number of mobile elements in the endophyte BH72 might indicate a low rate of recent gene transfer and genome rearrangements, which is presumably due to adaptation to a stable, low-stress microenvironment inside plants.

Comparative genomics

Genome comparison revealed a surprisingly low degree of synteny between genomes of strain BH72 and the *Azoarcus*-related strain EbN1 (**Fig. 2**). At a low cutoff e-value of e^{-30}, the majority of predicted proteins (58%) in strain BH72 have some counterparts in strain EbN1 (**Fig. 1b**, circle 1). However, only 43% of these proteins were more closely related to those of EbN1 than to proteins of other strains. Other pathogenic or plant symbiotic proteobacteria have even less related genomes (**Supplementary Table 2** online). Because strains BH72 and EbN1 have a very different ecology, the differences may give important hints as to which genes are required specifically for the endophytic lifestyle. Several gene clusters of strain BH72 that are

Table 1 Genome features of the N_2-fixing endophyte *Azoarcus* sp. strain BH72 in comparison to the denitrifying soil bacterium *Azoarcus* sp. strain EbN1

Feature	*Azoarcus* sp. BH72	*Azoarcus* sp. EbN1
Size of chromosome (bp)	4,376,040	4,296,230
Plasmids	0	2[a]
G+C content, %	67.92	65.12
Coding sequences	3,992	4,133
Function assigned	3,418	2,560
Conserved hypothetical protein	517	628
Hypothetical protein	57	945
% of genome coding	91.2	90.9
Average length (bp)	999	945
Maximal length (bp)	6,330	6,132
% ATG initiation codons	86.57	76.46
% GTG initiation codons	10.40	16.01
% other initiation codons	n.d.[b]	n.d.
RNA elements		
rRNA operons	4	4
tRNAs	56[c]	58

[a]Plasmid 1 (207,355 bp), plasmid 2 (223,670 bp). [b]n.d., not determined. [c]One $tRNA^{Ile}$ (azo_tRNA_0051) is disrupted by a self-splicing group I intron in the CAT anticodon loop[42].

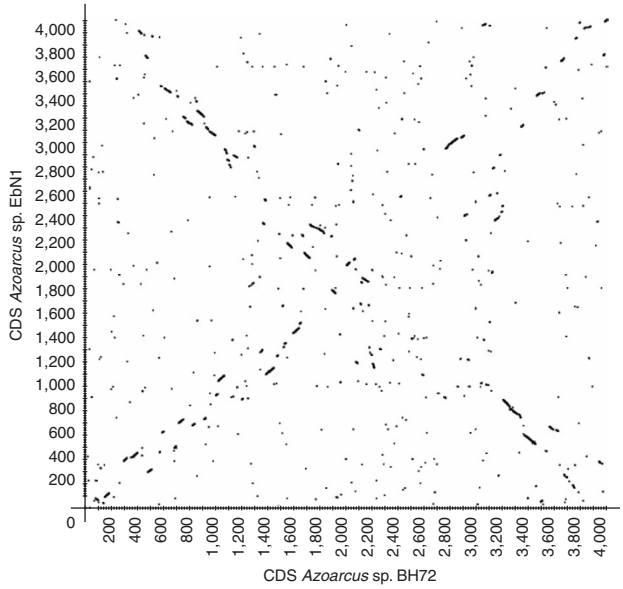

Figure 2 Synteny between the genomes of *Azoarcus* sp. strain BH72 and *Azoarcus* sp. strain EbN1. The genome of *Azoarcus* sp. strain EbN1 is adjusted to *dnaA* with its start codon as zero point of the chromosome. The diagram depicts *x-y* plots of dots forming syntenic regions between the two *Azoarcus* genomes. Each dot represents an *Azoarcus* sp. strain BH72 CDS having an ortholog in *Azoarcus* sp. strain EbN1, with coordinates corresponding to the CDS number in each genome. The orthologs were identified by best BLASTP matches of amino acid sequences (e-value $< e^{-30}$).

lacking in EbN1 or that are more similar to genes of other bacteria (**Fig. 1b**) harbor genes that encode proteins putatively involved in cell surface components or other features that may be required for the endophytic lifestyle (see below, **Fig. 3** and **Supplementary Table 1** online).

Carbon metabolism and signal transduction

Strain BH72 has a strictly respiratory type of metabolism and does not grow on any carbohydrate tested[9,15]. Aspects of putative carbon metabolism are shown in **Figure 4**. The inability to utilize common carbohydrates might contribute to a plant-compatible endophytic lifestyle because, in contrast to phytopathogens, the bacteria cannot grow and proliferate on the major cell wall constituents although a cellulase is present[16].

The major carbon sources for strain BH72 are dicarboxylic acids and ethanol[9]. Transport systems for C4-dicarboxylates (**Fig. 4**) might be of vital importance during the association with host plants, as in symbiotic rhizobia[17]. Ethanol might be important for association with flooded plants like rice, which accumulate ethanol under anoxic conditions, especially at root tips—one of the typical sites of colonization of strain BH72. Correspondingly, its genome harbors ten genes encoding putative alcohol dehydrogenases.

Strain BH72, despite being adapted to a relatively stable, low-stress microenvironment, shows a remarkable density of signal transduction systems (see details in **Supplementary Table 3** online). Thus it may be a good example for sophisticated signal transduction networks.

N_2 fixation and nitrogen metabolism

Azoarcus sp. strain BH72 appears to be highly adapted to environments poor in available nitrogen sources, which correlates with its role as an N_2-fixing endophyte (**Fig. 4**). (i) A low-affinity glutamate dehydrogenase (GDH) for ammonium assimilation is lacking, a feature highly unusual in free-living bacteria, whereas it is present in strain EbN1. Only the high-affinity ATP-consuming assimilation system (GS[2x]-GOGAT) is present. (ii) Four genes encoding high-affinity ammonia transporters exist (*amtB/Y/D/E*), one of them with an additional regulatory domain. (iii) In contrast to the soil strain EbN1, structural genes for the molybdenum-dependent nitrogenase complex and all genes required for cofactor synthesis and maturation of the nitrogenase are present in strain BH72, one of them in two copies (*nifY*). Several putative low-potential electron donors for N_2 fixation were identified including two flavodoxin-encoding genes (*nifF1, nifF2*), 12 genes for ferredoxin-like proteins; two clusters encoding putative electron transport systems (*rnf1, rnf2*) might be instrumental for electron supply to ferredoxin during N_2 fixation. Several genes likely to be involved in the regulatory cascade are also listed in **Supplementary Table 1** online. Although in pure culture, N_2-fixing strain BH72 does not excrete substantial amounts of nitrogenous compounds[18], it supplies fixed nitrogen to its grass host[2]. The four ammonia transport proteins are putative candidates for export to the plant. Two transport systems for glutamate or glutamine as well as nine for branched-chain amino acids were also identified; however, the presence of periplasmic substrate–binding proteins suggests that these systems are used for import and not for export.

About 38 genes encoding enzymes and transporters involved in nitrate metabolism were identified (**Fig. 4** and **Supplementary Table 1** online). As in strain EbN1, genes required for assimilatory nitrate and

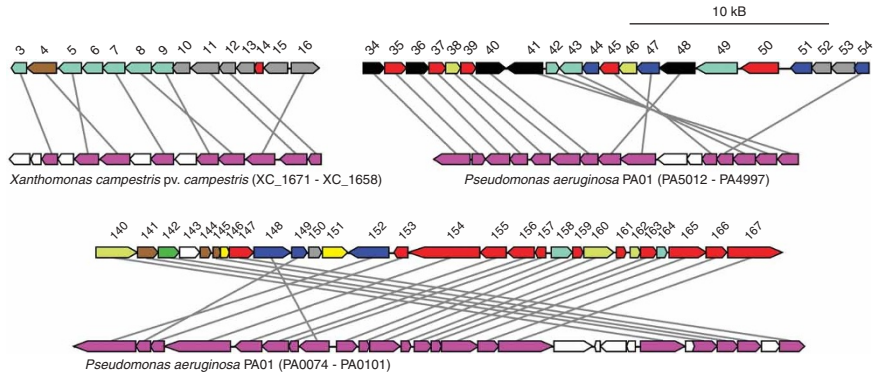

Figure 3 Gene clusters in *Azoarcus* sp. strain BH72 that are lacking in EbN1 or are more similar to genes of other bacteria. Synteny of selected clusters with gene clusters of other bacteria: *gum*-cluster, *Xanthomonas campestris* pv. *campestris* (locus_tag numbers XC_1671 - XC_1658); lipopolysaccharide-related cluster, *Pseudomonas aeruginosa* PA01 (acc. numbers PA5012 - PA4997); *sci*-cluster, *Pseudomonas aeruginosa* PA01 (acc. numbers PA0074 - PA0101); pink/white, genes present or not present in the gene cluster of strain BH72, respectively. Numbers refer to genes of *Azoarcus* sp. strain BH72 listed in **Supplementary Table 1** online. Highest similarities to proteins of other bacteria are depicted by the following colors: red, human or animal pathogens; green, root nodule symbionts (rhizobia); yellow-green, root-associated bacteria; turquoise, plant pathogens; other bacteria according to their phylogenetic affiliations: black, *Azoarcus* sp. strain EbN1; gray, beta-subgroup of *Proteobacteria*; blue, gamma-subgroup of *Proteobacteria*; yellow, alpha subgroup of *Proteobacteria*; brown, others.

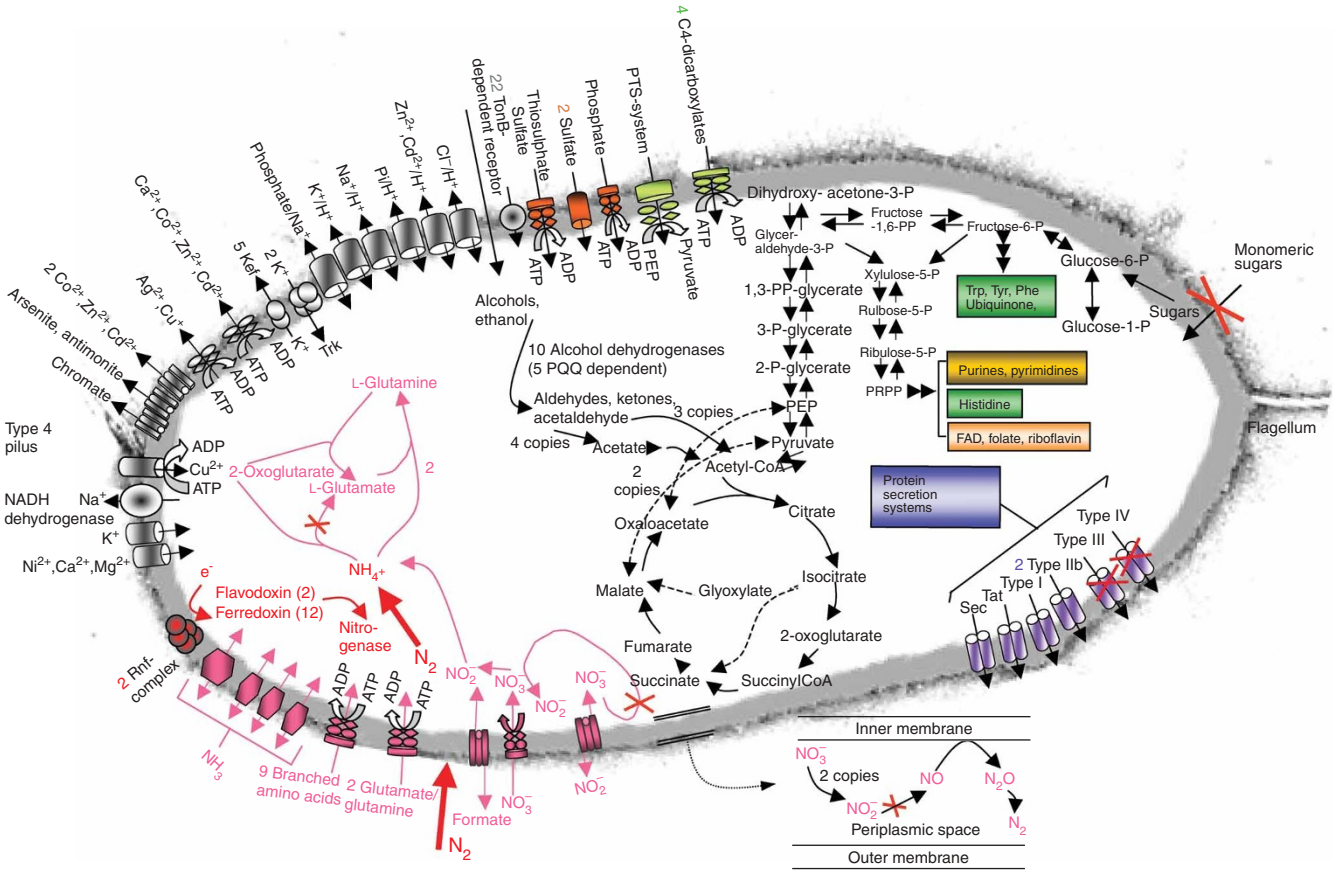

Figure 4 Overview of physiological features of *Azoarcus* sp. strain BH72. Depicted are carbohydrate metabolism, selected transporters and catabolic pathways for organic acids, nitrogen assimilatory and dissimilatory pathways, inorganic ion transport systems, and protein translocation systems. Crosses indicate pathways or reactions that are apparently not present in strain BH72. Details of metabolic features: the genome does not contain genes required for a functional Entner-Doudoroff pathway or the oxidative branch of the pentose phosphate pathway, however, the nonoxidative pentose phosphate pathway is complete. All enzymes required for glycolysis via the Embden-Meyerhoff pathway, gluconeogenesis and the TCA cycle are present, including genes encoding a typical phosphoenolpyruvate: sugar phosphotransferase system (PTS), but no specific outer membrane transporter for carbohydrates. Four complete and two incomplete copies of the TRAP-transport system (DctPQM) for C4-dicarboxylates are present. C4-dicarboxylates are probably metabolized via the glyoxylate shunt pathway, malate being decarboxylated by the malic enzyme (MaeB, two gene copies). Ten putative alcohol dehydrogenases are encoded in the genome (see also **Supplementary Table 1** online), five of which are PQQ-dependent enzymes not present in strain EbN1.

nitrite transport and reduction are present in strain BH72 in one copy (*nasFED*, *nirC*, *nasA*, *nasC*, *nirBD*). In contrast to strain EbN1, strain BH72 cannot conduct denitrification to N_2, as genes required for a nitrite reductase are missing in strain BH72. However, genes for subsequent denitrification reactions (*norCBQD* and *nosZRDFYL* encoding reductases for nitric oxide and nitrous oxide, respectively) are present. The periplasmic localization of the nitrate reductase components (duplicated *nap* operon) and detoxification of NO_2^- via an NO_3^-/NO_2^- antiporter (NarK) might decrease toxicity of nitrite. In contrast, the denitrification pathway is complete in strain EbN1 for which denitrification appears to be a typical feature[10].

Iron-transport related proteins

In Gram-negative bacteria, TonB-dependent, outer-membrane proteins are responsible for the specific uptake of ferric-siderophore complexes, high-affinity iron chelators. They are also important for perception of environmental signals and are associated with pathogenicity of plant pathogens[19]. Strain BH72 possesses 22 genes encoding proteins related to iron transport (**Supplementary Table 1** online), twice as many as described for strain EbN1 and even more than other N_2-fixing endosymbionts (*Bradyrhizobium japonicum*, 13; *Sinorhizobium meliloti*, 9; and *Mesorhizobium loti*, 1). Two genes (*azo2156*, *azo3836*) are not even present in the *P. fluorescens* Pf5 genome, a plant-associated bacterium known for its capacity to produce and take up a wide range of siderophores[20], which contains 45 such genes. Although putative receptors for hydroxamate- and catechol-type siderophores, ferricitrate, vitamin B12, colicins and unknown substances are present, there was no evidence for biosynthetic pathways for known hydroxamate or catecholate siderophores[21]. Moreover, production of siderophores was not detected experimentally (**Supplementary Fig. 1a** online). Apparently, this strain is highly adapted to obtaining chelated iron from other sources, as fungi and monocotyledonous plants also produce siderophores[22]. The high number of putative receptors suggests a role not only for rhizosphere competence of strain BH72, but also for biocontrol.

Plant-associated lifestyle

Surface characteristics of bacteria are important factors for recognition by and interaction with the host. Several gene clusters putatively related to surface components of strain BH72 are lacking in strain EbN1, or are more highly related to genes of plant-associated or pathogenic bacteria.

Type IV pili are among the few factors known to affect endophytic colonization of grass diazotrophs[23]. Establishment of microcolonies on roots and fungal mycelium, and systemic spreading in rice are mediated by type IV pili[23]. The strain BH72 genome harbors 41 genes encoding proteins putatively involved in pilus assembly and regulation, whereas only 30 such genes were found in strain EbN1. Genes highly similar in both species encode proteins with conserved function such as assembly or regulation (for example, PilBCD-PilF-PilM-NOPQ-PilTU-PilZ or PilSR-PilGHIL). Other pilus proteins related mostly to phytopathogenic bacteria might be either pseudopilins involved in secretion (PilV/W/X) or putative tip adhesins (PilY1A/B, 31% and 39% sequence identity to *Ralstonia solanacearum* and *Xylella fastidosa*, respectively) that are lacking even in strain EbN1 and might be characteristic for interaction with plant surfaces (**Supplementary Table 1** online).

Other cell surface components that are often involved in recognition or virulence of pathogens are lipopolysaccharides, exopolysaccharides and capsular material. Intriguingly, many gene products putatively involved in their synthesis in strain BH72 are not highly related to those of the soil isolate EbN1, but to proteins of plant symbionts, pathogens or gamma- or alpha-proteobacteria (**Supplementary Table 1** online). Several genes of strain BH72 are similar to the *gum* operon for exopolysaccharide production in phytopathogenic *Xanthomonas campestris*[24]; genes encoding putative glycosyl transferases have considerable similarity to the rhizobial *pss* gene cluster (**Supplementary Table 1** online, 20–23), which is involved in exopolysaccharide polymerization, translocation and thus in plant-microbe interaction; a gene cluster related mainly to lipopolysaccharide synthesis is most similar to genes of gamma-proteobacteria including pathogens; these clusters did not show sufficient synteny to support the assumption of a very recent gene transfer (**Fig. 3**).

Motility. Flagella are pivotal for motility, adhesion, biofilm formation and colonization of the host. Strain BH72 is highly motile by means of a polar flagellum. At least 48 genes were identified that are generally required for biosynthesis and function of flagella and chemotaxis. They are located in three different noncontiguous clusters and are mostly related to genes of other beta-proteobacteria and a few pathogens of the gamma-subgroup (**Supplementary Table 1** online). There are three genes encoding flagellins (*fliC1, fliC2, fliC3*) and two encoding flagellar motor proteins, suggesting an important role for motility in the plant-associated lifestyle. In contrast, the nonmotile strain EbN1 does not possess a complete flagellar regulon[10].

Secretion and communication. Several genes encoding potential protein secretion systems were identified in the genome of strain BH72 (**Fig. 4**) for a sec-dependent pathway, a signal recognition particle (SRP)-mediated translocation and a twin arginine translocation (Tat) system, all of them targeting proteins through the inner membrane. Secretion of proteins through the entire cell envelope seems to be limited to only three varieties of pathways. Genes were identified that encode one type I secretion system and one autotransporter. Two gene clusters were detected that encode a type II secretion-related system (type IIb secretion system[25]), consisting only of GspDEFG.

Two other secretion systems are common to plant-associated bacteria, type III and IV secretion systems, which transport a wide variety of effector proteins into the extracellular medium or into the cytoplasm of eukaryotic host cells and affect interaction[26–29]. Intriguingly, neither system is present in strain BH72, probably preventing the export of toxic proteins to the host.

'Quorum sensing' is a common way of bacteria to communicate with each other or hosts by means of autoinducers that accumulate in the extracellular environment in a cell density-dependent manner[30]. Although there is evidence that autoinducer-dependent gene regulation occurs in *Azoarcus* sp. strain BH72 (Böhm, M. & Reinhold-Hurek, B., unpublished data), this strain appears to escape the usual communication systems. Widespread autoinducers of Gram-negative bacteria are N-acyl homoserine lactones (AHLs)[30]. There is no evidence for genes encoding an AHL-based quorum-sensing system in strain BH72; genes encoding the autoinducer synthetase (LuxI/LasI–type) or the responsible cytoplasmic autoinducer receptor (LuxR/LasR-type) are lacking. Furthermore, different bacterial sensor strains detecting presence of short- or long-chain AHLs did not yield a positive response toward strain BH72 (**Supplementary Fig. 1b** online). Also genes encoding the autoinducer-2 synthetase LuxS[30] are lacking. Gram-positive bacteria usually use peptides as autoinducers[30]. Genes expected for this system were not detected in strain BH72 either.

Virulence and interaction factors. The strain BH72 genome stands out by the lack of obvious genes involved in production of toxins. Moreover, common hydrolytic enzymes that macerate plant cell wall polymers and thus contribute to a phytopathogenic lifestyle and plant damage are rare: pectinase-encoding genes are absent; only a few genes encode putative glycosidases (*palZ*, *spr1*, *ndvC*, *eglA*), some of them with transmembrane helices. Detection of genes for membrane-bound enzymes is in agreement with the observation that strain BH72 does not secrete cellulases into the culture medium, but shows activities of a cell surface-bound endoglucanase (EglA) and exoglycanase that also hydrolyses xylosides[16], a major component of primary cell walls in grasses. A low production of macerating enzymes is likely to contribute to compatibility with the plant host, however these hydrolases might assist in endophytic colonization, as shown for the endoglucanase EglA[31].

There is no genomic evidence for the central process in the rhizobium-legume symbiosis, the induction of nodulation. Common *nodABC* genes required for the biosynthesis of the Nod-factor backbone are not present in strain BH72; only a few genes show some sequence similarity to other *nod* or *nol* genes (**Supplementary Table 1** online). However, like other grass-associated microbes such as *Azospirillum* a gene similar to *nodD* is present, in rhizobia encoding a central regulator for flavonoid-inducible gene expression.

Some other gene clusters in strain BH72 are also interesting targets for putative roles in plant-microbe interactions. One cluster shows similarity to genes that are localized in the *sci*-genomic island, affecting virulence of human pathogens. The genomic organization shows remarkable synteny with genes of *P. aeruginosa* (**Fig. 3**), arguing for a more recent gene transfer. Interestingly some homologs are also present in rhizobia. Other noticeable gene clusters code for mainly conserved hypothetical proteins or metabolism-related proteins that have orthologs in rhizobia, or of regulatory proteins (ColRS) important for root colonization of commensals (**Supplementary Table 1** online, no. 177–185, 188–196, 198–200).

DISCUSSION

The complete genome sequence of *Azoarcus* sp. strain BH72 offers insights into genomic strategies for an endophytic life style, and allows identification of various features that may contribute to their interaction with plants. The strain appears to be adapted to a relatively stable, low-stress microenvironment, since its genome contains remarkably few phage- or transposon-related genes in comparison to many soil bacteria or pathogens, indicating a low plasticity of the

genome. The lack of the typical communication system of Gram-negative bacteria based on AHLs also argues for a rather exclusive microhabitat.

Strain BH72 appears to be disarmed compared to plant pathogens by its in ability to metabolize carbohydrates coupled with the lack of a massive occurrence of cell-wall degrading enzymes. Moreover, this bacterium lacks known toxins and type III and IV secretion systems that are typically used by pathogens to transport effector molecules to their host. This might be instrumental for avoiding damage to the plant host despite a dense internal colonization, and only the small set of hydrolases identified may be required for penetrating into the plant tissue. Some specific features of nodule symbionts are also lacking, like most *nod*/*nol* genes required for nodule induction.

Genome comparison with the *Azoarcus*-related, nondiazotrophic, nonendophytic soil bacterium strain EbN1 revealed features likely to be important for plant-microbe interaction. Strain BH72 appears to be highly adapted to environments of low nitrogen availability, N_2-fixation playing a key role in its ecology. Several gene clusters that were lacking in strain EbN1 or were highly similar to genes of plant-associated or pathogenic bacteria were related to cell surface components that are often involved in recognition—gene products participating in the synthesis of exopolysaccharide, lipopolysaccharide, type IV pilus tips, the flagellar and chemotaxis apparatus. Further targets for studying interaction mechanisms were also identified by comparative genomics, such as virulence-related *sci* genes or genes encoding conserved hypothetical proteins shared with nodule symbionts. A large and diverse set of TonB-dependent receptors (22) might play a role in iron acquisition and biocontrol. In future functional genomic analyses, the role of these target genes for host compatibility will be elucidated, which is crucial for a wider agro-biotechnological application of N_2-fixing endophytes.

METHODS

Whole genome shotgun sequencing. DNA shotgun libraries with insert sizes of 1 kb and 2–3 kb in pGEM-T (Promega), and 8-kb fragments in pTrueBlue-rop (MoBiTec) vectors were constructed by MWG Biotech. Plasmid clones were end-sequenced on ABI 3700 sequencers (ABI) by MWG Biotech AG. Base-calling was carried out using PHRED[32,33]. High-quality reads were defined by a minimal length of 250 bp with an averaging quality value of $\geq$20. Finally, 60,715 high-quality reads, a total of 39,266 (5.26 ×), 18,070 (2.32 ×) and 3,379 (0.40 ×) end sequences (×'s indicate genome equivalents) from libraries with 1-kb, 2- to 3-kb and 8-kb inserts, respectively, were obtained.

Sequence assembly and assembly validation. Basecalling, quality control and elimination of vector DNA sequences of the shotgun-sequences were performed by using the software package BioMake (Bielefeld University) as previously described[34]. Sequence assembly was performed by using the PHRAP assembly tool (http://www.phrap.org). The CONSED/AUTOFINISH software package[35,36], supplemented by the in-house tool BACCardI[37], was used for the finishing of the genome sequence.

For gap closure, a BAC library with inserts of ~90 kb in pBeloBAC11 was constructed and BAC contigs were assembled[12]. Remaining gaps of the whole genome shotgun assembly were closed by sequencing on shotgun and BAC clones carried out by IIT GmbH on LI-COR 4200L and ABI 377 sequencing machines. So that we would obtain a high quality genome sequence, all regions of the consensus sequence were polished to at least phred40 quality by primer walking. Collectively, 1,374 sequencing reads were added to the shotgun assembly for finishing and polishing of the genomic sequence. Repetitive elements, that is, rRNA operons, were sequenced completely by primer walking on BAC clones. For assembly validation, a fosmid library with inserts of ~35–38 kb was constructed by IIT GmbH using the EpiFOS Fosmid Library Production Kit (Epicentre). End-sequencing of 672 fosmids was carried out on ABI 377 and MegaBACE 1000 (Amersham Biosciences) sequencing machines by IIT GmbH and on ABI 3730XL DNA analyzers by the sequencing group of the Max Planck Institut für Molekulare Genetik. For assembly validation, fosmid end sequences were mapped onto the genome sequence by employing the BACCardI tool.

Genome analysis and annotation. In a first step automatic gene prediction and annotation were performed using the genome annotation system GenDB 2.0 (ref. 38) as previously described[14]. In a second step the coding sequences (CDS) prediction was validated: a position weight matrix (PWM) was generated to score all translation starts, and visualization of CDS was done using GeneQuest (DNASTAR Inc.). Reinspection of starts was coupled to recomputing of homology (BlastP) and assessment of function. In this way 4.3% more ORFs were detected, and 15.5% of the start sites were changed in comparison to the prediction obtained by a combined GLIMMER and CRITICA approach[39]. Intergenic regions were checked again for CDS missed probably by the automatic annotation using the BLAST programs[40]. During manual annotation, the following criteria were applied: (i) for hypothetical proteins, amino acid identities to other proteins were less than 30% over the entire length of the protein; (ii) for conserved hypothetical proteins, amino acid identities to proteins of unknown function were more than 30%; (iii) for proteins with putative or probable functions, sequence identities to named gene products were >20% or 40%, respectively.

Genomic comparison. For comparative analyses, the annotated genome sequence of *Azoarcus* strain EbN1 (acc. nos. CR555306, CR555307, CR555308) was imported into GenDB. Comparisons of chromosomal sequences were carried out with GenDB[38].

Detection of regions with atypical G+C content. For detection of anomalous gene clusters or putative pathogenicity islands in bacterial genomes, the Pai-Ida program 1.1 (http://compbio.sibsnet.org/projects/pai-ida/) based on an iterative discriminant analysis[41] was used.

Database submission. The nucleotide sequence of the *Azoarcus* sp. strain BH72 chromosome was submitted to EBI under accession number AM406670–*Azoarcus* sp. BH72.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The authors grateful thank all the people involved in this project. This work was supported by grants of the German Federal Ministry of Education and Research (BMBF) (031U113D, 031U213D and 0313105) in the frame of the GenoMik network "Genome Research on Bacteria Relevant for Agriculture, Environment and Biotechnology" and the Fonds der Chemischen Industrie (to R.T.).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Barac, T. *et al.* Engineered endophytic bacteria improve phytoremediation of water-soluble, volatile, organic pollutants. *Nat. Biotechnol.* **22**, 583–588 (2004).
2. Hurek, T., Handley, L., Reinhold-Hurek, B. & Piché, Y. *Azoarcus* grass endophytes contribute fixed nitrogen to the plant in an unculturable state. *Mol. Plant Microbe Interact.* **15**, 233–242 (2002).
3. Sevilla, M., Burris, R.H., Gunapala, N. & Kennedy, C. Comparison of benefit to sugarcane plant growth and 15N2 incorporation following inoculation of sterile plants with Acetobacter diazotrophicus wild-type and *nif*- mutant strains. *Mol. Plant-Microbe Interact.* **14**, 358–366 (2001).
4. Hurek, T. & Reinhold-Hurek, B. *Azoarcus* sp. strain BH72 as a model for nitrogen-fixing grass endophytes. *J. Biotechnol.* **106**, 169–178 (2003).
5. Reinhold, B., Hurek, T., Niemann, E-G. & Fendrik, I. Close association of *Azospirillum* and diazotrophic rods with different root zones of Kallar grass. *Appl. Environ. Microbiol.* **52**, 520–526 (1986).
6. Hurek, T., Reinhold-Hurek, B., Van Montagu, M. & Kellenberger, E. Root colonization and systemic spreading of *Azoarcus* sp. strain BH72 in grasses. *J. Bacteriol.* **176**, 1913–1923 (1994).
7. Miché, L., Battistoni, F., Gemmer, S. & Belghazi, M. Reinhold-Hurek, B. Differential colonization and up-regulation of jasmonate-inducible defence proteins in roots of *Oryza sativa* cultivars upon interaction with the endophyte *Azoarcus* sp. *Mol. Plant-Microbe Interact.* **19**, 502–511 (2006).

8. Egener, T., Hurek, T. & Reinhold-Hurek, B. Endophytic expression of *nif* genes of *Azoarcus* sp. strain BH72 in rice roots. *Mol. Plant-Microbe Interact.* **12**, 813–819 (1999).
9. Reinhold-Hurek, B., Tan, Z. & Hurek, T. in Bergey's Manual of Systematic Bacteriology. Vol. 2. (ed. G.M. Garrity) 890–901, (Springer Verlag, New York, 2005).
10. Rabus, R. *et al.* The genome sequence of an anaerobic aromatic-degrading denitrifying bacterium, strain EbN1. *Arch. Microbiol.* **183**, 27–36 (2005).
11. Reinhold-Hurek, B. & Hurek, T. Life in grasses: diazotrophic endophytes. *Trends Microbiol.* **6**, 139–144 (1998).
12. Battistoni, F. *et al.* Physical map of the *Azoarcus* sp. strain BH72 genome based on a bacterial artificial chromosome (BAC) library as a platform for genome sequencing and functional analysis. *FEMS Microbiol. Lett.* **249**, 233–240 (2005).
13. James, E.K. & Olivares, F.L. Infection and colonization of sugar cane and other graminaceous plants by endophytic diazotrophs. *Crit. Rev. Plant Sci.* **17**, 77–119 (1998).
14. Thieme, F. *et al.* Insights into genome plasticity and pathogenicity of the plant pathogenic bacterium *Xanthomonas campestris pv. vesicatoria* revealed by the complete genome sequence. *J. Bacteriol.* **187**, 7254–7266 (2005).
15. Reinhold-Hurek, B. *et al. Azoarcus* gen. nov., nitrogen-fixing proteobacteria associated with roots of Kallar grass (*Leptochloa fusca* (L.) Kunth) and description of two species *Azoarcus indigens* sp. nov. and *Azoarcus communis* sp. nov. *Int. J. Syst. Bacteriol.* **43**, 574–584 (1993).
16. Reinhold-Hurek, B., Hurek, T., Claeyssens, M. & Van Montagu, M. Cloning, expression in *Escherichia coli*, and characterization of cellulolytic enzymes of *Azoarcus* sp., a root-invading diazotroph. *J. Bacteriol.* **175**, 7056–7065 (1993).
17. Yurgel, S.N. & Kahn, M.L. Dicarboxylate transport by rhizobia. *FEMS Microbiol. Rev.* **28**, 489–501 (2004).
18. Hurek, T., Reinhold, B., Fendrik, I. & Niemann, E.G. Root-zone-specific oxygen tolerance of *Azospirillum* spp. and diazotrophic rods closely associated with Kallar grass. *Appl. Environ. Microbiol.* **53**, 163–169 (1987).
19. Koebnik, R. TonB-dependent trans-envelope signalling: the exception or the rule? *Trends Microbiol* **13**, 343–347 (2005).
20. Paulsen, I.T. *et al.* Complete genome sequence of the plant commensal *Pseudomonas fluorescens* Pf-5. *Nat. Biotechnol.* **23**, 873–878 (2005).
21. Crosa, J.H. & Walsh, C.T. Genetics and assembly line enzymology of siderophore biosynthesis in bacteria. *Microbiol. Mol. Biol. Rev.* **66**, 223–249 (2002).
22. Crowley, D.E., Wang, Y.C., Reid, C.P.P. & Szaniszlo, P.J. Mechanism of iron acquisition from siderophores by microorganism and plants. *Plant Soil* **130**, 179–198 (1991).
23. Dörr, J., Hurek, T. & Reinhold-Hurek, B. Type IV pili are involved in plant-microbe and fungus-microbe interactions. *Mol. Microbiol.* **30**, 7–17 (1998).
24. Katzen, F. *et al. Xanthomonas campestris* pv. campestris *gum* mutants: effects on xanthan biosynthesis and plant virulence. *J. Bacteriol.* **180**, 1607–1617 (1998).
25. Filloux, A. The underlying mechanisms of type II protein secretion. *Biochim. Biophys. Acta* **1694**, 163–179 (2004).
26. He, S.Y., Nomura, K. & Whittam, T.S. Type III protein secretion mechanism in mammalian and plant pathogens. *Biochim. Biophys. Acta* **1694**, 181–206 (2004).
27. Krause, A., Doerfel, A. & Göttfert, M. Mutational and transcriptional analysis of the type III secretion system of *Bradyrhizobium japonicum*. *Mol. Plant-Microbe Interact.* **15**, 1228–1235 (2002).
28. Christie, P.J. Type IV secretion: the *Agrobacterium* VirB/D4 and related conjugation systems. *Biochim. Biophys. Acta* **1694**, 219–234 (2004).
29. Hubber, A., Vergunst, A.C., Sullivan, J.T., Hooykaas, P.J. & Ronson, C.W. Symbiotic phenotypes and translocated effector proteins of the *Mesorhizobium loti* strain R7A VirB/D4 type IV secretion system. *Mol. Microbiol.* **54**, 561–574 (2004).
30. Camilli, A. & Bassler, B.L. Bacterial small-molecule signaling pathways. *Science* **311**, 1113–1116 (2006).
31. Reinhold-Hurek, B., Maes, T., Gemmer, S., Van Montagu, M. & Hurek, T. An endoglucanase is involved in infection of rice roots by the not cellulose-metabolizing endophyte *Azoarcus* sp. BH72. *Mol. Plant-Microbe Interact.* **19**, 181–188 (2006).
32. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
33. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
34. Kaiser, O. *et al.* Whole genome shotgun sequencing guided by bioinformatics pipelines–an optimized approach for an established technique. *J. Biotechnol.* **106**, 121–133 (2003).
35. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).
36. Gordon, D., Desmarais, C. & Green, P. Automated finishing with autofinish. *Genome Res.* **11**, 614–625 (2001).
37. Bartels, D. *et al.* BACCardI–a tool for the validation of genomic assemblies, assisting genome finishing and intergenome comparison. *Bioinformatics* **21**, 853–859 (2005).
38. Meyer, F. *et al.* GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.* **31**, 2187–2195 (2003).
39. McHardy, A.C., Goesmann, A., Pühler, A. & Meyer, F. Development of joint application strategies for two microbial gene finders. *Bioinformatics* **20**, 1622–1631 (2004).
40. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
41. Tu, Q. & Ding, D. Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. *FEMS Microbiol. Lett.* **221**, 269–275 (2003).
42. Reinhold-Hurek, B. & Shub, D.A. Self-splicing introns in tRNA genes of widely divergent bacteria. *Nature* **357**, 173–176 (1992).