nature biotechnology

Genome sequence of the chlorinated compound–respiring bacterium *Dehalococcoides* species strain CBDB1

Michael Kube[1], Alfred Beck[1], Stephen H Zinder[2], Heiner Kuhl[1], Richard Reinhardt[1] & Lorenz Adrian[3]

Dehalococcoides **species are strictly anaerobic bacteria, which catabolize many of the most toxic and persistent chlorinated aromatics and aliphatics by reductive dechlorination and are used for *in situ* bioremediation of contaminated sites. Our sequencing of the complete 1,395,502 base pair genome of *Dehalococcoides* strain CBDB1 has revealed the presence of 32 reductive-dehalogenase-homologous (*rdh*) genes, possibly conferring on the bacteria an immense dehalogenating potential. Most *rdh* genes were associated with genes encoding transcription regulators such as two-component regulatory systems or transcription regulators of the MarR-type. Four new paralog groups of *rdh*-associated genes without known function were detected. Comparison with the recently sequenced genome of *Dehalococcoides ethenogenes* strain 195 reveals a high degree of gene context conservation (synteny) but exceptionally high plasticity in all regions containing *rdh* genes, suggesting that these regions are under intense evolutionary pressure.**

In the past century, human activities have released large amounts of chlorinated organic compounds into the environment, and these compounds are among the most pervasive groundwater pollutants. Chloroorganics with fewer chlorines can usually be biodegraded by aerobic microorganisms, whereas more highly chlorinated ones can be reductively dechlorinated by organisms using them as electron acceptors for anaerobic respiration. Several microbial groups carry out respiratory reductive dechlorination[1], but one genus, *Dehalococcoides*, seems particularly adapted to a unique niche, and the bacteria are only known to use chloroorganics as electron acceptors and hydrogen as an electron donor for growth. Several studies have now shown successful application of *Dehalococcoides*-containing cultures for *in situ* remediation of contaminated sites[2,3]. The recently described genome sequence of *D. ethenogenes*, strain 195 (ref. 4), an organism that reductively dechlorinates the solvents tetrachloroethene and trichloroethene to vinyl chloride and ethene[5], reflects this specialization: although only 1.47-megabases (Mb) long, the genome possesses 17 *rdhAB* pairs potentially encoding reductive dehalogenases and genes predicted to encode five different hydrogenase complexes. We describe here the genome sequence of *Dehalococcoides* sp. strain CBDB1, a bacterium that has in many ways a different dechlorination spectrum from strain 195. For instance, strain CBDB1 (refs. 6,7) dechlorinated 1,2,3-trichlorobenzene, 1,2,4-trichlorobenzene, 2,3-dichloro-*p*-dibenzodioxin and the ‘Seveso dioxin’ 2,3,7,8-tetrachloro-*p*-dibenzodioxin, but strain 195 (ref. 8) did not. In contrast to strain 195, strain CBDB1 did not dechlorinate tetrachloroethene past dichloroethene[6]. By sequencing a second *Dehalococcoides* genome, we can now show that strain CBDB1 has an even greater potential as a reductive dechlorinator than strain 195 and, in addition, the comparison of the two genomes with each other provides insights into the evolution of reductive dechlorination in this unusual microbial group.

RESULTS

General description of the genome of strain CBDB1

Strain CBDB1 contains a single circular chromosome with 1,395,502 base pairs, encoding 1,458 predicted protein coding sequences (**Table 1**). Of the 1,051 coding sequences that were annotated to known functions, 32 *rdhAB* pairs and the two-component regulatory systems form the largest paralog groups. Twenty-eight histidine kinase and 34 response-regulator genes as well as 16 regulators of the MarR-family were annotated in the genome. Fifteen of the two-component regulatory systems and 13 of the MarR-type regulators (DNA-binding transcriptional repressors or activators that often regulate degradation pathways) are associated with *rdhAB* genes (see **Supplementary Fig. 1** online). Four more paralog groups containing between five and eight proteins of unknown function each are also mainly associated with *rdhAB* genes and were therefore denominated *rdhF*, *rdhG*, *rdhH* and *rdhI*. Large paralog groups not associated with *rdhAB* genes include ABC transporters (25 ATP-binding subunits, 17 permease subunits, 5 substrate-binding subunits), SAM-dependent methyltransferases (seven coding sequences), aminotransferases (five coding sequences), major facilitator family transporters (five coding sequences) and degV proteins (seven coding sequences).

[1]Max-Planck-Institut für Molekulare Genetik, Ihnestr. 63-73, 14195 Berlin-Dahlem, Germany. [2]Dept. of Microbiology, Cornell University, 272 Wing Hall, Ithaca, New York 14853, USA. [3]FG Technische Biochemie, Technische Universität Berlin, Seestr. 13, 13353 Berlin, Germany. Correspondence should be addressed to L.A. (lorenz.adrian@tu-berlin.de).

Received 23 May; accepted 12 July; published online 21 August 2005; doi:10.1038/nbt1131

Table 1 Genome overview of the two *Dehalococcoides* strains

	Dehalococcoides strain	
	CBDB1	195^{a}
Size (base pairs, bp)	1,395,502	1,469,720
G+C content (%)	47.0	48.9
Stable RNAs		
rRNAs	3	3
tRNAs	47	46
Protein-coding sequences (CDS)	1,458	1,591
Coding density (%)	89.7	90.5
Average CDS size (bp)	859	825
Assigned function	1,051	907
Conserved unknown	164	382
Predicted novel	243	302
Reductive-dehalogenase-homologous genes (*rdhAB*)	32	17

^{a}Data from ref. 4.

Comparison with *D. ethenogenes* strain 195

Of the 1,458 coding sequences in strain CBDB1, 1,217 (83.5%) have orthologous genes in strain 195. The median amino acid sequence identity between orthologs is 91.1% and the median amino acid sequence similarity is 95.7%. Also the nucleotide sequence identities of rRNAs (16S rRNA, 98.8%; 5S rRNA, 99.2%; 23S rRNA, 98.9%) and common marker genes such as *gyrA* (89.7%), translation elongation factor Tu (92.6%) and *groEL* (93.7%) are high. Genes of the tryptophan operon, genes encoding ribosomal proteins and several genes without known products are among the most highly conserved. Because of the high number of orthologous genes, most enzymes previously described for strain 195 (ref. 4) are also present in strain CBDB1. These include enzymes for acetate uptake via acetyl-CoA, pyruvate and phosphoenolpyruvate; enzymes for amino acid syntheses, a membrane-bound hydrogen-uptake [NiFe]-hydrogenase (*hup*), two membrane-bound (*ech* and *hyc*) and one soluble [NiFe] hydrogenase complex (*vhu*); a membrane-bound multisubunit complex with close similarity to bacterial NADH-ubiquinone-oxidoreductase (*nuo*–complex I); a formate dehydrogenase; an enzyme complex that contains molybdopterin-guanidine dinucleotide whose electron donor and acceptor cannot be predicted; vitamin B_{12} uptake and salvage genes; various transporters; and genes involved in DNA replication, transcription and translation. Neither of the two strains encodes cytochromes, flagella genes or the genes necessary to synthesize a peptidoglycan cell wall. One copy of each rRNA gene is present in strain CBDB1 and strain 195. In both strains the 16S rRNA gene is spatially separated from 5S and 23S rRNA genes.

The genomes of strains CBDB1 and 195 are organized similarly based on a comparison of nucleotide sequences (**Fig. 1**) or amino acid sequences of translated coding sequences (see **Supplementary Fig. 2** online). Major differences between the two genomes are concentrated in distinct regions (**Fig. 1c** and **Supplementary Tables 1** and **2** online). In strain CBDB1, three regions contain large putative integrated elements (region 1: bp 58,790–108,048; region 3: bp 570,912–579,362; region 7: bp 1,166,180–1,213,880), which are characterized by exceptional di- and trinucleotide composition[4,9] (see **Supplementary Fig. 3** online), deviation from the average GC content found across the entire genome (**Supplementary Fig. 3** online) and/or the occurrence of genes often associated with mobile elements. Region 1 is composed of two parts. The first encodes four *rdhAB* pairs; the second contains several genes often associated with mobile elements such as a restriction modification system[10], a helicase or an ATP-dependent exoDNAse. Only the second part of region 1 shows irregular Karlin-signature difference values suggesting that the *rdh* genes originate from *Dehalococcoides* species or that the *rdh* genes were integrated much earlier. Directly preceding region 1 is a putative recombinase fragment (cbdbB2) that was probably involved in the recombination event. Region 3 contains a site-specific recombinase as the only predictable protein together with 18 hypothetical peptides. Region 7 is framed on one side by a series of genes often associated with mobile elements including a site-specific integrase, DNA invertase/resolvase and transposase and on the other side by repetitive elements called clustered regularly interspaced short palindromic repeats (CRISPR) together with CRISPR-associated genes that are thought to be acquired efficiently by horizontal gene transfer[11]. Both flanking parts of region 7, but not the enclosed region containing four *rdhAB* pairs, show increased Karlin-signature-difference values (**Fig. 1b**). The presence of eight *rdhAB* pairs in presumed mobile elements corroborates

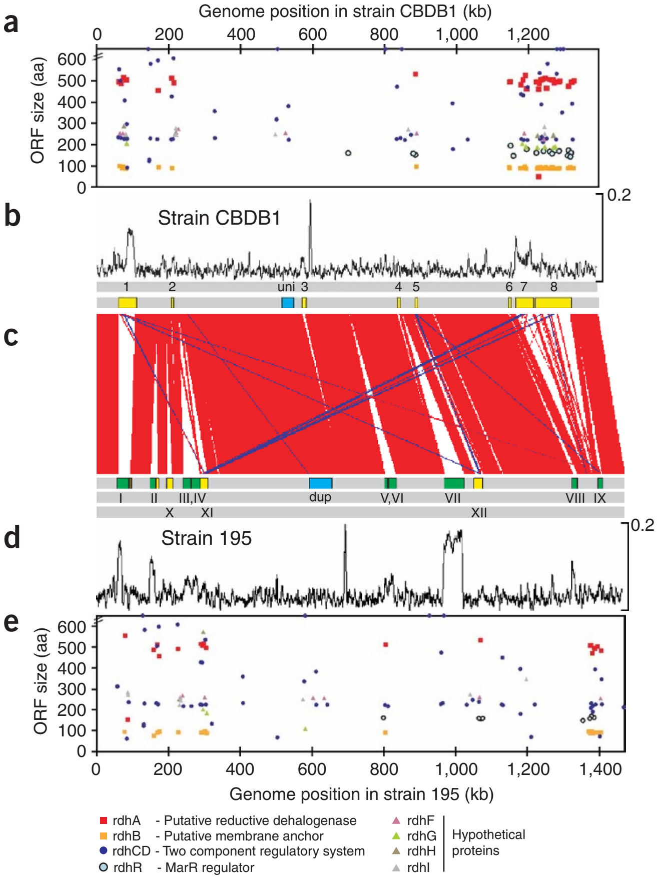

Figure 1 Comparison of the genomes of strain CBDB1 and strain 195. (**a**) Location of *rdhAB* pairs and *rdhAB*-associated genes in strain CBDB1. Coding sequence positions are plotted versus the predicted amino acid length to improve resolution and to illustrate the similar size of proteins of the same paralog group. (**b**) Karlin signature difference in the genome of strain CBDB1. (**c**) Regions of significant sequence identity between the nucleotide sequences of strain CBDB1 and strain 195 (score value ≥200) connected by red and blue (inversion) lines. Boxes indicate regions that are organized differently in the two genomes. Colors: green (roman numbers I–IX), predicted putative integrated regions[4]; yellow, further regions only present in one of the two genomes; blue, 31-kb region duplicated in strain 195 (dup), unique in strain CBDB1 (uni). (**d**) Karlin signature difference in the genome of strain 195. (**e**) Location of *rdhAB* pairs and *rdhAB*-associated genes in strain 195.

hypotheses, developed from data on other organisms, that suggest mobile elements played an important role in the evolution of *rdh* genes[1,12,13]. However, other *rdh* genes in the genome of strain CBDB1 that are absent from the genome of strain 195 are located in regions without clear characteristics of mobile elements. Region 6 for example is present only in strain CBDB1 and encodes two *rdhAB* pairs, and one MarR-type regulator gene (*rdhR*). Sixteen *rdhAB* pairs are accumulated in region 8. One transposase fragment (cbdbB42) is present in the center of this region, but its involvement in recombination events is unclear. Several *rdh* genes in this region seem to originate from recent duplication events as they show noticeable high sequence similarity to each other (cbdbA1550–cbdbA1570; cbdbA1546–cbdbA1575).

Nine regions predicted previously as integrated elements[4] are present in strain 195 but not in strain CBDB1 (**Fig. 1c**, green boxes, regions I–IX). Three of these regions are copies of a 22-kb integrated element that contains phage-like genes (regions III, IV and VI). Region VII contains a 55-kb prophage. As a result of extensive integration of mobile elements, the genome of strain 195 is 5% (74 kb) larger than that of strain CBDB1. A 31-kb duplicated region in the genome of strain 195 encoding many genes connected with CO_2-fixation, corrinoid-factor uptake and corrinoid-factor salvage is present in strain CBDB1 in a single copy. Apart from the predicted integrated elements, several more coding regions were identified in strain 195 that are not present in strain CBDB1 (see **Supplementary Table 2** online). These regions contain the *tceA*-gene DET0079 responsible for dechlorination of trichloroethene via vinyl chloride to ethene in strain 195 (ref. 14), a putatively nonfunctional *rdhA*-gene DET0162 and four additional *rdhA* genes (DET0173, DET0876, DET1528, DET1559) together with their *rdhB* and regulatory genes. All nitrogen fixation–related genes in strain 195 (DET1148–1158), genes for a three-subunit molybdenum ABC transporter (DET1159–1161), a Fec-type ABC transporter (DET1174–1176), some genes involved in carbohydrate metabolism/glycosylation (DET0202–0211, DET0215) and a gene similar to the β-subunit of cofactor F_{420}-reducing hydrogenase (DET0214) are missing in strain CBDB1. Also many genes related to DNA recombination such as site-specific integrase, recombinase, transposase or resolvase, a gene encoding an acetyltransferase (DET0144) and many genes encoding unknown proteins are present only in strain 195. The absence of the *tceA*-gene can explain the inability of strain CBDB1 to dechlorinate trichloroethene to vinyl chloride and ethene. The physiological impact of the absence of the other genes is unclear as strain CBDB1 grows in completely synthetic medium containing several vitamins and fixed nitrogen. Although some nitrogen fixation–related genes are present in strain 195, no diazotrophic growth is described for this strain, which is presently grown in a complex medium.

rdh genes in CBDB1

All 32 *rdh* loci in the genome of strain CBDB1 are composed of two genes, *rdhA* and *rdhB*. *rdhA* is thought to encode the active subunit of the enzyme, *rdhB* encodes a small hydrophobic protein, which possibly acts as a membrane anchor for RdhA. All these 32 *rdhAB* pairs share numerous characteristics with the biochemically characterized trichloroethene reductive dehalogenase of strain 195 (*tceA*)[14], the vinyl chloride reductive dehalogenase of *Dehalococcoides* sp. strain BAV1[15], a vinyl chloride reductive dehalogenase isolated from *Dehalococcoides* sp. strain VS in highly enriched culture[16] and *rdh* loci in other bacteria growing by respiratory dehalogenation. Similarities include the lengths of the encoded proteins RdhA (455–532 amino acids) and RdhB (89–100 amino acids), the twin arginine leader sequence on the N terminus of the *rdhA* product, two iron-sulfur-cluster-forming motifs on the C terminus of the *rdhA* product and three transmembrane helices in the encoded amino acid sequence of *rdhB*. In addition to 32 *rdhAB* pairs, one 50 amino acid, N-terminal *rdhA* fragment is present in the genome of strain CBDB1 (cbdbA1540). With one exception (cbdbA1092) all *rdhAB* pairs are located between 63 and 250 kb before or after the origin of replication (**Fig. 1a** and **Supplementary Figs. 1** and **4** online). Seven *rdhAB* pairs are located in the first quarter of the genome (bp 61,000–216,000), 24 *rdhAB* pairs in the last quarter of the genome (bp 1,146,000–1,327,000). The highest concentration of *rdhAB* pairs is in region 8 with 16 *rdhAB* pairs within 97 kb, averaging 1 *rdhAB* pair every 6 kb. In contrast to other genes in the genome, *rdh* genes show strong genome orientation. Of the 32 *rdhAB* genes, 28 are located on the leading strand. The associated regulatory genes *rdhC*, *rdhD* and *rdhR* are predominantly located on the lagging strand. Analyses of *rdhA* (see **Supplementary Fig. 5** online) and *rdhB* genes (see **Supplementary Fig. 6** online) show that phylogenetically closely related *rdh* genes are usually not located in close proximity on the genome, indicating that recombination events mostly occurred between distal regions on the genome.

With few exceptions the topologies of *rdhA* and *rdhB* trees are identical (see **Supplementary Figs. 5** and **6** online) corroborating that the two genes are linked in their evolution and that each *rdhB* product is specific for its associated *rdhA* gene. One important exception is the *rdhA*-gene cbdbA1508 that appears to be a chimera between a 347-bp N-terminal fragment unique in strain CBDB1 and a 116-bp C-terminal fragment that is identical with the C-terminal part of cbdbA1588. Also the associated *rdhB* genes of the *rdhA* genes cbdbA1508 and cbdbA1588 are identical, indicating a recent recombination. RdhB proteins in strain CBDB1 are exceptionally high in tryptophan content showing an average of 5.7%. In contrast, the average tryptophan content of other conserved proteins in strain CBDB1 is 0.94%, and the average tryptophan content of proteins with predicted transmembrane helices excluding *rdhB* gene products and hypothetical proteins is 1.6%. The positions of tryptophan residues at both ends of predicted transmembrane helices in *rdhB* products are highly conserved, suggesting a crucial role of tryptophan residues, possibly in the interaction with the RdhA subunit or with an electron carrier in the membrane. This elevated tryptophan content can also be found in RdhB proteins from other *Dehalococcoides* species and to a lesser extent in most RdhB proteins from other organisms growing by respiratory dehalogenation.

Twelve of the 32 *rdhAB* pairs in strain CBDB1 have orthologous genes in strain 195 (see **Supplementary Fig. 5** online). Amino acid sequence identities of orthologs are between 86.4% and 95.4%. However, position and orientation is conserved in only six of the orthologs. The other six conserved *rdhAB* pairs have an inverted orientation relative to each other, different positions in the genomes or both. Therefore, only 6 out of the 32 *rdhAB* pairs in strain CBDB1 are located in conserved regions whereas the positions and orientations of more than 90% of the genes not associated with *rdh* loci are conserved. This illustrates how strong genomic variation in the two strains is focused on *rdh* loci. A remarkable example for this high plasticity of *rdh* loci is region XI in strain 195. All four *rdhAB* pairs in this region have orthologs in strain CBDB1; however, the four orthologs in strain CBDB1 are scattered through the genome (regions 1, 2, 7 and 8), three of them with an inverted orientation.

rdh-associated genes

Nearly all *rdhAB* pairs have two-component regulatory systems (*rdhCD*) or MarR-type transcriptional regulators (*rdhR*) nearby, indicating that expression of *rdhAB* genes is tightly regulated. As in strain 195, all *rdhAB*-associated two-component systems have cytoplasmic histidine kinase subunits indicating that a cytoplasmic signal triggers expression. For unknown reasons, MarR-type regulators are present only in the second half of the genome in both organisms. Phylogenetic analysis of *Dehalococcoides* MarR-type regulators reveals that ten MarR-type regulators of strain CBDB1 together with three MarR-type regulators of strain 195 form a yet-unknown cluster of MarR-type regulators (see **Supplementary Fig. 7** online). All MarR-type regulators of this cluster are closely associated with *rdhAB* genes, being mostly located directly upstream of the associated *rdhAB* pair, and with one exception (cbdbA1583) being oriented in the opposite direction to the associated *rdhAB* pair. In a typical *rdh* locus with associated MarR-type regulator, the size of the intergenic region between *rdhA* and *rdhR* is between 200 and 320 bp. This gene organization is identical with many MarR-regulated loci described in the literature. Several studies have described aromatic compounds as effectors of MarR-type regulators[17]. Therefore, the ten anticipated MarR-regulated *rdhAB* genes in strain CBDB1 are particularly interesting candidates for reductive dehalogenases that dechlorinate aromatic compounds. Other MarR-type regulators in the genomes of strains CBDB1 and 195 cluster with MarR-type regulators of other bacteria (see **Supplementary Fig. 7** online) and are mostly not closely associated with *rdhAB* genes in *Dehalococcoides* strains. Apart from regulatory functions, MarR-type regulators seem also to have played a crucial role in genomic rearrangement events in *Dehalococcoides* species, as they are very often located at the edges of integrated or excised regions (e.g., cbdbA1085, which is duplicated in strain 195; cbdbA1091; cbdbA1456).

Four *rdhAB*-associated paralog groups with unknown functions were discovered in the genome of strain CBDB1 (**Fig. 1e**): *rdhF*-encoded proteins (eight coding sequences) with around 250 amino acids are among the most highly conserved proteins in the genome. Five members of this group are directly associated with *rdhA* genes, three others (cbdbA249, cbdbA258 and cbdbA637) are associated with cyanocobalamin biosynthesis genes and therefore functionally connected with *rdh* genes. *rdhG* genes (seven coding sequences) encode proteins of about 190 amino acids. One of these proteins matched to a Zinc-carboxypeptidase pattern from the PROSITE database of protein families and domains (http://ca.expasy.org/prosite/) suggesting that *rdhG*-encoded proteins could be dehalogenase processing proteases. *rdhH* genes (five coding sequences) code for proteins with around 250 amino acids and are all closely associated with *rdhAB* genes. All eight encoded proteins of the paralog group *rdhI* are members of the radical S-adenosyl methionine (SAM) Pfam-superfamily and contain about 250–280 amino acids. These proteins may serve as corrinoid cofactor–modifying enzymes.

DISCUSSION

Bacteria of the genus *Dehalococcoides* are extraordinarily specialized for a unique physiological niche, detoxifying compounds that are otherwise persistent for decades. This specialization can also be observed in the genome, which contains less than 1.4 Mb and is among the smallest for free-living prokaryotes; however, a remarkable 32 *rdhAB* genes are encoded. The presence of these 32 *rdhAB* genes implies that strain CBDB1 can reductively dehalogenate an enormous variety of halogenated organic compounds, or that there are non-halogenated targets for some of them.

Detailed comparisons of prokaryotic genomes have been done previously including those between different *Escherichia coli* strains[18] and between *Thermococcus kodakaraensis* and three *Pyrococcus* species genomes[19]. Similar to those studies, we found, by comparing the genomes of strains CBDB1 and 195, long regions of colinearity of gene order (synteny) among housekeeping genes, punctuated by islands that represent mobile genetic elements or sets of genes encoding functions present in one organism but not the other. Of particular interest are the regions encoding *rdhAB* genes. In both organisms, nearly all the *rdhAB*-pairs are located close to the origin of replication, are transcribed in the direction of DNA synthesis and are associated with genes encoding either two-component or MarR-type regulators, all indications that they encode important, highly regulated functions in these organisms. Several observations in our study, including those that reveal that most major differences between the two genomes are within regions that contain *rdhAB* genes, that many of the *rdhAB* pairs present in both organisms have different locations in the genome and/or different orientations, that genes encoding integrases and recombinases are often present in regions with *rdhAB* genes as well as the apparently recent recombination event between the *rdhA* genes cbdbA1508 and cbdbA1588, show that *rdh* genes in *Dehalococcoides* species are under intense evolutionary pressure.

Besides *rdh* genes, nearly all other genes predicted to participate in electron transport were present in both organisms including those predicted to encode five different hydrogenase multisubunit complexes. Strain 195 has a complex nutrition, requiring acetate, vitamins and extracts from mixed cultures[5] and its genome annotation predicted an incomplete pathway to methionine, lesions in the tricarboxylic acid cycle leading to glutamate, as well as incomplete pathways for the synthesis of vitamin B_{12}, biotin and quinones[4]. Strain CBDB1 is able to grow in a defined medium containing only acetate and vitamins, yet these same lesions are still predicted to be present. The higher amount of bacteriophage-related integrated sequences observed in strain 195 might reflect the higher burden of phages in the habitat from which strain 195 was isolated, sewage sludge[5], whereas strain CBDB1 was isolated from anoxic river sediment[6].

In future work, we will investigate the biochemical activities of the different *rdhAB* genes and the functions of the newly discovered *rdhAB*-associated paralog groups. Knowledge of the encoded gene functions can be used for a rational prediction of dechlorinating potential in *Dehalococcoides* populations at contaminated sites and might help in the further promotion of *in situ* application of *Dehalococcoides*-enriched cultures for bioremediation.

METHODS

Cultivation and DNA extraction. *Dehalococcoides* sp. strain CBDB1 was cultivated in synthetic mineral medium under strictly anaerobic conditions with Ti(III) citrate as the reducing agent[6]. Hydrogen was used as the electron donor (7.5 mM nominal concentration), 5 mM acetate as carbon source, and 15 μM 1,2,3-trichlorobenzene and 15 μM 1,2,4-trichlorobenzene as the electron acceptors. After two weeks of static incubation in the dark, 10 mM of 1,2,3-trichlorobenzene was added as a 1-M solution in hexadecane. After two more weeks of incubation, cultures had reached cell numbers of about 5×10^7 ml^{-1} and cells were harvested by centrifugation (15 min, 10,000*g*). DNA was isolated from 30-ml cultures using the Qiagen Mini Kit (Qiagen) according to the instructions of the manufacturer.

Sequencing. Two whole-genome shotgun libraries with average insert sizes of 1.5 and 3.5 kb were generated. DNA was sonicated, and fragment ends were polished with T4 and Klenow polymerase (New England Biolabs). Size-selected fragments were ligated into pUC19 vector (Fermentas) and transferred into *E. coli* strain DH10B (Invitrogen) by electroporation. Templates for sequencing were obtained by insert amplification via PCR or by plasmid isolation. DNA was sequenced using Big Dye chemistry, and ABI3730XL capillary sequencer systems (ABI) up to a 12-fold sequencing coverage. Gaps were eliminated and regions of weak quality were improved by resequencing of selected plasmids, primer walking and sequencing of long-range PCR products. The quality of raw sequence data was determined with PHRED[20,21]. Sequences were assembled with Phrap (http://www.genome.bnl.gov/Software/UW/). Consed (Version 14.00[22]) was used for final editing of the sequence. Sequence data contain less than one error within 100,000 bases.

Genome analysis. Glimmer 2.0 (ref. 23) was used for the prediction of open reading frames (ORFs) on the finished chromosome. ORF prediction was manually refined using ARTEMIS[24]. Most overlapping ORFs without functional assignment or BLAST hits were removed. To identify ORFs that were not noticed during initial prediction, the translated nucleotide sequence of the genome was screened against the UniProt database. Similarity searches for annotation were carried out with BLASTP[25] using the translated amino acid sequences of predicted ORFs as queries. Functional assignments were performed with the INTERPRO system[26] using the modules PROSITE, Pfam, PRINTS, ProDom, SMART, TIGRFAMs and SIGNALP[27,28]. In addition, predicted ORFs were screened against the database of Clusters of Orthologous Groups of proteins (COGs[27]). These methods were implemented in the web-based platform HTGA (High-Throughput Genome Annotation[29]) and used for annotation. The annotated genome of *D. ethenogenes* strain 195 (NCBI accession number CP000027) was used for comparison and refinement of the annotation of strain CBDB1. Genome comparison was done with the Artemis Comparison Tool (ACT, http://www.sanger.ac.uk/Software/ACT/). Karlin signature difference, the difference of the relative abundance of dinucleotides between a sliding window and the whole sequence[9], was calculated with ACT using a window size of 5,000 bp and a step size of 2,500 bp. Trinucleotide composition was calculated by SWAAP using a window size of 5,000 bp and a step size 2,500 bp (http://www.bacteriamuseum.org/SWAAP/SwaapPage.htm).

Coding sequences in strains CBDB1 and 195 were automatically assigned as orthologs if the following five criteria applied: (i) the best hits of BLASTP analyses were to each other, (ii) the amino acid sequence identity of the two sequences was $>50\%$, (iii) both BLASTP e-values were $>10^{-10}$, (iv) the ratio of the coding sequence lengths was between 0.9 and 1.1, and (v) the ratio of the alignment length to the coding sequence length in strain CBDB1 was >0.9. If only some of these five criteria matched, global alignments, genome context and results of an InterProScan (http://www.ebi.ac.uk/InterProScan/) were taken into consideration. Sequence identities and similarities between orthologous gene pairs were calculated from global alignments using Stretcher from the EMBOSS package with the BLOSUM62 matrix[30,31]. Paralog groups were formed from gene groups that had BLASTP hits to each other with e-values better than (that is, less than) 10^{-6} and an alignment length of at least 60% of the amino acid sequence[32]. If some of the e-values were between 10^{-3} and 10^{-6}, more parameters were taken into consideration including coding sequence length, genome context and domain structure.

Tree calculation was done with MEGA version 3.0 (ref. 33) from ClustalX-alignments[34] using the Neighbor-Joining method with Poisson correction. Stability of tree topology was tested by bootstrapping (1,000 replications), and by calculating maximum parsimony trees (MEGA 3.0). Gene sequences from other strains were retrieved from NCBI http://www.ncbi.nlm.nih.gov/).

Accession number. EMBL/GenBank/DDBJ: AJ965256.

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

ACKNOWLEDGMENTS
We thank Ines Müller, Monique Rönick, Janina Thiel, Beatrice Baumann and Daniela Gröger for technical assistance, Dirk Jacobs for computing assistance and Helmut Görisch for continuing support. This study was supported by the Technical University Berlin, Cornell University and the Max-Planck Society. S.Z.'s research on *Dehalococcoides* is supported by National Science Foundation grant MCB 0236044. Publication was supported by DFG project AD178/1-2.

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. Smidt, H. & de Vos, W.M. Anaerobic microbial dehalogenation. *Annu. Rev. Microbiol.* **58**, 43–73 (2004).
2. Major, D.W. *et al.* Field demonstration of successful bioaugmentation to achieve dechlorination of tetrachloroethene to ethene. *Environ. Sci. Technol.* **36**, 5106–5116 (2002).
3. Lendvay, J.M. *et al.* Bioreactive barriers: A comparison of bioaugmentation and biostimulation for chlorinated solvent remediation. *Environ. Sci. Technol.* **37**, 1422–1431 (2003).
4. Seshadri, R. *et al.* Genome sequence of the PCE-dechlorinating bacterium *Dehalococcoides ethenogenes. Science* **307**, 105–108 (2005).
5. Maymó-Gatell, X., Chien, Y.T., Gossett, J.M. & Zinder, S.H. Isolation of a bacterium that reductively dechlorinates tetrachloroethene to ethene. *Science* **276**, 1568–1571 (1997).
6. Adrian, L., Szewzyk, U., Wecke, J. & Görisch, H. Bacterial dehalorespiration with chlorinated benzenes. *Nature* **408**, 580–583 (2000).
7. Bunge, M. *et al.* Reductive dehalogenation of chlorinated dioxins by an anaerobic acterium. *Nature* **421**, 357–360 (2003).
8. Fennell, D.E., Nijenhuis, I., Wilson, S.F., Zinder, S.H. & Häggblom, M.M. *Dehalococcoides ethenogenes* strain 195 reductively dechlorinates diverse chlorinated aromatic pollutants. *Environ. Sci. Technol.* **38**, 2075–2081 (2004).
9. Karlin, S. Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr. Opin. Microbiol.* **1**, 598–610 (1998).
10. Sekizaki, T., Otani, Y., Osaki, M., Takamatsu, D. & Shimoji, Y. Evidence for horizontal transfer of SsuDAT1I restriction-modification genes to the *Streptococcus suis* genome. *J. Bacteriol.* **183**, 500–511 (2001).
11. Jansen, R., van Embden, J.A., Gaastra, W. & Schouls, L.M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43**, 1565–1575 (2002).
12. Rhee, S.K., Fennell, D.E., Häggblom, M.M. & Kerkhof, L.J. Detection by PCR of reductive dehalogenase motifs in a sulfidogenic 2-bromophenol-degrading consortium enriched from estuarine sediment. *FEMS Microbiol. Ecol.* **43**, 317–324 (2003).
13. Maillard, J., Regeard, C. & Holliger, C. Isolation and characterization of Tn-Dha1, a transposon containing the tetrachloroethene reductive dehalogenase of *Desulfitobacterium hafniense* strain TCE1. *Environ. Microbiol.* **7**, 107–117 (2005).
14. Magnuson, J.K., Romine, M.F., Burris, D.R. & Kingsley, M.T. Trichloroethene reductive dehalogenase from *Dehalococcoides ethenogenes*: sequence of *tceA* and substrate range characterization. *Appl. Environ. Microbiol.* **66**, 5141–5147 (2000).
15. Krajmalnik-Brown, R. *et al.* Genetic identification of a putative vinyl chloride reductase in *Dehalococcoides* sp. strain BAV1. *Appl. Environ. Microbiol.* **70**, 6347–6351 (2004).
16. Müller, J.A. *et al.* Molecular identification of the catabolic vinyl chloride reductase from *Dehalococcoides* sp. strain VS and its environmental distribution. *Appl. Environ. Microbiol.* **70**, 4880–4888 (2004).
17. Tropel, D. & van der Meer, J.R. Bacterial transcriptional regulators for degradation pathways of aromatic compounds. *Microbiol. Mol. Biol. Rev.* **68**, 474–500 (2004).
18. Perna, N.T. *et al.* Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**, 529–533 (2001).
19. Fukui, T. *et al.* Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1 and comparison with *Pyrococcus* genomes. *Genome Res.* **15**, 352–363 (2005).
20. 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).
21. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
22. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).
23. Delcher, A.L., Harmon, D., Kasif, S., White, O. & Salzberg, S.L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999).
24. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944–945 (2000).
25. 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).
26. Apweiler, R. *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29**, 37–40 (2001).
27. Baxevanis, A.D. The molecular biology database collection: 2003 update. *Nucleic Acids Res.* **31**, 1–12 (2003).
28. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1–6 (1997).
29. Rabus, R. *et al.* The genome sequence of an anaerobic aromatic-degrading denitrifying bacterium, strain EbN1. *Arch. Microbiol.* **183**, 27–36 (2004).
30. Myers, E.W. & Miller, W. Optimal alignments in linear space. *Comput. Appl. Biosci.* **4**, 11–17 (1988).
31. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European molecular biology open software suite. *Trends Genet.* **16**, 276–277 (2000).
32. Chien, M. *et al.* The genomic sequence of the accidental pathogen *Legionella pneumophila. Science* **305**, 1966–1968 (2004).
33. Kumar, S., Tamura, K. & Nei, M. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* **5**, 150–163 (2004).
34. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. & Higgins, D.G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882 (1997).