The complete genome of Rhodococcus sp. RHA1 provides insights into a catabolic powerhouse

Michael P. McLeod, René L. Warren, William W. L. Hsiao, Naoto Araki, Matthew Myhre, Clinton Fernandes, Daisuke Miyazawa, Wendy Wong, Anita L. Lillquist, Dennis Wang, Manisha Dosanjh, Hirofumi Hara, Anca Petrescu, Ryan D. Morin, George Yang, Jeff M. Stott, Jacqueline E. Schein, Heesun Shin, Duane Smailus, Asim S. Siddiqui, Marco A. Marra, Steven J. M. Jones, Robert Holt, Fiona S. L. Brinkman, Keisuke Miyauchi, Masao Fukuda, Julian E. Davies, William W. Mohn, and Lindsay D. Eltis

PNAS published online Oct 9, 2006;
doi:10.1073/pnas.0607048103

The complete genome of *Rhodococcus* sp. RHA1 provides insights into a catabolic powerhouse

Michael P. McLeod*, René L. Warren†, William W. L. Hsiao‡, Naoto Araki§, Matthew Myhre*, Clinton Fernandes*, Daisuke Miyazawa*, Wendy Wong*, Anita L. Lillquist*, Dennis Wang*, Manisha Dosanjh*, Hirofumi Hara*, Anca Petrescu†, Ryan D. Morin†, George Yang†, Jeff M. Stott†, Jacqueline E. Schein†, Heesun Shin†, Duane Smailus†, Asim S. Siddiqui†, Marco A. Marra†, Steven J. M. Jones†, Robert Holt†, Fiona S. L. Brinkman‡, Keisuke Miyauchi§, Masao Fukuda§, Julian E. Davies*, William W. Mohn*, and Lindsay D. Eltis*¶

*Department of Microbiology and Immunology, Life Sciences Institute, University of British Columbia, Vancouver, BC, Canada V6T 1Z3; †Michael Smith Genome Sciences Centre, Vancouver, BC, Canada V5Z 1L3; ‡Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada V5A 1S6; and §Department of Bioengineering, Nagaoka University of Technology, Nagaoka 940-2118, Japan

Communicated by James M. Tiedje, University of Michigan, East Lansing, MI, August 15, 2006 (received for review January 28, 2006)

***Rhodococcus* sp. RHA1 (RHA1) is a potent polychlorinated biphenyl-degrading soil actinomycete that catabolizes a wide range of compounds and represents a genus of considerable industrial interest. RHA1 has one of the largest bacterial genomes sequenced to date, comprising 9,702,737 bp (67% G+C) arranged in a linear chromosome and three linear plasmids. A targeted insertion methodology was developed to determine the telomeric sequences. RHA1's 9,145 predicted protein-encoding genes are exceptionally rich in oxygenases (203) and ligases (192). Many of the oxygenases occur in the numerous pathways predicted to degrade aromatic compounds (30) or steroids (4). RHA1 also contains 24 nonribosomal peptide synthase genes, six of which exceed 25 kbp, and seven polyketide synthase genes, providing evidence that rhodococci harbor an extensive secondary metabolism. Among sequenced genomes, RHA1 is most similar to those of nocardial and mycobacterial strains. The genome contains few recent gene duplications. Moreover, three different analyses indicate that RHA1 has acquired fewer genes by recent horizontal transfer than most bacteria characterized to date and far fewer than *Burkholderia xenovorans* LB400, whose genome size and catabolic versatility rival those of RHA1. RHA1 and LB400 thus appear to demonstrate that ecologically similar bacteria can evolve large genomes by different means. Overall, RHA1 appears to have evolved to simultaneously catabolize a diverse range of plant-derived compounds in an O_2-rich environment. In addition to establishing RHA1 as an important model for studying actinomycete physiology, this study provides critical insights that facilitate the exploitation of these industrially important microorganisms.**

biodegradation | actinomycete | linear chromosome | aromatic pathways | oxygenase

Actinomycetales are an order of nonmotile Gram-positive bacteria that live in a broad range of environments, including soil, water, and eukaryotic cells. This order includes some of the most important organisms known to humankind, including streptomycetes, which produce most of the antibiotics in use today, and *Mycobacterium tuberculosis*, responsible for the largest number of human deaths by bacterial infection. The most industrially important genus of actinomycetes not used for antibiotic production is arguably *Rhodococcus* (1). Applications of rhodococci include bioactive steroid production, fossil fuel biodesulfurization, and the production of acrylamide and acrylic acid, the most commercially successful application of a microbial biocatalyst (2).

The biotechnological importance of rhodococci derives from their lifestyles; these heterotrophs commonly occur in soil where they degrade a wide range of organic compounds. Their assimilatory abilities have been attributed to their diversity of enzymatic activities as well as their mycolic acids, proposed to facilitate the uptake of hydrophobic compounds (3). In addition to their industrial importance, rhodococci offer advantages as experimental systems over more familiar actinomycetes. For example, rhodococci grow faster than *M. tuberculosis* and have a simpler developmental cycle than streptomycetes. Despite their importance, rhodococci have not been well characterized. Indeed, the strains identified as rhodococci may not all belong to the same genus (3).

Rhodococcus sp. RHA1 (RHA1) was isolated from lindane-contaminated soil (4) and is best known for its potent ability to transform polychlorinated biphenyls (PCBs). Its 16*S* RNA sequence indicates that RHA1 is closely related to *Rhodococcus opacus* (3). RHA1 utilizes a wide range of aromatic compounds, carbohydrates, nitriles, steroids, and other compounds as sole sources of carbon and energy. Ongoing proteomic, transcriptomic, and gene-disruption studies of RHA1 are identifying catabolic pathways and characterizing their regulation (5–7). RHA1 contains four replicons, including three large linear plasmids (8). The plasmid genes encode important catabolic capabilities, including apparently redundant biphenyl and alkyl benzene pathways that cometabolize PCBs. The smallest plasmid, pRHL3, is an actinomycete invertron containing large terminal inverted repeats (9).

We report the complete nucleotide sequence of the RHA1 genome. The genome was manually annotated and compared with respect to gene content, organization, and recent horizontal gene transfer (HGT) with the genomes of other actinomycetes and ecologically similar soil bacteria. In addition, the phylogenies of key replicon elements were investigated. These analyses provide important insights into the biology and biotechnological applications of rhodococci.

Results and Discussion

Genome Anatomy. The genome of RHA1 contains 9,702,737 bp arranged in four linear replicons: one chromosome and three plasmids: pRHL1, pRHL2, and pRHL3 (Fig. 1, Table 1; refs. 8 and 10). The final assembly contained 154,878 shotgun reads and 3,887 finishing reads, yielding an overall coverage of 9-fold and a predicted error rate of <1 in 100 kbp. The %G+C of the RHA1 genome is 67.0% (Table 1). Historically, sequencing of such high

Author contributions: M.P.M., M.A.M., S.J.M.J., R.H., F.S.L.B., M.F., J.E.D., W.W.M., and L.D.E. designed research; M.P.M., R.L.W., W.W.L.H., N.A., D.M., M.D., A.P., R.D.M., G.Y., J.M.S., J.E.S., H.S., D.S., A.S.S., and K.M. performed research; M.P.M., R.L.W., W.W.L.H., M.M., C.F., D.M., W.W., A.L.L., D.W., H.H., F.S.L.B., and L.D.E. analyzed data; and M.P.M., F.S.L.B., W.W.M., and L.D.E. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: CDS, protein-coding sequence; DIMOB, dinucleotide mobility; DIMOB-I, DIMOB island; GI, genomic island; HGT, horizontal gene transfer; LB400, *Burkholderia xenovorans* LB400; NCBI, National Center for Biotechnology Information; RHA1, *Rhodococcus* sp. RHA1; MFS, Major Facilitator Superfamily.

Data deposition: The sequences reported in this paper have been deposited in the National Center for Biotechnology Information database (NCBI accession nos. NC_008268, NC_008269, NC_008270, and NC_008271).

¶To whom correspondence should be addressed. E-mail: leltis@interchange.ubc.ca..

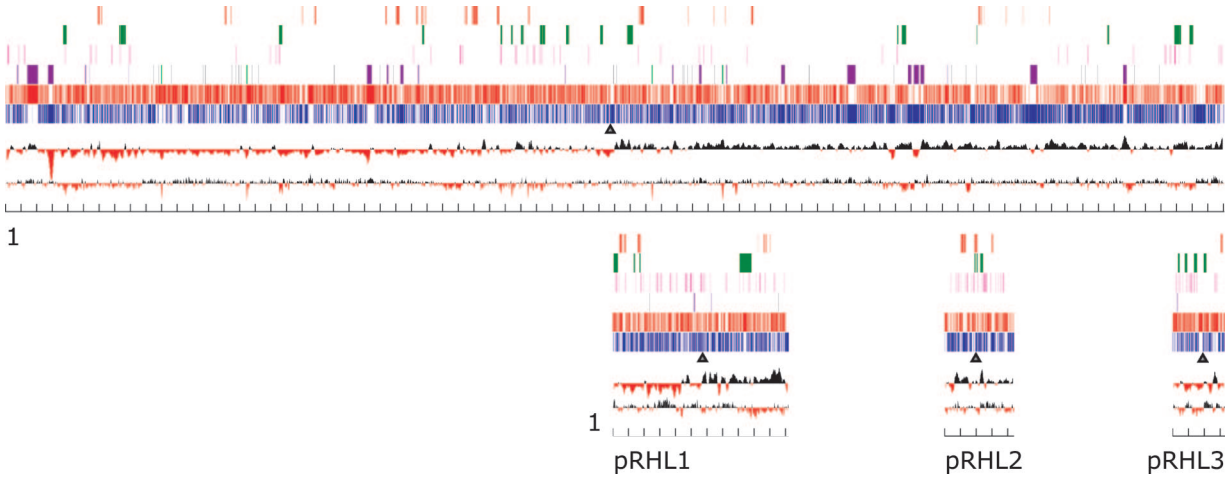

Fig. 1. Physical map of the *Rhodococcus* sp. RHA1 genome. The four maps represent each of the replicons. Ticks on the bottom scale are spaced every 100 kbp. The rows represent (counting from the bottom scale): 1, %G+C calculated over 20 kbp using a 2-kbp sliding window (red and black are below and above the mean, respectively); 2, G+C skew calculated over 5 kbp using a 1-kbp sliding window (G−C)/(C+G) (red $<0 \leq$ black); 3 and 4, genes in the forward (blue) and reverse (red) strands (box width reflects gene size); 5, tRNA (black), rRNA (green), and nonribosomal peptide synthetase and polyketide synthase (purple) genes; 6, genes associated with mobility (pink); 7, manually curated GIs (green) (Table 4); and 8, aromatic pathway genes (orange) (Table 3). Black triangles indicate the predicted *oriC* positions.

%G+C organisms has been difficult. Linear genetic elements pose an additional challenge, because the telomeres are underrepresented in most shotgun libraries, as was the case for the right pRHL1 telomere and left chromosomal telomere. Because these telomeres could not be resolved by PCR-based amplification of genomic DNA either, they were cloned by genetically tagging them as described in *Supporting Text*, which is published as supporting information on the PNAS web site. This targeted insertion methodology will be useful for sequencing other linear replicons. The linearity of the chromosome was confirmed by using PCR and substantiated by the telomere analysis described below.

The RHA1 replicons are typical actinomycete invertrons. The chromosomal inverted repeats (five mismatches in 11,943 bp) are much longer than those of the plasmids (e.g., the pRHL1 telomeres are ≈500 bp). Nevertheless, the RHA1 telomeric inverted repeats are similar to those of streptomycetes, particularly over the first 300 bp. It has been proposed that linear plasmids evolved from bacteriophages, and that linear chromosomes arose from the recombination of linear plasmids with circular chromosomes (11). Of the sequenced actinomycete genomes (www.ncbi.nlm.nih.gov/genomes/lproks.cgi), only the streptomycetes contain linear chromosomes (12, 13); actinomycetes more closely related to RHA1 possess circular chromosomes. Although the prevalence of linear chromosomes in rhodococci is unclear, the current results suggest that chromosome linearization occurred more than once during the evolution of actinomycetes.

To investigate the evolution of the RHA1 plasmids, we performed phylogenetic analyses of the ParA proteins and the telomeres, respectively. ParA is a component of the widely conserved ParABS system that partitions replicons during cell division (14).

Table 1. Summary of genomic content of *Rhodococcus* sp. RHA1

Replicon	Size, bp	G+C %	Coding* %	Proteins
Chromosome	7,804,765	67.5	91.4	7,211
pRHL1	1,123,075	65.0	82.1	1,146
pRHL2	442,536	64.0	84.0	454
pRHL3	332,361	64.9	84.8	334
Genome	9,702,737	67.0	87.2	9,145

*Does not include the regions before the first gene or after the last gene on each replicon.

Each RHA1 replicon carries *parA* and copies of *parS*. By contrast, only the chromosome and pRHL1 carry *parB* homologs; these may help partition the other two plasmids. Consistent with previous phylogenetic analyses (14), the chromosomally encoded ParA homologs had phylogenies congruous with that of their hosts. The plasmid-encoded ParAs are more diverse, and their clustering did not always follow organismal phylogeny (Fig. 2; Fig. 3, which is published as supporting information on the PNAS web site), consistent with the transmission of plasmids between species. ParA of pRHL3 clustered with a different set of actinomycete plasmid ParAs than did the more closely related ParAs of pRHL1 and pRHL2. By contrast, seven of the eight telomeric sequences in

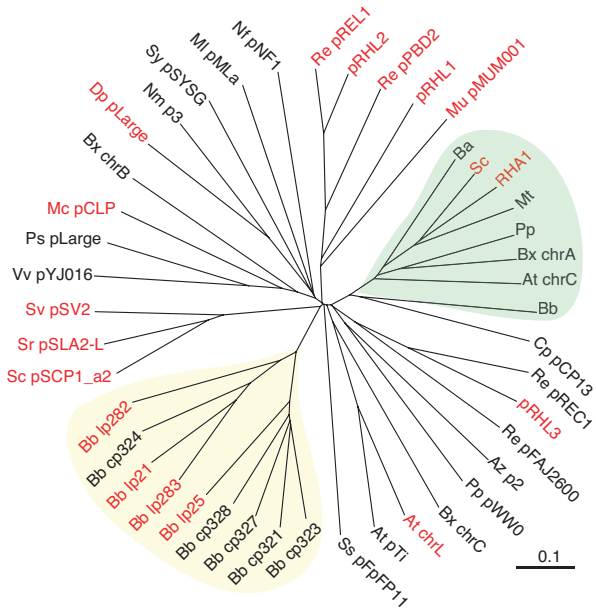

Fig. 2. Phylogenetic relationship of ParA proteins from 44 different replicons. Names in black and red identify ParAs from circular and linear replicons, respectively. Green shading identifies chromosomal ParAs. Yellow shading identifies ParAs from *B. burgdorferi* plasmids. A tree based on 90 sequences is provided in Fig. 3, together with the methods used.

MICROBIOLOGY

Table 2. Comparison of *Rhodococcus* sp. RHA1 genome with those of selected actinomycetes and ecologically related bacteria

Organism	Size, Mb*	% GC*	Proteins*	Sigma factors*	Two-component systems*†	Regulators‡	Transport proteins§	Oxygenases	σ %G+C¶	Percent DIMOB‖
Rhodococcus sp. RHA1	9.7	67.0	9,145	34	33	705	890	203	2.89	6.1
B. xenovorans LB400	9.7	62.6	8,702	22	57	745	1139	134	4.14	8.5
Bradyrhizobium japonicum USDA 110	9	64.0	8,317	23	79	541	1185	96	3.53	8.6
S. coelicolor A3(2)	8.7	72.1	7,769	58	74	671	838	70	3.42	4.0
N. farcinica IFM 10152	6.3	70.8	5,936	31	20	448	447	96	3.3	4.7
Agrobacterium tumefaciens C58	5.6	59.0	5,402	12	35	395	1010	32	2.69	3.4
Pseudomonas putida KT2440	6.2	61.5	5,350	24	58	420	721	46	3.82	5.9
Polaromonas sp. JS666	5.2	62.5	4817	12	48	405	801	45	3.96	4.2
M. tuberculosis H37Rv	4.4	53.8	3,991	13	11	166	295	45	3.32	4.5
Acinetobacter sp. ADP1	3.6	40.4	3,325	8	16	188	368	52	3.82	0.8
C. glutamicum ATCC 13032	3.3	65.6	2,993	7	10	148	402	27	3.71	5.1

Analysis of the *S. avermitilis* MA-4680 and *M. tuberculosis* CDC1551 genomes yielded results similar to those of *S. coelicolor* A3(2) and *M. tuberculosis* H37Rv, respectively.

*From www.cbs.dtu.dk/services/GenomeAtlas.

†To consistently reflect only complete systems, the number is the smaller number of predicted RR or Hsk from CBS.

‡From keyword parsing of protein annotation.

§Number of proteins in each genome with an expected cutoff of e-10 or less against the complete TransportDB (33).

¶Calculations used only CDSs > 300 bp. Values for largest replicon (chromosome) given.

‖Percent of largest replicon occurring in DIMOB-Is.

RHA1 are highly similar (Fig. 4*b*, which is published as supporting information on the PNAS web site). These results suggest that the replicons of RHA1 have diverse origins and were linearized at different times, but that similar types of linearizing elements (e.g., a particular kind of phage) were involved in linearizing the replicons and forming the telomeres.

The ParA phylogenies provide evidence for several other evolutionary events. First, the clustering of *Borrelia burgdorferi* ParAs indicates these plasmids have also interconverted their topologies. Second, as noted by others (14), the ParAs of some secondary chromosomes cluster with extrachromosomal ParAs, suggesting that such secondary chromosomes were formed by integrating chromosomal genes into a plasmid rather than duplication of the primary chromosome followed by reduction. Finally, our analyses suggest that actinomycete plasmids linearize relatively frequently in nature, consistent with the ease with which linear and circular forms can interconvert in the laboratory (15).

Several observations indicate that the RHA1 replicons depend on each other for stable maintenance. As noted above, pRHL2 and pRHL3 lack *parB* homologs. Second, pRHL2 lacks a *rep* homolog, suggesting this plasmid may require elements from other replicons for replication. Finally, each of the three RHA1 plasmids carries distant homologs of *tpg* and *tap*, encoding the telomere-associated proteins of streptomycete chromosomes. By contrast, the RHA1 chromosome does not appear to contain these genes, raising the intriguing possibility that stabilization of the chromosomal telomeres requires plasmid-encoded proteins. Consistent with the interdependence of the replicons, efforts to cure RHA1 of all plasmids have yielded only derivatives lacking pRHL2 or pRHL3.

Genome Content. RHA1 contains 9,145 predicted protein-coding sequences (CDS), of which 38.4% encode proteins of unknown function (Table 1). The latter include 2,538 conserved hypothetical proteins and 971 proteins with no hits in the National Center for Biotechnology Information (NCBI) NR database. We have transcriptomic and/or proteomic evidence for 1,025 of the genes of unknown function. An additional 1,578 proteins belong to known families, but their physiological function is unknown. The RHA1 chromosome contains four ribosomal RNA operons (Fig. 1) and encodes all ribosomal proteins, with the exception of S21, as in other actinomycetes. The 52 tRNAs, including two on pRHL1, correspond to all 20 natural amino acids and include two $tRNA^{fMet}$.

Transcriptional Control. The number of transcriptional regulatory devices in RHA1 is proportional to its genome size (see *Nocardia*; Table 2). These include 34 sigma factors and 665 genes encoding either two-component signal transduction systems or transcriptional regulators. Streptomycetes of similar genome size have significantly more, presumably because of their sporulation capacity. *Burkholderia xenovorans* LB400 (LB400), which occupies a similar ecological niche, contains 13% more genes responsible for transcriptional regulation, perhaps because of motility and chemotaxis.

Transport. Using the TransportDB, we identified 890 proteins in RHA1 associated with transport. This is >20% fewer than in LB400 (Table 2). RHA1 transporters include at least 78 members of the ATP-binding cassette (ABC) superfamily and 128 members of the Major Facilitator Superfamily (MFS). The best-represented families of ABC and MFS transporters in RHA1 are most similar to ones involved in drug export, the drug exporter-1 (Drug E1; 11% of all ABC transporters) and drug resistance-associated MFS family (37% of MFS transporters), respectively. This may reflect the fact that RHA1 shares its environment with streptomycetes and other organisms that produce antimicrobial metabolites, although it is likely that some of these transporters take up substrates. In addition, RHA1 likely produces several bioactive compounds of its own (see below). Indeed, DrrA (ro06841) and DrrB (ro06840) homologs share 65% and 51% amino acid sequence identity with the daunorubicin transporters from *Streptomyces peucetius* (16). Although many MFS transporters in RHA1 are homologs of tetracycline transporters (e.g., ro04399), RHA1 is not resistant to tetracycline, suggesting that these proteins transport other compounds. Finally, RHA1 contains proteins involved in Types I and II secretion systems as well as the components for each of the three major protein secretion pathways of bacteria; the Sec translocon (ro06153, ro01981, and ro07180), the signal recognition particle protein and RNA (ro06533 and ro04192), and the twin-arginine translocation pathway (ro00837, ro05980, and ro00836). Using PSORTb, we predicted that RHA1 exports at least 118 proteins.

Metabolism. RHA1 contains pathways to degrade many types of mono- and disaccharides, as well as multiple pathways for glucose degradation. The central metabolism of RHA1 includes gluconeogenesis, glycolysis, the Entner–Doudoroff pathway, the pentose phosphate pathway, and the tricarboxylic acid cycle. RHA1 also contains complete biosynthetic pathways for all nucleotides, nucleosides, and natural amino acids.

We analyzed the genome of RHA1 and 11 phylogenetically or ecologically related bacteria to investigate the occurrence of certain classes of enzymes based on EC number. In general, the number of oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases increases in proportion to genome size (Fig. 5, which is published as supporting information on the PNAS web site). However, certain genomes encode disproportionately large numbers of enzymes of certain classes; for RHA1, these are oxidoreductases (1,085) and ligases (192); for two streptomycetes, hydrolyases; for LB400, lyases; and for two mycobacteria, ligases. Closer examination revealed that RHA1 has disproportionate numbers of four EC classes of oxidoreductases: 1.1.-.-, 1.3.-.-, 1.13.-.-, and 1.14.-.-. The latter are oxygenases, of which RHA1 has 203 (Table 2). Numbering 19, the cyclohexanone monooxygenases (EC 1.14.13.22) are particularly overrepresented. By contrast, the number of cytochromes P450 in RHA1, 25, is typical of actinomycetes. Consistent with the limited ability of RHA1 to grow on linear alkanes, the genome encodes few monooxygenases predicted to transform such compounds. It is unclear which of the RHA1 oxygenases are catabolic and which are involved in secondary metabolism. Nevertheless, their abundance is consistent with RHA1's ability to degrade an exceptional range of aromatic compounds and steroids; oxygenases catalyze the hydroxylation and cleavage of such compounds. Furthermore, ≈77% of the oxygenases are chromosomally encoded, and only 7% are predicted to have been acquired through recent HGT (see below), suggesting that their prevalence reflects a fundamental aspect of RHA1, and perhaps rhodococcal, physiology. LB400 has 33% fewer oxygenases than RHA1 despite its similar genome size and polychlorinated biphenyl-degrading ability. This may reflect LB400's adaptation to lower O_2 availability.

Catabolism of Aromatic Compounds. The catabolism of aromatic compounds in rhodococci is organized as in the better-studied pseudomonads (17); a large number of "peripheral aromatic" pathways funnel a broad range of natural and xenobiotic compounds into a restricted number of "central aromatic" pathways (5). The latter, exemplified by the β-ketoadipate pathway, complete the transformation of these compounds to TCA cycle intermediates. RHA1 encodes at least 26 peripheral aromatic pathways and 8 central aromatic pathways (Fig. 1 and Table 3, which is published as supporting information on the PNAS web site). RHA1 and LB400 contain a similar number of such pathways, more than any genome sequenced to date.

Steroid Catabolism. The RHA1 genome is remarkably rich in steroid catabolic genes, containing up to four distinct pathways, each of which is predicted to involve aromaticization. Each of four gene clusters in RHA1 encodes homologs of each of four core structure-degrading enzymes, KshAB, KstD, TesAaAb, and TesB (18, 19). For example, the KshA homologs (ro02490, ro04538, ro05811, and ro09003) share at least 52% amino acid sequence identity with KshA of *Rhodococcus erythropolis* SQ1 (19). Similarly, the TesB homologs (ro02488, ro04541, ro05803, and ro09005) share at least 37% amino acid sequence identity and key active site residues with TesB of *Comamonas testosteroni* TA441. Finally, some of the cyclohexanone monooxygenase homologs noted above may be involved in the catabolism of rings C and D. The different clusters appear to catabolize different steroids (R. van der Geize, L.D.E., and W.W.M., unpublished work), and none appear to have been recently acquired by HGT (Fig. 1; Table 4, which is published as supporting information on the PNAS web site).

Secondary Metabolism. The RHA1 genome encodes 24 nonribosomal peptide synthetases (NRPSs) and 7 polyketide synthases. Environmental isolates, particularly actinomycetes, contain high numbers of such secondary metabolic genes (Table 2), which are involved in the biosynthesis of siderophores, cell signaling molecules, and antibiotics. Nevertheless, the relatively high number of such genes in RHA1 was somewhat unexpected considering that no rhodococcal secondary metabolite has been reported. Six of the RHA1 NRPS genes are >25 kbp in length. These RHA1 genes provide evidence that rhodococci contain an extensive secondary metabolism. The high number of transporters potentially involved in drug export is consistent with such extensive secondary metabolism of RHA1. Similarly, RHA1 contains an *opp*-encoded oligopeptide transporter (ro05349–ro05352). The peptides transported by this system regulate the expression of secondary metabolic genes (20).

HGT. Clusters of horizontally acquired genes, or genomic islands (GIs), are frequently associated with particular physiologic adaptation such as virulence, catabolism, or resistance to a toxic compound (21). To evaluate the role of HGT in shaping the RHA1 genome, we performed three analyses of increasing sophistication on increasingly restricted data sets: (*i*) a %G+C-based analysis of available bacterial genomes; (*ii*) a dinucleotide-mobility (DIMOB) analysis of 13 phylogenetically or ecologically related bacterial genomes; and (*iii*) a manually curated list of GIs in RHA1 identified by using IslandPath. The DIMOB analyses use dinucleotide bias and mobility genes to more accurately predict regions, termed DIMOB islands (DIMOB-Is), that may have been acquired by HGT (22). IslandPath (23) combined these parameters with additional tRNAs and %G+C information. Manual curation of GIs incorporated sequence similarity searches. Finally, we investigated the content of DIMOB-Is and GIs to compare the kinds of genes likely to have been acquired by HGT in RHA1 and other species.

For 345 replicons from 314 bacteria, we calculated the %G+C of each gene, the mean %G+C, and the standard deviation ($\sigma_{\%G+C}$). Although %G+C alone is insufficient to identify horizontally acquired genes (24), $\sigma_{\%G+C}$ for all genes in a replicon correlates surprisingly well with the degree of apparent recent HGT despite the influence of other factors on %G+C (25). Thus, the $\sigma_{\%G+C}$ values ranged from 2.20 to 6.86 (Table 5 and Fig. 6, which are published as supporting information on the PNAS web site). Species with the highest $\sigma_{\%G+C}$ values, such as *Neisseria* spp., are naturally competent and very nonclonal (26). Conversely, the species with the lowest $\sigma_{\%G+C}$ values, such as *Chlamydiaceae*, are ecologically isolated because of their obligate intracellular lifestyle (25, 27). The $\sigma_{\%G+C}$ values correlate with neither genome size (Fig. 7, which is published as supporting information on the PNAS web site) nor mean %G+C (Fig. 8, which is published as supporting information on the PNAS web site). Collectively, these studies suggest that $\sigma_{\%G+C}$ is a useful initial indication of the degree to which recent HGT may have shaped a genome. The $\sigma_{\%G+C}$ of RHA1, 2.89, is in the low range of the analyzed replicons, suggesting that RHA1 has been subject to relatively little recent HGT. By contrast, the $\sigma_{\%G+C}$ values for phylogenetically and ecologically similar bacteria vary considerably and is particularly high for LB400 (Table 2).

The DIMOB analyses and manually curated GIs corroborate the $\sigma_{\%G+C}$ prediction that RHA1 has been subject to relatively little recent HGT (Table 2). Together, the 17 GIs on the chromosome, 4 on pRHL1, 3 on pRHL2, and 4 on pRHL3 (Fig. 1; Table 6, which is published as supporting information on the PNAS web site) comprise ≈5% of the genome, which is lower than is observed in most bacteria (28). By contrast, our DIMOB analysis indicates that LB400 has acquired ≈10.5% of its large chromosome through

recent HGT. This is consistent with, but more conservative than, the values estimated by Chain *et al.* (29). Examination of gene content revealed no overrepresentation of any major Clusters of Orthologous Genes category in RHA1 DIMOB-Is/GIs. However, the proportion of unconserved hypothetical genes on the predicted islands is twice that of the rest of the genome. The disproportionate occurrence of such "novel" genes in GIs is consistent with what has been reported for other species and further supports the proposal that GIs may be associated with a different gene pool than the rest of the genome (22). By contrast, LB400 DIMOB-Is contained a disproportionate number of genes associated with cell motility, secretion, and defense mechanisms, as well as novel hypotheticals (Table 6).

Despite the high numbers of aromatic pathways and oxygenases in RHA1, genes encoding oxygenases were slightly underrepresented in DIMOB-Is (Table 2 and Table 7, which is published as supporting information on the PNAS web site), and <2% of the aromatic pathway genes were found in DIMOB-Is (Table 3), suggesting they were not acquired through recent HGT. By contrast, DIMOB-Is contained 26% of the aromatic catabolic pathways in LB400 (i.e., 2.7 times the density of the rest of the genome), consistent with HGT playing a more significant role in establishing the catabolic capabilities of LB400, as concluded by Chain *et al.* (29).

Many of the predicted horizontally acquired genes in RHA1 have high sequence similarity to proteins from β-proteobacteria such as *Burkholderia*. For example, the GI spanning ro08175 to ro08180 on pRHL1 specifies terephthalate catabolism and is duplicated on pRHL2. Three observations suggest this GI was acquired from an ancestor of LB400. First, the gene order of the cluster is conserved between RHA1 and LB400 but not between RHA1 and other β-proteobacteria. Second, the dinucleotide bias of the cluster is consistent with HGT into RHA1 and not into LB400. Finally, five of the six encoded proteins share higher amino acid sequence identity with homologs from β-proteobacteria (with highest identities to LB400 homologs) than from actinomycetes. As noted below, this GI is part of the largest duplication in the RHA1 genome.

Overall, RHA1 appears to have primarily gained its large genome and diverse metabolic capacity through more ancient gene duplications or acquisitions. Consistent with the conclusion that the chromosome of RHA1 has undergone relatively little recent genetic flux, this replicon contains only two intact insertion sequences, relatively few transposase genes (80), and only one identifiable pseudogene. This conclusion is in contrast to the larger role that HGT appears to have played in shaping the LB400 genome and indicates that large bacterial genomes can arise through different evolutionary processes.

Plasmid Function. Many of the above-described analyses were also performed to evaluate the gene content and function of the plasmids. No major Cluster of Orthologous Genes category is disproportionately represented on the RHA1 plasmids, but a full 50% of plasmid genes are of unknown function (Table 1). Nevertheless, the plasmids carry 11 of the 26 peripheral aromatic pathways of RHA1 (i.e., three times the density of the chromosome), suggesting the plasmids have a significant catabolic role. The plasmids also contain more insertion sequences (6), transposase genes (120), and pseudogenes (2) than the chromosome, as well as the only major duplication of the genome. Taken together, these results are consistent with the plasmids' roles as reservoirs and workshops for the evolution of novel catabolic capabilities.

Duplication and Redundancy. Catabolic redundancy is often cited as a basis for the catabolic versatility of *Rhodococcus* (1). Consistent with this notion, many of the paralogous transporter and transcriptional regulators in RHA1 are likely involved in substrate uptake and regulating catabolic pathways, respectively. Nevertheless, true functional redundancy is difficult to ascertain, because enzymes possessing similar sequences often have different substrate specificities or act under different conditions. The RHA1 genome contains short duplications, including 48 tandem duplications of genes, but only one duplication with >90% identity that exceeds 5 kbp. The latter is a 28.2-kbp region of pRHL1 and pRHL2 that shares >99% nucleotide sequence identity. Dinucleotide bias data and the occurrence of a 4-kbp insertion in the pRHL2 copy indicate that pRHL1 contains the ancestral copy. The first 19.5 kbp contain the phthalate and terephthalate catabolic genes potentially acquired from LB400, noted above.

The relatively low sequence identities of most of the RHA1 paralogs suggest these proteins have evolved distinct functions. For example, BphAaAb and EtbAaAb, two ring-hydroxylating dioxygenases, share 37% amino acid sequence identity and transform distinct but overlapping sets of aromatic compounds (M. Patrauchan and L.D.E., unpublished work). Similarly, some extradiol dioxygenases previously annotated as BphC are not upregulated on biphenyl (7) and are likely involved in degrading other compounds. Nevertheless, RHA1 clearly contains some redundant genes, including genes encoding two identical EtbAaAb dioxygenases, identical copies of the phthalate and terephthalate genes, and eight clusters of *bph*EFG homologs (20–80% amino acid sequence identity; ref. 7). The latter clusters encode a central aromatic pathway that transforms 2-hydroxypentadienoate to central metabolites. The collocation of some clusters with different peripheral aromatic pathways genes, together with emerging functional data, indicate they are involved in the catabolism of different growth substrates. Nevertheless, these pathways likely transform similar, if not identical, metabolites. The redundant *bph*EFG pathways appear to result from the acquisition of distinct peripheral aromatic pathways: there is no evidence that RHA1 benefits from the functional redundancy.

Evolutionary Divergence and Specialization. The predicted proteome of RHA1 was compared with each of those of *Nocardia farcinica* IFM 10152, *M. tuberculosis* H37Rv, *Corynebacterium glutamicum* ATCC 13032, *Streptomyces coelicolor* A3 (2) and *Frankia* sp. EAN1pec (Fig. 9, which is published as supporting information on the PNAS web site). Normalizing the numbers of reciprocal best hits with at least 30% identity over 70% of the length of both proteins to the geometric mean of the proteome size indicates that RHA1 shares 43.3%, 35.3%, 30.9%, 29.3%, and 23.3% of its proteome with each of these organisms, respectively. This result, demonstrating the overall relationship of these organisms, is consistent with 16*S* rRNA phylogeny studies (3). Our analysis identified 36 hypothetical proteins in RHA1 whose homologs are unique to actinomycetes (NCBI NR database; expected cutoff of e-04).

Syntenic plots identify four regions of conservation of gene order in the chromosomes of RHA1, *N. farcinica* IFM 10152 and *M. tuberculosis* H37Rv (Fig. 10, which is published as supporting information on the PNAS web site). The two clearest regions of conservation in RHA1 are within 1.5 Mb of a chromosomal telomere. Interestingly, the nonsyntenic regions in RHA1 contain proportionally more genes whose closest homologs are not found in actinomycetes. This suggests these regions may be unique to rhodococci and not shared with other actinomycetes. Finally, the conserved chromosomal organization in these three strains further indicates that linearization of the rhodococcal and streptomycete chromosomes were separate events.

The similarity between RHA1 and *M. tuberculosis* may provide information on the slow-growing pathogen. Conserved genes of particular interest include transporters, regulators, catabolic enzymes, cell-wall proteins, and 623 proteins of unknown function. Shared transporters include those encoded by the mammalian cell entry (*mce*) genes; RHA1 has two *mce* gene clusters compared with four in *M. tuberculosis* (30). The MCE proteins are critical virulence factors in *M. tuberculosis* that are thought to transport molecules

between the bacterium and its host. Shared regulators include the SenX3/RegX3 and MtrAB two-component systems, whose respective regulators share >88% sequence identity in the two bacteria. Shared cell-wall proteins include mycolic acid biosynthetic enzymes and "antigenic 85 complex." The latter are major components of the cell wall, possess mycolytransferase activity, and help maintain cell-wall integrity (31). Finally, the two organisms share steroid catabolic genes (R. van der Geize, L.D.E., and W.W.M., unpublished work), some of which are required for growth of *M. tuberculosis* in the macrophage. On a more general level, the numerous oxygenases in RHA1 suggest it may share adaptation to an O_2-rich environment with *M. tuberculosis*. Studies of the O_2 stress response in the two actinomycetes may yield important insights.

Concluding Remarks

Currently, the five largest complete microbial genomes (>9 Mbp each) are soil-dwelling heterotrophs (www.ncbi.nlm.nih.gov/genomes/lproks.cgi). Soil is a chemically complex environment, due in part to the wide range of compounds produced by plants. Further complexity is introduced by decomposition and soil biogenesis. Many compounds of plant origin are chemically related, although they are likely present at concentrations that individually do not support bacterial growth. A potentially successful competitive strategy for bacteria in this environment is to use the diverse compounds simultaneously. This strategy, rather than rapid response to transient nutrient sources, may underlie selective pressure for large genomes having numerous paralogous genes. The relatively low levels of recent duplication and HGT in the chromosome of RHA1 suggest that this genome is quite stable, and that RHA1's catabolic versatility has evolved primarily through ancient acquisition or duplication processes. Although recent genetic flux appears to have played a more significant role in the evolution of other large genomes, such as LB400's, the emerging trend is that the large gene repertoires of these organisms have also evolved principally through ancient processes. The finding that this is true in species as phylogenetically diverse as rhodococci and pseudomonads is remarkable and further suggests the ancient origin of this catabolic capacity. Nevertheless, the examples of functional redundancy in RHA1 central aromatic pathways suggest that selective pressure for removing such genes is low, as suggested (32). Finally, it appears that in RHA1, the plasmids represent the most rapidly evolving parts of the genome and are important reservoirs for beneficial catabolic functions. On a more applied level, the availability of the RHA1 genome sequence facilitates the exploitation of rhodococci for industrial uses such as the production of novel secondary metabolites and bioactive steroids. In addition, the genome provides surprising insights into *M. tuberculosis*, an important pathogen.

Materials and Methods

Sequencing, Assembly, Finishing, and Validation. Bacterial growth, harvesting, DNA extraction, and shotgun sequencing, including base calls, were performed by standard techniques, as described in ref. 9 and *Supporting Text*. Assembly, finishing, and validation were performed as described (9), and in *Supporting Text*.

Annotation. The genome was initially annotated by using the Oak Ridge National Laboratory Microbial Genome Pipeline (Oak Ridge, TN) and subsequently annotated as summarized below. Coding regions and ribosomal binding sites were detected as described in *Supporting Text*. The resulting list of predicted proteins was compared against a number of databases by using NCBI BLASTP (see *Supporting Text*). tRNAs were annotated with tRNAscan-SE, and rRNAs were detected by sequence similarity with known rRNAs. Insertion sequences were found by using the ISfinder database (www-is.biotoul.fr).

Analysis of Enzyme Types, Protein Families, Synteny, and GIs. Numbers of enzymes and transporters were calculated by using rpsBLAST and either the PRIAM or TransportDB databases. Protein familes were identified by using NCBI BLASTclust and a criterion of 30% amino acid sequence identity across 70% of their length. The latter criterion was also used in synteny analyses. DIMOB islands and GIs were analyzed by using IslandPath (23). Additional details are provided in *Supporting Text*.

Public Database Submission and Clone Access. Additional data and information regarding clone access are available at www.rhodococcus.ca.

Dr. Frank Larimer, Dr. Loren Hauser, Dr. David Nelson, Dr. George Fox, Dr. Marianna Patrauchan, Dr. Edmilson Gonçalves, Christine Florizone, Sachi Okamoto, Evelyn Gunn, and Dr. Robert van der Geize assisted with annotation. Dr. Louis Tisa (U.S. Department of Energy Joint Genome Institute) provided the *Frankia* sp. EAN1pec genome. Work in Canada was funded by Genome Canada and Genome BC. Work in Japan was funded by the Program for Promotion of Basic Research Activities for Innovative Biosciences and the Ministry of Education, Culture, Sports, Science, and Technology of Japan. W.W.L.H. and F.S.L.B. are recipients of fellowships from the Michael Smith Foundation for Health Research and the Canadian Institutes of Health Research.

1. van der Geize R, Dijkhuizen L (2004) *Curr Opin Microbiol* 7:255–261.
2. Banerjee A, Sharma R, Banerjee UC (2002) *Appl Microbiol Biotechnol* 60:33–44.
3. Gurtler V, Mayall BC, Seviour R (2004) *FEMS Microbiol Rev* 28:377–403.
4. Seto M, Kimbara K, Shimura M, Hatta T, Fukuda M, Yano K (1995) *Appl Environ Microbiol* 61:3353–3358.
5. Patrauchan MA, Florizone C, Dosanjh M, Mohn WW, Davies J, Eltis LD (2005) *J Bacteriol* 187:4050–4063.
6. Navarro-Llorens JM, Patrauchan MA, Stewart GR, Davies JE, Eltis LD, Mohn WW (2005) *J Bacteriol* 187:4497–4504.
7. Gonçalves ER, Hara H, Miyazawa D, Davies J, Eltis LD, Mohn WW (2006) *Appl Environ Microbiol*, 72:6183–6193.
8. Shimizu S, Kobayashi H, Masai E, Fukuda M (2001) *Appl Environ Microbiol* 67:2021–2028.
9. Warren R, Hsiao WW, Kudo H, Myhre M, Dosanjh M, Petrescu A, Kobayashi H, Shimizu S, Miyauchi K, Masai E, *et al*. (2004) *J Bacteriol* 186:7783–7795.
10. Masai E, Sugiyama K, Iwashita N, Shimizu S, Hauschild JE, Hatta T, Kimbara K, Yano K, Fukuda M (1997) *Gene* 187:141–149.
11. Chen CW, Huang CH, Lee HH, Tsai HH, Kirby R (2002) *Trends Genet* 18:522–529.
12. Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, *et al*. (2002) *Nature* 417:141–147.
13. Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M, Omura S (2003) *Nat Biotechnol* 21:526–531.
14. Yamaichi Y, Niki H (2000) *Proc Natl Acad Sci USA* 97:14656–14661.
15. Shiffman D, Cohen SN (1992) *Proc Natl Acad Sci USA* 89:6129–6133.
16. Gandlur SM, Wei L, Levine J, Russell J, Kaur P (2004) *J Biol Chem* 279:27799–27806.
17. Jiménez JI, Minambres B, García JL, Díaz E (2004) in *Pseudomonas*, ed Ramos J-L (Kluwer Academic/Plenum, New York), Vol II, pp 425–462.
18. Horinouchi M, Hayashi T, Kudo T (2004) *J Steroid Biochem Mol Biol* 92:143–154.
19. van der Geize R, Hessels GI, van Gerwen R, van der Meijden P, Dijkhuizen L (2002) *Mol Microbiol* 45:1007–1018.
20. Lorenzana LM, Perez-Redondo R, Santamarta I, Martin JF, Liras P (2004) *J Bacteriol* 186:3431–3438.
21. Hentschel U, Hacker J (2001) *Microbes Infect* 3:545–548.
22. Hsiao WW, Ung K, Aeschliman D, Bryan J, Finlay BB, Brinkman FS (2005) *PLoS Genet* 1:e62.
23. Hsiao W, Wan I, Jones SJ, Brinkman FS (2003) *Bioinformatics* 19:418–420.
24. Nakamura Y, Itoh T, Matsuda H, Gojobori T (2004) *Nat Genet* 36:760–766.
25. Brinkman FS, Blanchard JL, Cherkasov A, Av-Gay Y, Brunham RC, Fernandez RC, Finlay BB, Otto SP, Ouellette BF, Keeling PJ, *et al.* (2002) *Genome Res* 12:1159–1167.
26. Elkins C, Thomas CE, Seifert HS, Sparling PF (1991) *J Bacteriol* 173:3911–3913.
27. Read TD, Brunham RC, Shen C, Gill SR, Heidelberg JF, White O, Hickey EK, Peterson J, Utterback T, Berry K, *et al.* (2000) *Nucleic Acids Res* 28:1397–1406.
28. Philippe H, Douady CJ (2003) *Curr Opin Microbiol* 6:498–505.
29. Chain PSG, Denef VJ, Konstantinidis KT, Vergez LM, Agulló L, Reyes VL, Hauser L, Córdova M, Gómez L, González M, *et al*. (2006) *Proc Natl Acad Sci USA* 103:15280–15287.
30. Cole ST, Barrell BG (1998) *Novartis Found Symp* 217, 160–72; discussion 172–177.
31. Belisle JT, Vissa VD, Sievert T, Takayama K, Brennan PJ, Besra GS (1997) *Science* 276:1420–1422.
32. Konstantinidis KT, Tiedje JM (2004) *Proc Natl Acad Sci USA* 101:3160–3165.
33. Ren Q, Kang KH, Paulsen IT (2004) *Nucleic Acids Res* 32:D284–D288.