The complete genome sequence of the nitrile biocatalyst Rhodocccus rhodochrous ATCC BAA-870

Joni Frederick
Protein Technologies, CSIR Biosciences
Fritha Hennessy
Protein Technologies, CSIR Biosciences
Uli Horn
Meraka, CSIR
Pilar de la Torre Cortés
Technische Universiteit Delft
Marcel van den Broek
Technische Universiteit Delft
Ulrich Strych
Biology and Biochemistry, University of Houston
Richard C. Willson
University of Houston
Charles A. Hefer
Department of Biochemistry, University of Pretoria
Jean-Marc G. Daran
Technische Universiteit Delft
Trevor Sewell
Electron Microscope Unit, University of Cape town
Linda G. Otten (l.g.otten@tudelft.nl)
Delft University of Technology https://orcid.org/0000-0002-8344-9836
Dean Brady
Protein Technologies, CSIR Biosciences

Research article

Keywords:

Posted Date: December 3rd, 2019

DOI: https://doi.org/10.21203/rs.2.9383/v2

Version of Record: A version of this preprint was published on January 2nd, 2020. See the published version at https://doi.org/10.1186/s12864-019-6405-7.

Abstract

Background Rhodococci are industrially important soil-dwelling Gram-positive bacteria that are well known for both nitrile hydrolysis and oxidative metabolism of aromatics. Rhodococcus rhodochrous ATCC BAA-870 is capable of metabolising a wide range of aliphatic and aromatic nitriles and amides. The genome of the organism was sequenced and analysed in order to better understand this whole cell biocatalyst. Results The genome of R. rhodochrous ATCC BAA-870 is the first Rhodococcus genome fully sequenced using Nanopore sequencing. The circular genome contains 5.9 megabase pairs (Mbp) and includes a 0.53 Mbp linear plasmid, that together encode 7548 predicted protein sequences according to BASys annotation, and 5535 predicted protein sequences according to RAST annotation. The genome contains numerous oxidoreductases, 15 identified antibiotic and secondary metabolite gene clusters, several terpene and nonribosomal peptide synthetase clusters, as well as 6 putative clusters of unknown type. The 0.53 Mbp plasmid encodes 677 predicted genes and contains the nitrile converting gene cluster, including a nitrilase, a low molecular weight nitrile hydratase, and an enantioselective amidase. Although there are fewer biotechnologically relevant enzymes compared to those found in rhodococci with larger genomes, such as the well-known Rhodococcus jostii RHA1, the abundance of transporters in combination with the myriad of enzymes found in strain BAA-870 might make it more suitable for use in industrially relevant processes than other rhodococci. Conclusions The sequence and comprehensive description of the R. rhodochrous ATCC BAA-870 genome will facilitate the additional exploitation of rhodococci for biotechnological applications, as well as enable further characterisation of this model organism. The genome encodes a wide range of enzymes, many with unknown substrate specificities supporting potential applications in biotechnology, including nitrilases, nitrile hydratase, monooxygenases, cytochrome P450s, reductases, proteases, lipases, and transaminases.

Background

Rhodococcus is arguably the most industrially important actinomycetes genus [1] owing to its wide-ranging applications as a biocatalyst used in the synthesis of pharmaceuticals [2], in bioactive steroid production [3], fossil fuel desulphurization [4], and the production of kilotons of commodity chemicals [5]. Rhodococci have been shown to have a variety of important enzyme activities in the field of biodegradation (for reviews see [6] and [7]). These activities could also be harnessed for synthesis of various industrially relevant compounds [8]. One of the most interesting qualities of rhodococci that make them suitable for use in industrial biotechnology is their outer cell wall [9]. It is highly hydrophobic through a high percentage of mycolic acid, which promotes uptake of hydrophobic compounds. Furthermore, upon contact with organic solvents, the cell wall composition changes, becoming more resistant to many solvents and more stable under industrially relevant conditions like high substrate concentration and relatively high concentrations of both water-miscible and -immiscible solvents. This results in a longer lifetime of the whole cell biocatalyst and subsequent higher productivity.

Rhodococcal species isolated from soil are known to have diverse catabolic activities, and their genomes hold the key to survival in complex chemical environments [10]. The first full *Rhodococcus* genome

sequenced was that of *Rhodococcus jostii* RHA1 (NCBI database: NC_008268.1) in 2006 [10]. *R. jostii* RHA1 was isolated in Japan from soil contaminated with the toxic insecticide lindane (g-hexachlorocyclohexane) [11] and was found to degrade a range of polychlorinated biphenyls (PCBs) [12]. Its full genome is 9.7 Mbp, inclusive of the 7.8 Mbp chromosome and 3 plasmids (pRHL1, 2 and 3). Since then, many additional rhodococci have been sequenced by various groups and consortia (Supp. Info. Table S1). One sequencing effort to improve prokaryotic systematics has been implemented by the University of Northumbria, which showed that full genome sequencing provides a robust basis for the classification and identification of rhodococci that have agricultural, industrial and medical/veterinary significance [13].

A few rhodococcal genomes have been more elaborately described (Table 1), including *R. erythropolis* PR4 (NC_012490.1) [14] which degrades long alkanes [15]. Multiple monooxygenases and fatty acid β-oxidation pathway genes were found on the *R. erythropolis* PR4 genome and several plasmids, making this bacterium a perfect candidate for bioremediation of hydrocarbon-contaminated sites and biodegradation of animal fats and vegetable oils. The related *R. rhodochrous* ATCC 17895 (NZ_ASJJ01000002) [16] also has many mono- and dioxygenases, as well as interesting hydration activities which could be of value for the organic chemist. The oleaginous bacterium *R. opacus* PD630 is a very appealing organism for the production of biofuels and was sequenced by two separate groups. Holder *et al*. used enrichment culturing of *R. opacus* PD630 to analyse the lipid biosynthesis of the organism, and the ~300 or so genes involved in oleaginous metabolism [17]. This sequence is being used in comparative studies for biofuel development. The draft sequence of the *R. opacus* PD630 genome was only recently released (NZ_AGVD01000000) and appears to be 9.15 Mbp, just slightly smaller than that of *R. jostii* RHA1. The full sequence of the same strain was also deposited in 2012 by Chen *et al*. (NZ_CP003949) [18], who focused their research on the lipid droplets of this strain. Twenty strains of *R. fascians* were sequenced to understand the pathogenicity of this species for plants [19], which also resulted in the realisation that sequencing provides additional means to traditional ways of determining speciation in the very diverse genus of *Rhodococcus* [20]. The clinically important pathogenic strain *R. hoagii* 103S (formerly known as *R. equi* 103S) was also fully sequenced in order to understand its biology and virulence evolution (NC_014659.1) [21]. In this and other pathogenic *R. hoagii* strains, virulence genes are usually located on plasmids, which was well described for several strains including ATCC 33701 and 103 [22], strain PAM1593 [23] and 96 strains isolated from Normandy (France) [24]. As many important traits are often located on (easily transferable) plasmids, numerous rhodococcal plasmid sequences have been submitted to the NCBI (Supp. Info. Table S2). More elaborate research has been published on the virulence plasmid pFiD188 from *R. fascians* D188 [25], pB264, a cryptic plasmid from *Rhodococcus* sp. B264-1 [26], pNC500 from *R. rhodochrous* B-276 [27], and several plasmids from *R. opacus* B4 [28] and PD630 [18]. *R. erythropolis* harbours many plasmids besides the three from strain PR4, including pRE8424 from strain DSM8424 [29], pFAJ2600 from NI86/21 [30] and pBD2 from strain BD2 [31]. All these sequences have highlighted the adaptability of rhodococci and explain the broad habitat of this genus.

The versatile nitrile-degrading bacterium, *R. rhodochrous* ATCC BAA-870 [32], was isolated through enrichment culturing of soil samples from South Africa on nitrile nitrogen sources. *R. rhodochrous* ATCC BAA-870 possesses nitrile-hydrolysing activity capable of metabolising a wide range of aliphatic and aromatic nitriles and amides through the activity of nitrilase, nitrile hydratase and amidase [32-35]. These enzymes can also perform enantioselective hydrolysis of nitrile compounds selected from classes of chemicals used in pharmaceutical intermediates, such as β-adrenergic blocking agents, antitumor agents, antifungal antibiotics and antidiabetic drugs. Interestingly, the nitrile hydratase-amidase system can enantioselectively hydrolyse some compounds, while the nitrilase hydrolyses the opposite enantiomer of similar nitriles [36]. Biocatalytic nitrile hydrolysis affords valuable applications in industry, including production of solvents, extractants, pharmaceuticals, drug intermediates, and pesticides [37-40]. Herein, we describe the sequencing and annotation of *R. rhodochrous* ATCC BAA-870, identifying the genes associated with nitrile hydrolysis as well as other genes for potential biocatalytic applications. The extensive description of this genome and the comparison to other sequenced rhodococci will add to the knowledge of the *Rhodococcus* phylogeny and its industrial capacity.

Results

Genome preparation, sequencing and assembly

The genome of *R. rhodochrous* ATCC BAA-870 was originally sequenced in 2009 by Solexa Illumina with sequence reads of average length 36 bps, resulting in a coverage of 74%, with an apparent raw coverage depth of 36x. An initial assembly of this 36-cycle, single-ended Illumina library, together with a mate-pair library, yielded a 6 Mbp genome of 257 scaffolds. A more recently performed paired-end Illumina library combined with the mate-pair library reduced this to only 6 scaffolds (5.88 Mbp). Even after several rounds of linking the mate-pair reads, we were still left with 3 separate contiguous sequences (contigs). The constraint was caused by the existence of repeats in the genome of which one was a 5.2 kb contig that, based on sequence coverage, must exist in four copies, containing 16S-like genes. Applying third generation sequencing (Oxford Nanopore Technology) enabled the full assembly of the genome, while the second generation (Illumina) reads provided the necessary proof-reading. This resulted in a total genome size of 5.9 Mbp, consisting of a 5.37 Mbp circular chromosome and a 0.53 Mbp linear plasmid. The presence of the plasmid was confirmed by performing Pulse Field Gel Electrophoresis using non-digested DNA [41]. The complete genome sequence of *R. rhodochrous* ATCC BAA 870 is deposited at NCBI GenBank, with Bioproject accession number PRJNA487734, and Biosample accession number SAMN09909133.

Taxonomy and lineage of *R. rhodochrous* ATCC BAA-870

The *R. rhodochrous* ATCC BAA-870 genome encodes four 16S rRNA genes, consistent with the average 16S gene count statistics of *Rhodococcus* genomes. From a search of The Ribosomal RNA Database, of the 28 *Rhodococcus* genome records deposited in the NCBI database, 16S rRNA gene counts range from

3-5 copies, with an average of 4 [42]. Of the four 16S rRNA genes found in *R. rhodochrous* ATCC BAA-870, two pairs are identical (i.e. there are two copies of two different 16S rRNA genes). One of each identical 16S rRNA gene was used in nucleotide-nucleotide BLAST for highly similar sequences [43]. BLAST results (complete sequences with percentage identity greater than 95.5%) were used for comparison of *R. rhodochrous* ATCC BAA-870 to other similar species using 16S rRNA multiple sequence alignment and phylogeny in ClustalO and ClustalW respectively [44-46] (Figure 1). Nucleotide BLAST results of the two different *R. rhodochrous* ATCC BAA-870 16S rRNA genes show closest sequence identities to *Rhodococcus* sp. 2G and *R. pyridinovorans* SB3094, with either 100% or 99.74% identities to both strains depending on the 16S rRNA copy.

We used the *in silico* DNA-DNA hybridisation tool, the Genome-to-Genome Distance Calculator (GGDC) version 2.1 [47-49], to assess the genome similarity of *R. rhodochrous* ATCC BAA-870 to its closest matched strains based on 16S rRNA alignment (*R. pyridinovorans* SB3094 and *Rhodococcus* sp. 2G). The results of genome based species and subspecies delineation, and difference in GC content, is summarised (Supp. Info Table S3), with *R. jostii* RHA1 additionally shown for comparison. GC differences of below 1% would indicate the same species, and therefore *R. rhodochrous* ATCC BAA-870 cannot be distinguished from the other strains based on GC content. Digital DNA-DNA hybridisation values of more than 70% and 79% are the threshold for delineating type strains and subspecies. While 16S rRNA sequence alignment and GC content suggest that *R. rhodochrous* ATCC BAA-870 and *R. pyridinovorans* SB3094 and *Rhodococcus* sp. 2G are closely related strains, the GGDC supports their delineation at the subspecies level.

Genome annotation

The assembled genome sequence of *R. rhodochrous* ATCC BAA-870 was submitted to the Bacterial Annotation System web server, BASys, for automated, in-depth annotation [50]. The BASys annotation was performed using raw sequence data for both the chromosome and plasmid of *R. rhodochrous* ATCC BAA-870 with a total genome length of 5.9 Mbp, in which 7548 genes were identified and annotated (Figure 2, Table 1). The plasmid and chromosome encode a predicted 677 and 6871 genes, respectively. 56.9% of this encodes previously identified proteins of unknown function and includes 305 conserved hypothetical proteins. A large proportion of genes are labelled ‘hypothetical’ based on sequence similarity and/or the presence of known signature sequences of protein families (Figure 3). Out of 7548 BASys annotated genes, 1481 are annotated enzymes that could be assigned an EC number (20%). Confirmation of annotation was performed manually for selected sequences. In BASys annotation, COGs (Clusters of Orthologous Groups) were automatically delineated by comparing protein sequences encoded in complete genomes representing major phylogenetic lineages [51]. As each COG consists of individual proteins or groups of paralogs from at least 3 lineages, it corresponds to an ancient conserved domain [52, 53]. A total of 3387 genes annotated in BASys were assigned a COG function (44.9% of annotated

genes), while 55 and 59% of annotated genes on the chromosome and plasmid respectively have unknown function.

The genome sequence run through RAST (Rapid Annotation using Subsystem Technology) predicted fewer (5535) protein coding sequences than BASys annotation (Figure 4), showing the importance of the bioinformatics tool used. The RAST subsystem annotations are assigned from the manually curated SEED database, in which hypothetical proteins are annotated based only on related genomes. RAST annotations are grouped into two sets (genes that are either in a subsystem, or not in a subsystem) based on predicted roles of protein families with common functions. Genes belonging to recognised subsystems can be considered reliable and conservative gene predictions. Annotation of genes that do not belong to curated protein functional families however (i.e. those not in the subsystem), may be underpredicted by RAST, since annotations belonging to subsystems are based only on related neighbours. Based on counts of total genes annotated in RAST (5535), only 26% are classified as belonging to subsystems with known functional roles, while 74% of genes do not belong to known funtional roles. Overall 38% of annotated genes were annotated as hypothetical irrespective of whether or not they were included in subsystems. The use of two genome annotation pipelines allowed us to manually compare and search for enzymes, or classes of enzymes, using both the subsystem based, known functional pathway categories provided by RAST (Figure 4), as well as the COG classification breakdowns provided by BASys (Figure 3 and Supp. Info. Table S4). From both the RAST and BASys annotated gene sets, several industrially relevant enzyme classes are highlighted and discussed further in the text.

The average GC content of the *R. rhodochrous* ATCC BAA-870 chromosome and plasmid is 68.2% and 63.8%, respectively. The total genome has a 90.6% coding ratio, and on average large genes, consisting of ~782 bps per gene. Interestingly, the distribution of protein lengths on the chromosome is bell-shaped with a peak at 350 bps per gene, while the genes on the plasmid show two size peaks, one at 100 bps and one at 350 bps.

Transcriptional control

Transcriptional regulatory elements in *R. rhodochrous* ATCC BAA-870 include 18 sigma factors, at least 8 regulators of sigma factor, and 118 other genes involved in signal transduction mechanisms (COG T), 261 genes encoding transcriptional regulators and 47 genes encoding two-component signal transduction systems. There are 129 proteins in *R. rhodochrous* ATCC BAA-870 associated with translation, ribosomal structure and biogenesis (protein biosynthesis). The genome encodes all ribosomal proteins, with the exception of S21, as occurs in other actinomycetes. RAST annotation predicts 66 RNAs. The 56 tRNAs correspond to all 20 natural amino acids and include two $tRNA^{fMet}$. Additional analysis of the genome sequence using the tRNA finding tool tRNAScan-SE v. 2.0 [54, 55] confirms the presence of 56 tRNA genes in the *R. rhodochrous* ATCC BAA-870 genome, made up of 52

tRNA genes encoding natural amino acids, 2 pseudogenes, one tRNA with mismatched isotype and one +9 Selenocysteine tRNA.

Protein location in the cell

It is often critical to know where proteins are located in the cell in order to understand their function [56], and prediction of protein localization is important for both drug targeting and protein annotation. In this study, prediction was done using the BASys SignalP signal prediction service [50]. The majority of annotated proteins are soluble and located in the cytoplasm (83%), while proteins located at the cellular membrane make up 16% of the total. Cell membrane proteins include proteins that form part of lipid anchors, peripheral and integral cell membrane components, as well as proteins with single or multiple pass functions. Of the membrane proteins in *R. rhodochrous* ATCC BAA-870, 47% constitute single-pass, inner or peripheral membrane proteins, while 41% are multi-pass membrane proteins. Most of the remaining proteins will be transported over the membrane. The periplasm contains proteins distinct from those in the cytoplasm which have various functions in cellular processes, including transport, degradation, and motility. Periplasmic proteins would mostly include hydrolytic enzymes such as proteases and nucleases, proteins involved in binding of ions, vitamins and sugar molecules, and those involved in chemotaxic responses. Detoxifying proteins, such as penicillin binding proteins, are also presumed to be located mostly in the periplasm.

Transport and Metabolism

A total of 1504 genes are implicated in transport. Numerous components of the ubiquitous transporter families, the ATP-Binding Cassette (ABC) superfamily and the Major Facilitator Superfamily (MFS), are present in *Rhodococcus* strain BAA-870. MFS transporters are single-polypeptide secondary carriers capable only of transporting small solutes in response to chemiosmotic ion gradients [57, 58]. *R. rhodochrous* ATCC BAA-870 has 81 members of the MFS, mostly from the phthalate permease and sugar transporter families. There are dozens of families within the ABC superfamily, and each family generally correlates with substrate specificity. Transporters of *R. rhodochrous* ATCC BAA-870 include at least 122 members of the ABC superfamily, which includes both uptake and efflux transport systems. Out of 3387 genes assigned a COG function, 1486 (44%) are associated with transport and metabolism. These include 206 carbohydrate, 271 amino acid, 121 coenzyme, 236 inorganic ion, 411 lipid and 67 nucleotide transport and metabolism gene functions, and 174 secondary metabolite biosynthesis, transport and catabolism genes.

The complete biosynthetic pathways for all nucleotides, nucleosides and natural amino acids are also contained in the genome of *R. rhodochrous* ATCC BAA-870. The central metabolism of strain BAA-870 includes glycolysis, gluconeogenesis, the pentose phosphate pathway, and the tricarboxylic acid cycle, a

typical metabolic pathway for an aerobic organism. There is no evidence for the Entner-Doudoroff pathway (including 6-phosphogluconate dehydratase and 2-keto-3-deoxyphosphogluconate aldolase) in *R. rhodochrous* ATCC BAA-870. General metabolic enzymes such as lipases and esterases [59, 60] are, however, present in this strain.

Aromatic Catabolism and oxidoreductases

As deduced from the better characterized pseudomonads [61], a large number of 'peripheral aromatic' pathways funnel a broad range of natural and xenobiotic compounds into a restricted number of 'central aromatic' pathways. Analysis of the *R. rhodochrous* ATCC BAA-870 genome suggests that at least four major pathways exist for the catabolism of central aromatic intermediates. The dominant portion of annotated enzymes is involved in oxidation and reduction, which is typical for catabolism. There are about 500 oxidoreductase related genes including oxidases, hydrogenases, reductases, oxygenases, dioxygenases, cytochrome P450s, catalases and peroxiredoxins. Furthermore, there are 71 monooxygenase genes, 11 of which are on the plasmid.

In *R. rhodochrous* ATCC BAA-870 there are 14 cytochrome P450 genes and 87 oxygenase genes. It is unclear which oxygenases are catabolic and which are involved in secondary metabolism. Oxygenase genes include three cyclopentanone monooxygenases (EC 1.14.13.16) and a phenol monooxygenase (EC 1.14.13.7) on the plasmid, a methane monooxygenase (EC 1.14.13.25), two alkane 1-monooxygenases (EC 1.14.15.3) and five phenylacetone monooxygenases (EC 1.14.13.92), one of which is on the plasmid.

Nitrile biocatalysis

Rhodococci are well known for their application in the commercial manufacture of amides and acids through hydrolysis of the corresponding nitriles. *R. rhodochrous* J1 can convert acrylonitrile to the commodity chemical acrylamide [62], and both Mitsubishi Rayon Co., Ltd (Japan) and Senmin (South Africa) are applying this biocatalytic reaction at the multi-kiloton scale. Lonza Guangzhou Fine Chemicals use the same biocatalyst for large-scale commercial synthesis of nicotinamide from 3-cyanopyridine [63]. Both processes rely on rhodococcal nitrile hydratase activity [64].

As *R. rhodochrous* ATCC BAA-870 was isolated from a nitrile enrichment culture [32], we were very interested in its nitrile degrading enzymes. As expected, strain BAA-870 contains several nitrile converting enzymes: a low molecular weight cobalt-containing nitrile hydratase and two nitrilases, along with several amidases. The low molecular weight nitrile hydratase and two amidase genes form a cluster, along with their associated regulatory elements, including cobalt transport genes necessary for uptake of cobalt for inclusion in the nitrile hydratase active site. Interestingly, this cluster is found on the plasmid. The alternative nitrile hydrolysis enzyme, nitrilase, is also found in *R. rhodochrous* ATCC BAA-870. It expresses an enantioselective aliphatic nitrilase encoded on the plasmid, which is induced by dimethylformamide

[36]. Another nitrilase/cyanide hydratase family protein is also annotated on the plasmid (this study) but has not been characterised.

Secondary metabolism and metabolite biosynthesis clusters

The ongoing search for new siderophores, antibiotics and antifungals has led to a recent explosion of interest in mining bacterial genomes [65], and the secondary metabolism of diverse soil-dwelling microbes remains relatively underexplored despite their huge biosynthetic potential [66]. Evidence of an extensive secondary metabolism in *R. rhodochrous* ATCC BAA-870 is supported by the presence of at least 227 genes linked to secondary metabolite biosynthesis, transport and catabolism. The genome contains 15 biosynthetic gene clusters associated with secondary metabolites or antibiotics, identified by antiSMASH (antibiotics and Secondary Metabolite Analysis Shell pipeline, version 5.0.0) [67, 68]. Biosynthetic gene clusters identified in *R. rhodochrous* BAA-870 include ectoine (1,4,5,6-tetrahydro-2-methyl-4-pyrimidinecarboxylic acid), butyrolactone, betalactone, and type I polyketide synthase (PKS) clusters, as well as three terpene and seven nonribosomal peptide synthetase (NRPS) clusters. An additional six putative biosynthetic clusters were identified on the *R. rhodochrous* ATCC BAA-870 plasmid, four of an unknown type, and the other two with low similarity to enterobactin and lipopolysaccharide biosynthetic clusters.

Soil-dwelling rhodococci present rich possible sources of terpenes and isoprenoids which are implicated in diverse structural and functional roles in nature. AntiSMASH analysis revealed 3 terpene biosynthetic clusters in the genome of *R. rhodochrous* ATCC BAA-870. Some of the examples of annotated *R. rhodochrous* ATCC BAA-870 genes related to terpene and isoprenoid biosynthesis include phytoene saturase and several phytoene synthases, dehydrogenases and related proteins, as well as numerous diphosphate synthases, isomerases and epimerases. The genome also contains lycopene cyclase, a novel non-redox flavoprotein [69], farnesyl diphosphate synthase, farnesyl transferase, geranylgeranyl pyrophosphate synthetases and digeranylgeranylglycerophospholipid reductase. Farnesyl diphosphate synthase and geranylgeranyl pyrophosphate synthases are potential anticancer and anti-infective drug targets [70]. In addition, the *R. rhodochrous* ATCC BAA-870 plasmid encodes a lactone ring-opening enzyme, monoterpene epsilon-lactone hydrolase.

The *R. rhodochrous* ATCC BAA-870 genome has two PKS genes, one regulator of PKS expression, one exporter of polyketide antibiotics, as well as three polyketide cyclase/dehydrases involved in polyketide biosynthesis. In addition, there are two actinorhodin polyketide dimerases. A total of five NRPS genes for secondary metabolite synthesis can be found on the chromosome. *R. rhodochrous* ATCC BAA-870 contains 4 probable siderophore-binding lipoproteins, 3 probable siderophore transport system permeases, and two probable siderophore transport system ATP-binding proteins. Other secondary metabolite genes found in *R. rhodochrous* ATCC BAA-870 include a dihydroxybenzoic acid-activating enzyme (2,3-dihydroxybenzoate-AMP ligase bacillibactin siderophore), phthiocerol/phenolphthiocerol

synthesis polyketide synthase type I, two copies of linear gramicidin synthase subunits C and D genes, and tyrocidine synthase 2 and 3.

CRISPR

One putative clustered regularly interspaced short palindromic repeat (CRISPR) is contained in the *R. rhodochrous* ATCC BAA-870 genome, according to analysis by CRISPRCasFinder [71]. Associated CRISPR genes are not automatically detected by the CRISPRCasFinder tool, but manual searches of the annotated genome for Cas proteins reveal possible Cas9 candidate genes within the *R. rhodochrous* ATCC BAA-870 genome, including a *ruv*C gene, and HNH endonuclease and nuclease genes.

Horizontal gene transfer

Organisms acquire diverse metabolic capacity through gene duplications and acquisitions, typically mediated by transposases. Analysis using IslandViewer (for computational identification of genomic islands) [72] identifies 10 possible large genomic island regions in *R. rhodochrous* ATCC BAA-870 which may have been obtained through horizontal mobility. Half of these genomic islands are located on the plasmid and make up 90% of the plasmid coding sequence. The low molecular weight cobalt-containing nitrile hydratase operon is located on an 82.5 kbp genomic island that includes 57 predicted genes in total. Other genes of interest located on this same genomic island include crotonase and enoyl-CoA hydratase, 10 dehydrogenases including four acyl-CoA dehydrogenases and two aldehyde dehydrogenases, four hydrolases including 5-valerolactone hydrolase and amidohydrolase, beta-mannosidase, haloacid dehalogenase and five oxidoreductases. The *R. rhodochrous* ATCC BAA-870 genome contains 31 transposase genes found in the genomic regions identified by IslandViewer, one of which is from the IS30 family, a ubiquitous mobile insertion element in prokaryotic genomes [73]. Other transposase genes belonging to at least 10 different families of insertion sequences were identified in *R. rhodochrous* ATCC BAA-870, including ISL3, IS5, IS701, two IS1634, three IS110, three IS3, three IS256, five IS21, and six IS630 family transposases. The majority of these transposons (27 of the 31 identified by IslandViewer) are located on the plasmid.

Discussion

Sequencing and annotation

New sequencing technology has revolutionized the cost and pace of obtaining genome information, and there has been a drive to sequence the genomes of organisms which have economic applications, as well as those with environmental interest [74, 75]. This holds true for *Rhodococcus* genomes, of which only two were sequenced in 2006, while 13 years later 353 genomes are now available, mainly due to Whole Genome Shotgun sequencing efforts (Supp. Info. Table S1). The impact of better and faster sequencing,

using improved sequencing techniques, is evident in this case of sequencing the *R. rhodochrous* ATCC BAA-870 genome: an initial assembly of a 36-cycle, single-ended Illumina library sequence performed in 2009, together with a mate-pair library, yielded a 6 Mbp genome of 257 scaffolds. A more recently performed paired-end Illumina library combined with the previous mate-pair library reduced this to only 6 scaffolds (5.88 Mbp), showing the improved second-generation sequencing results in only 10 years' time. The presence of four copies of 16S-like genes was the main reason for the assembly to break into 6 scaffolds. Using third generation sequencing (Nanopore), this problem was overcome, and the genome could be fully assembled. Hence, we see second generation sequencing evolving to produce higher quality assemblies, but the combination with 3rd generation sequencing was necessary to obtain the full-length closed bacterial genome.

It has been assumed that the annotation of prokaryotic genomes is simpler than that of the intron-containing genomes of eukaryotes. However, annotation has been shown to be problematic, especially with over- or under-prediction of small genes where the criterion used to decide the size of an open reading frame (ORF) can systematically exclude annotation of small proteins [76]. Warren *et al*. 2010, used high performance computational methods to show that current annotated prokaryotic genomes are missing 1153 candidate genes that have been excluded from annotations based on their size [76]. These missing genes do not show strong similarities to gene sequences in public databases, indicating that they may belong to gene families which are not currently annotated in genomes. Furthermore, they uncovered ~38,895 intergenic ORFs, currently labelled as 'putative' genes only by similarity to annotated genes, meaning that the annotations are absent. Therefore, prokaryotic gene finding and annotation programs do not accurately predict small genes, and are limited to the accuracy of existing database annotations. Hypothetical genes (genes without any functional assignment), genes that are assigned too generally to be of use, misannotated genes and undetected real genes remain the biggest challenges in assigning annotations to new genome data [77-80]. As such, there is the possibility that we are under-estimating the number of genes present on this genome.

Apart from possible misannotation, the algorithm or software used for the annotation plays a huge role in the outcome. In this research both BASys (Figure 2) and RAST (Figure 4) were used as annotation tools, resulting in 7548 and 5535 predicted genes respectively. BASys annotation may provide an overprediction of gene numbers, due to sensitive GLIMMER *ab initio* gene prediction methods that can give false positives for higher GC content sequences [81]. This shows the importance of the bioinformatics tool used, which makes comparison to other genomes more difficult.

Size and content of the genome

The genomic content of *R. rhodochrous* ATCC BAA-870 was outlined and compared to other rhodococcal genomes. Sequences of other *Rhodococcus* genomes were obtained from the Genome database at NCBI [82] and show a large variation in genome size between 4 and 10 Mbp (Supp. Info. Table S1), with an average of 6.1 ± 1.6 Mbp. The apparent total genome size of *R. rhodochrous* ATCC BAA-870, 5.9 Mbp

(consisting of a 5.37 Mbp genome and a 0.53 Mbp plasmid), is close to the average. From the well-described rhodococci (Table 1), the genome of *R. jostii* RHA1 is the largest rhodococcal genome sequenced to date (9.7 Mbp), but only 7.8 Mbp is chromosomal, while the pathogenic *R. hoagii* genomes are the smallest at ~5 Mbp. All rhodococcal genomes have a high GC content, ranging from 62 – 71%. The average GC content of the *R. rhodochrous* ATCC BAA-870 chromosome and plasmid is 68.2% and 63.8%, respectively. *R. jostii* RHA1 has the lowest percentage coding DNA (87%), which is predictable given its large overall genome size, while *R. rhodochrous* ATCC BAA-870 has a 90.6% coding ratio, which is in line with its smaller total size. Interestingly, the distribution of protein lengths on the chromosome is different from those on the plasmid. Together with the lower GC content, this shows that the plasmid content was probably acquired over different occasions [83].

Fundamental and applicable biocatalytic properties of rhodococci

Catabolism typically involves oxidative enzymes. The presence of multiple homologs of catabolic genes in all *Rhodococcus* species suggests that they may provide a comprehensive biocatalytic profile [1]. *R. rhodochrous* ATCC BAA-870 combines this with multiple transport systems (44% of total COG annotated genes), highlighting the metabolic versatility of this *Rhodococcus*, which facilitates the use of whole cells in biotechnological applications.

McLeod *et al*. reported that *R. jostii* RHA1 contains genes for the Entner-Doudoroff pathway (which requires 6-phosphogluconate dehydratase and 2-keto-3-deoxyphosphogluconate aldolase to create pyruvate from glucose) [10]. The Entner-Doudoroff pathway is, however, rare in Gram positive organisms which preferably use glycolysis for a richer ATP yield. There is no evidence of this pathway existing in *R. rhodochrous* ATCC BAA-870, indicating that it is not a typical rhodococcal trait, but the RHA1 strain must have acquired it rather recently.

Analysis of the *R. rhodochrous* ATCC BAA-870 genome suggests that at least four major pathways exist for the catabolism of central aromatic intermediates, comparable to the well-defined aromatic metabolism of *Pseudomonas putida* KT2440 strain [84]. In *R. rhodochrous* ATCC BAA-870 the dominant portion of annotated enzymes are involved in oxidation and reduction. There are about 500 oxidoreductase related genes, which is quite a high number compared to other bacteria of the same size, but in line with most other (sequenced) rhodococci [85]. *Rhodococcus* genomes usually encode large numbers of oxygenases [1], which is also true for strain BAA-870 (71). Some of these are flavonoid proteins with diverse useful activities [86], which includes monooxygenases capable of catalysing Baeyer–Villiger oxidations wherein a ketone is converted to an ester [87, 88].

The 14 cytochrome P450 genes in *R. rhodochrous* ATCC BAA-870 reflects a fundamental aspect of rhodococcal physiology. Similarly, the number of cytochrome P450 genes in *R. jostii* RHA1 is 25 (proportionate to the larger genome) which is typical of actinomycetes. Although it is unclear which oxygenases in *R. rhodochrous* ATCC BAA-870 are catabolic and which are involved in secondary

metabolism, their abundance is consistent with a potential ability to degrade an exceptional range of aromatic compounds (oxygenases catalyse the hydroxylation and cleavage of these compounds). Rhodococci are well known to have the capacity to catabolise hydrophobic compounds, including hydrocarbons and polychlorinated biphenyls (PCBs), mediated by a cytochrome P450 system [89-92]. Cytochrome P450 oxygenase is often found fused with a reductase, as in *Rhodococcus* sp. NCIMB 9784 [93]. Genes associated with biphenyl and PCB degradation are found in multiple sites on the *R. jostii* RHA1 genome, both on the chromosome as well as on linear plasmids [1]. *R. jostii* RHA1 was also found to show lignin-degrading activity, possibly based on the same oxidative capacity as that used to degrade biphenyl compounds [94].

The oxygenases found in rhodococci include multiple alkane monooxygenases (genes *alkB1*– *alkB4*) [95], steroid monooxygenase [96], styrene monooxygenase [97], peroxidase [98] and alkane hydroxylase homologs [99]. *R. rhodochrous* ATCC BAA-870 has 87 oxygenase genes while the PCB degrading *R. jostii* RHA1 has 203 oxygenases, including 19 cyclohexanone monooxygenases (EC 1.14.13.22), implying that of the two, strain BAA-870 is less adept at oxidative catabolism. Rhodococcal cyclohexanone monooxygenases can be used in the synthesis of industrially interesting compounds from cyclohexanol and cyclohexanone. These include adipic acid, caprolactone (for polyol polymers) and 6-hydroxyhexanoic acid (for coating applications) [64]. Chiral lactones can also be used as intermediates in the production of prostaglandins [100]. The same oxidative pathway can be used to biotransform cyclododecanone to lauryl lactone or 12-hydroxydodecanoic acid [101, 102]. Cyclododecanone monooxygenase of *Rhodococcus* SC1 was used in the kinetic resolution of 2-substituted cycloketones for the synthesis of aroma lactones in good yields and high enantiomeric excess [103]. Similar to *R. jostii* RHA1, *R. rhodochrous* ATCC BAA-870 encodes several monooxygenases. All these redox enzymes could be interesting for synthetic purposes in industrial biotechnological applications.

The presence of an ectoine biosynthesis cluster suggests that *R. rhodochrous* ATCC BAA-870 has effective osmoregulation and enzyme protection capabilities. *R. rhodochrous* ATCC BAA-870, together with other *Rhodococcus* strains, is able to support diverse environments and can tolerate harsh chemical reactions when used as whole cell biocatalysts, and it is likely that ectoine biosynthesis plays a role in this. Regulation of cytoplasmic solute concentration through modulation of compounds such as inorganic ions, sugars, amino acids and polyols provides a versatile and effective osmo-adaptation strategy for bacteria in general. Ectoine and hydroxyectoine are common alternate osmoregulation solutes found especially in halophilic and halotolerant microorganisms [104, 105], and hydroxyectoine has been shown to confer heat stress protection *in vivo* [106]. Ectoines provide a variety of useful biotechnological and biomedical applications [107], and strains engineered for improved ectoine synthesis have been used for the industrial production of hydroxyectoine as a solute and enzyme stabiliser [108, 109]. The special cell-wall structure of rhodococci might make these organisms a better choice as production organism.

Terpenes and isoprenoids provide a rich pool of natural compounds with applications in synthetic chemistry, pharmaceutical, flavour, and even biofuel industries. The structures, functions and chemistries

employed by the enzymes involved in terpene biosynthesis are well known, especially for plants and fungi [70, 110]. However, it is only recently that bacterial terpenoids have been considered as a possible source of new natural product wealth [111, 112], largely facilitated by the explosion of available bacterial genome sequences. Interestingly, bacterial terpene synthases have low sequence similarities, and show no significant overall amino acid identities compared to their plant and fungal counterparts. Yamada *et al*. used a genome mining strategy to identify 262 bacterial synthases, and subsequent isolation and expression of genes in a *Streptomyces* host confirmed the activities of these predicted genes and led to the identification of 13 previously unknown terpene structures [111]. The three biosynthetic clusters annotated in strain BAA-870 may therefore be an underrepresentation of possible pathways for these valuable compounds.

A total of five NRPS genes for secondary metabolite synthesis can be found on the chromosome, which is not much compared to *R. jostii* RHA1 that contains 24 NRPS and seven PKS genes [10]. Like strain ATCC BAA-870, *R. jostii* RHA1 was also found to possess a pathway for the synthesis of a siderophore [113]. The multiple PKS and NRPS clusters suggest that *R. rhodochrous* ATCC BAA-870 may host a significant potential source of molecules with immunosuppressing, antifungal, antibiotic and siderophore activities [114].

Nitrile conversion

Many rhodococci can hydrolyse a wide range of nitriles [115-118]. The locations and numbers of nitrile converting enzymes in the available genomes of *Rhodococcus* were identified and compared to *R. rhodochrous* ATCC BAA-870 (Table 2). *R. rhodochrous* ATCC BAA-870 contains several nitrile converting enzymes which is in line with previous activity assays using this *Rhodococcus* strain [33, 34]. However, in most *R. rhodochrous* strains these enzymes are on the chromosome, while in *R. rhodochrous* ATCC BAA-870, they are found on a plasmid. In *R. rhodochrous* ATCC BAA-870 the nitrile hydratase is expressed constitutively, explaining why this strain is an exceptional nitrile biocatalyst [36]. Environmental pressure through chemical challenge by nitriles may have caused the elimination of regulation of the nitrile biocatalyst by transferring it to a plasmid.

The *R. jostii* RHA1 16S RNA sequence indicates that it is closely related to *R. opacus* [10] according to the taxonomy of Gürtler *et al.* (Figure 1) [119]. *R. jostii* RHA1 expresses a nitrile hydratase (an acetonitrile hydratase) and utilises nitriles such as acetonitrile, acrylonitrile, propionitrile and butyronitrile [120], while *R. opacus* expresses nitrile hydrolysis activity [115]. *R. erythropolis* PR4 expresses a Fe-type nitrile hydratase [121], and *R. erythropolis* strains are well known for expressing this enzyme [115, 122, 123] as part of a nitrile metabolism gene cluster [119]. This enzyme has been repeatedly determined in this species from isolated diverse locations [124], expressing broad substrate profiles, including acetonitrile, propionitrile, acrylonitrile, butyronitrile, succinonitrile, valeronitrile, isovaleronitrile and benzonitrile [115].

The nitrile hydratase enzymes of *R. rhodochrous* have to date been shown to be of the Co-type [6, 123, 125], which are usually more stable than the Fe-type nitrile hydratases. They have activity against a broad range of nitriles, including phenylacetonitrile, 2-phenylpropionitrile, 2-phenylglycinonitrile, mandelonitrile, 2-phenylbutyronitrile, 3-phenylpropionitrile, *N*-phenylglycinonitrile, *p*-toluonitrile and 3-hydroxy-3-phenylpropionitrile [32]. *R. ruber* CGMCC3090 and other strains express nitrile hydratases [115, 126] while the nitrile hydrolysis activity of *R. hoagii* [115] is also attributed to a nitrile hydratase [127].

The alternative nitrile hydrolysis enzyme, nitrilase, is also common in rhodococci (Table 2), including *R. erythropolis* [128], *R. rhodochrous* [129-132], *R. opacus* B4 [133] and *R. ruber* [134, 135]. The nitrilase from *R. ruber* can hydrolyse acetonitrile, acrylonitrile, succinonitrile, fumaronitrile, adiponitrile, 2-cyanopyridine, 3-cyanopyridine, indole-3-acetonitrile and mandelonitrile [135]. The nitrilases from multiple *R. erythropolis* strains were active towards phenylacetonitrile [136]. *R. rhodochrous* nitrilase substrates include (among many others) benzonitrile for *R. rhodochrous* J1 [137] and crotononitrile and acrylonitrile for *R. rhodochrous* K22 [138]. *R. rhodochrous* ATCC BAA-870 expresses an enantioselective aliphatic nitrilase encoded on the plasmid, which is induced by dimethylformamide [36]. Another nitrilase/cyanide hydratase family protein is also annotated on the plasmid (this study) but has not been characterised. The diverse, yet sometimes very specific and enantioselective substrate specificities of all these rhodococci gives rise to an almost plug-and-play system for many different synthetic applications. Combined with their high solvent tolerance, rhodococci are very well suited as biocatalysts to produce amides for both bulk chemicals and pharmaceutical ingredients.

The large percentage of possible mobile genomic region making up the plasmid, together with the high number of transposon genes and the fact that the plasmid contains the machinery for nitrile degradation, strongly support our theory that *R. rhodochrous* ATCC BAA-870 has adapted its genome recently in response to the selective pressure of routine culturing in nitrile media in the laboratory. Even though isolated from contaminated soil, the much larger chromosome of *R. jostii* RHA1 in comparison has undergone relatively little recent genetic flux as supported by the presence of only two intact insertion sequences, relatively few transposase genes, and only one identified pseudogene [10]. The smaller *R. rhodochrous* ATCC BAA-870 genome, still has the genetic space and tools to adapt relatively easily in response to environmental selection.

CRISPR

CRISPRs are unusual finds in rhodococcal genomes. Based on literature searches to date, only two other sequenced *Rhodococcus* strains were reported to contain potential CRISPRs. *R. opacus* strain M213, isolated from fuel-oil contaminated soil, has one confirmed and 14 potential CRISPRs [139], identified using the CRISPRFinder tool [140]. Pathak *et al*. also surveyed several other *Rhodococcus* sequences and found no other CRISPRs. Zhao and co-workers state that *Rhodococcus* strain sp. DSSKP-R-001, interesting for its beta-estradiol-degrading potential, contains 8 CRISPRs [141]. However, the authors do not state how these were identified. Pathak *et al*. highlight the possibility that the CRISPR in *R. opacus*

strain M213 may have been recruited from *R. opacus* R7 (isolated from polycyclic aromatic hydrocarbon contaminated soil [142]), based on matching BLASTs of the flanking regions.

The *R. rhodochrous* ATCC BAA-870 CRISPR upstream and downstream regions (based on a 270- and 718 nucleotide length BLAST, respectively) showed significant, but not matching, alignment with several other *Rhodococcus* strains. The region upstream of the BAA-870 CRISPR showed a maximum 95% identity with that from *R. rhodochrous* strains EP4 and NCTC10210, while the downstream region showed 97% identities to *R. pyridinovorans* strains GF3 and SB3094, *R. biphenylivorans* strain TG9, and *Rhodococcus* sp. P52 and 2G. Analysis by PHAST phage search tool [143] identified the presence of 6 potential, but incomplete, prophage regions on the chromosome, and one prophage region on the plasmid, suggesting that the CRISPR acquisition in *R. rhodochrous* ATCC BAA-870 could also have arisen from bacteriophage infection during its evolutionary history.

Identification of target genes for future biotechnology applications

An estimated 150 biocatalytic processes are currently being applied in industry [144-146]. The generally large and complex genomes of *Rhodococcus* species afford a wide range of genes attributed to extensive secondary metabolic pathways that are presumably responsible for an array of biotransformations and bioremediations. These secondary metabolic pathways have yet to be characterised and offer numerous targets for drug design as well as synthetic chemistry applications, especially since enzymes in secondary pathways are usually more promiscuous than enzymes in the primary pathways.

A number of potential genes which could be used for further biocatalyses have been identified in the genome of *R. rhodochrous* ATCC BAA-870. A substantial fraction of the genes have unknown functions, and these could be important reservoirs for novel gene and protein discovery. Most of the biocatalytically useful classes of enzyme suggested by Pollard and Woodley [147] are present on the genome: proteases, lipases, esterases, reductases, nitrilase/cyanohydrolase/nitrile hydratases and amidases, transaminase, epoxide hydrolase, monooxygenases and cytochrome P450s. Only oxynitrilases (hydroxynitrile lyases) and halohydrin dehalogenase were not detected, although a haloacid dehalogenase is present. Rhodococci are robust industrial biocatalysts, and the metabolic abilities of the *Rhodococcus* genus will continue to attract attention for industrial uses as further bio-degradative [6] and biopharmaceutical [148] applications of the organism are identified. Preventative and remediative biotechnologies will become increasingly popular as the demand for alternative means of curbing pollution increases and the need for new antimicrobial compounds and pharmaceuticals becomes a priority.

Conclusions

The genome sequence of *R. rhodochrous* ATCC BAA-870 is one of 353 *Rhodococcus* genomes that are sequenced to date, but it is only the 4th sequence that has been fully characterised on a biotechnological level. Therefore, the sequence of the *R. rhodochrous* ATCC BAA-870 genome will facilitate the further

exploitation of rhodococci for biotechnology applications, as well as enable further characterisation of a biotechnologically relevant organism. The genome has at least 1481 enzyme encoding genes, many of which have potential application in industrial biotechnology. Based on comparative annotation of the genome, up to 50% of annotated genes are hypothetical, while as much as 74% of genes may have unknown metabolic functions, indicating there is still a lot to learn about rhodococci.

Methods

Strain and culture conditions

R. rhodochrous ATCC BAA-870, isolated from industrial soil in Modderfontein, Johannesburg, South Africa, was grown routinely on Tryptone Soya Agar medium. For genomic DNA preparation, the strain was grown in 50 mL Tryptone Soya Broth overnight at 37°C. Cells were centrifuged and the DNA purified using a Wizard® Genomic DNA Purification Kit (Promega, Madison, WI) or Ultraclean microbial DNA extraction kit (MoBio, Carlsbad, CA). DNA concentrations were measured spectrophotometrically by absorbance readings at 260 nm using a NanoDrop-1000 (Thermo Scientific, Wilmington, DE).

Illumina sequencing

Genomic DNA of *R. rhodochrous* BAA-870 was used to obtain two libraries with different insert sizes. One 300 cycle paired-end library with insert-size of 550 bp was sequenced in-house on a MiSeq sequencer (Illumina, San Diego, CA) using TruSeq PCR-free library preparation. The second, a 50 cycle mate pair library with 5kb insert-size, was performed at BaseClear (Leiden, The Netherlands). Data is available at NCBI under Bioproject accession number PRJNA487734.

MinION sequencing

For Nanopore sequencing a 1D sequencing library (SQK-LSK108) was loaded onto a FLO-MIN106 (R9.4) flowcell, connected to the MinION Mk1B (Oxford Nanopore Technology, Oxford, United Kingdom). MinKNOW software (version 1.11.5; Oxford Nanopore) was used for quality control of active pores and for sequencing. Raw files generated by MinKNOW were base called, on a local compute server (HP ProLiant DL360 G9, 2x XEON E5-2695v3 14 Cores and 256 RAM), using Albacore (version 1.2.5; Oxford Nanopore). Reads, in fastq format, with a minimum length of 1000 bps were extracted, yielding 5.45 Gigabase sequence with an average read length of 9.09 kb.

De novo assembly

De novo assembly was performed using Canu (v1.4, settings: genomesize=6m) [149] producing a 5.88 Mbp genome consisting of two contigs. One chromosome with a length of 5.35 Mbp, while the second covers a size of 0.531 Mbp which, based on the Canu assembly graph, is a linear plasmid. The paired-end Illumina library was aligned, using BWA [150], to the assembly and the resulting Binary Alignment Map file was processed by Pilon [151] for polishing the assembly (correcting assembly errors), using correction of only SNPs and short indels (–fix bases parameter).

Annotation

The assembled genome sequence of *R. rhodochrous* ATCC BAA-870 was submitted to the Bacterial Annotation System web server, BASys, for automated, in-depth annotation of the chromosomal and plasmid sequences [50]. BASys annotates based on microbial *ab initio* gene prediction using GLIMMER [81]. The genome sequence was also run on the RAST (Rapid Annotation using Subsystem Technology) server using the default RASTtk annotation pipeline for comparison [152, 153]. RAST annotation uses the manually curated SEED database to infer gene annotations based on protein functional roles within families [154]. The two annotation pipelines offered different but useful and complimentary input formats and results, and gene annotations of interest could be manually compared and confirmed.

Abbreviations

ABC: ATP-Binding Cassette

antiSMASH: antibiotics and Secondary Metabolite Analysis Shell pipeline

BASys: Bacterial Annotation System

bps: base pairs

COG: Cluster of Orthologous Groups

contig: contiguous sequence

CRISPR: clustered regularly interspaced short palindromic repeat

EC: enzyme commission

GGDC: Genome-to-Genome Distance Calculator

Mbp: megabase pairs

MFS: Major Facilitator Superfamily

NCBI: National Center for Biotechnology Information

NRPS: nonribosomal peptide synthetase

ORF: open reading frame

PCBs: polychlorinated biphenyls

PKS: polyketide synthase

RAST: Rapid Annotation using Subsystem Technology

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The complete genome sequence of *R. rhodochrous* ATCC BAA 870 is deposited at NCBI GenBank, with Bioproject accession number PRJNA487734, and Biosample accession number SAMN09909133.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was partly funded by The Department of Science and Technology (DST) Biocatalysis Initiative (Grant 0175/2013), South Africa, the CSIR Thematic Programme, and the Welch Foundation (grant E-1264). Also the European Science Foundation COST Action CM0701 funded part of the project. The funding bodies had no role in the design of the study and no role in the collection, analysis, and interpretation of data or writing of the manuscript.

Authors' contributions

JF: principal investigator and primary author. FH: annotation of the preliminary genome and contribution to the manuscript preparation. UH: computational support for the assembly of the first scaffolds and evaluated data integrity. MB: genome assembly and annotation. PT and US: genome sequencing and data acquisition. RW, CH and JD: interpretation of sequence data. TS: academic supervisor of principal investigator. LO: interpretation of data, method strategy and major contribution to manuscript

preparation. DB: supervisor of principal investigator with major contribution to manuscript preparation. All authors read and approved the final manuscript.

Acknowledgements

We would like to acknowledge use of the University of Houston Institute for Molecular Design Sequencing Center, and Michael Benedik for useful discussions on this paper. We would also like to thank Dr Robert Gordon for support of nitrile biocatalysis research in South Africa and AECI for its support of Biotechnology.

References

1. van der Geize R, Dijkhuizen L. Harnessing the catabolic diversity of rhodococci for environmental and biotechnological applications. Curr Opin Microbiol. 2004;7(3):255-261.
2. Banerjee A, Sharma R, Banerjee UC. The nitrile-degrading enzymes: current status and future prospects. Appl Microbiol Biotechnol. 2002;60(1-2):33-44.
3. Yam KC, Geize R, Eltis LD. Catabolism of aromatic compounds and steroids by *Rhodococcus*. In: Alvarez HM, editors. Biology of Rhodococcus. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p.133-169.
4. Gray KA, Pogrebinsky OS, Mrachko GT, Xi L, Monticello DJ, Squires CH. Molecular mechanisms of biocatalytic desulfurization of fossil fuels. Nat Biotech. 1996;14(13):1705-1709.
5. Kobayashi M, Nagasawa T, Yamada H. Enzymatic synthesis of acrylamide: a success story not yet over. Trends Biotechnol. 1992;10(11):402-408.
6. Martínková L, Uhnáková B, Pátek M, Nešvera J, Křen V. Biodegradation potential of the genus *Rhodococcus*. Environ Int. 2009;35(1):162-177.
7. de Carvalho CC, Costa SS, Fernandes P, Couto I, Viveiros M. Membrane transport systems and the biodegradation potential and pathogenicity of genus *Rhodococcus*. Front Physiol. 2014;5.
8. Brady D. Biocatalytic hydrolysis of nitriles. In: Anastas PT, editors. Handbook of Green Chemistry. vol. 3: Wiley-VCH Verlag GmbH & Co. KGaA; 2010. p.27-49.
9. Sokolovská I, Rozenberg R, Riez C, Rouxhet PG, Agathos SN, Wattiau P. Carbon source-induced modifications in the mycolic acid content and cell wall permeability of *Rhodococcus erythropolis* E1. Appl Environ Microbiol. 2003;69(12):7019-7027.
10. McLeod MP, Warren RL, Hsiao WWL, Araki N, Myhre M, Fernandes C, Miyazawa D, Wong W, Lillquist AL, Wang D *et al*. The complete genome of *Rhodococcus* sp. RHA1 provides insights into a catabolic powerhouse. Proc Natl Acad Sci USA. 2006;103(42):15582-15587.
11. Seto M, Kimbara K, Shimura M, Hatta T, Fukuda M, Yano K. A novel transformation of polychlorinated biphenyls by *Rhodococcus* sp. strain RHA1. Appl Environ Microbiol. 1995;61(9):3353-3358.

12. Masai E, Yamada A, Healy JM, Hatta T, Kimbara K, Fukuda M, Yano K. Characterization of biphenyl catabolic genes of gram-positive polychlorinated biphenyl degrader *Rhodococcus* sp. strain RHA1. Appl Environ Microbiol. 1995;61(6):2079-2085.
13. Sangal V, Goodfellow M, Jones AL, Schwalbe EC, Blom J, Hoskisson PA, Sutcliffe IC. Next-generation systematics: An innovative approach to resolve the structure of complex prokaryotic taxa. Sci Rep. 2016;6:38392.
14. Sekine M, Tanikawa S, Omata S, Saito M, Fujisawa T, Tsukatani N, Tajima T, Sekigawa T, Kosugi H, Matsuo Y *et al*. Sequence analysis of three plasmids harboured in *Rhodococcus erythropolis* strain PR4. Environ Microbiol. 2006;8(2):334-346.
15. Komukai-Nakamura S, Sugiura K, Yamauchi-Inomata Y, Toki H, Venkateswaran K, Yamamoto S, Tanaka H, Harayama S. Construction of bacterial consortia that degrade Arabian light crude oil. J Ferment Bioeng. 1996;82(6):570-574.
16. Chen B-S, Otten LG, Resch V, Muyzer G, Hanefeld U. Draft genome sequence of *Rhodococcus rhodochrous* strain ATCC 17895. Stand Genomic Sci. 2013;9(1):175-184.
17. Holder JW, Ulrich JC, DeBono AC, Godfrey PA, Desjardins CA, Zucker J, Zeng Q, Leach ALB, Ghiviriga I, Dancel C *et al*. Comparative and functional genomics of *Rhodococcus opacus* PD630 for biofuels development. PLoS Genet. 2011;7(9):e1002219.
18. Chen Y, Ding Y, Yang L, Yu J, Liu G, Wang X, Zhang S, Yu D, Song L, Zhang H *et al*. Integrated omics study delineates the dynamics of lipid droplets in *Rhodococcus opacus* PD630. Nuc Acids Res. 2014;42(2):1052-1064.
19. Creason AL, Vandeputte OM, Savory EA, Davis EW, II, Putnam ML, Hu E, Swader-Hines D, Mol A, Baucher M, Prinsen E *et al*. Analysis of genome sequences from plant pathogenic *Rhodococcus* reveals genetic novelties in virulence loci. PLoS ONE. 2014;9(7):e101996.
20. Creason AL, Davis EW, Putnam ML, Vandeputte OM, Chang JH. Use of whole genome sequences to develop a molecular phylogenetic framework for *Rhodococcus fascians* and the *Rhodococcus* genus. Front Plant Sci. 2014;5(406).
21. Letek M, González P, MacArthur I, Rodríguez H, Freeman TC, Valero-Rello A, Blanco M, Buckley T, Cherevach I, Fahey R *et al*. The genome of a pathogenic *Rhodococcus*: Cooptive virulence underpinned by key gene acquisitions. PLoS Genet. 2010;6(9):e1001145.
22. Takai S, Hines SA, Sekizaki T, Nicholson VM, Alperin DA, Osaki M, Takamatsu D, Nakamura M, Suzuki K, Ogino N *et al*. DNA sequence and comparison of virulence plasmids from *Rhodococcus equi* ATCC 33701 and 103. Infect Immun. 2000;68(12):6840-6847.
23. Letek M, Ocampo-Sosa AA, Sanders M, Fogarty U, Buckley T, Leadon DP, González P, Scortti M, Meijer WG, Parkhill J *et al*. Evolution of the *Rhodococcus equi vap* pathogenicity island seen through comparison of host-associated *vapA* and *vapB* virulence plasmids. J Bacteriol. 2008;190(17):5797-5805.
24. Duquesne F, Hébert L, Sévin C, Breuil M-F, Tapprest J, Laugier C, Petry S. Analysis of plasmid diversity in 96 *Rhodococcus equi* strains isolated in Normandy (France) and sequencing of the 87-kb type I

virulence plasmid. FEMS Microbiol Lett. 2010;311(1):76-81.

25. Francis I, De Keyser A, De Backer P, Simón-Mateo C, Kalkus J, Pertry I, Ardiles-Diaz W, De Rycke R, Vandeputte OM, El Jaziri M *et al*. pFiD188, the linear virulence plasmid of *Rhodococcus fascians* D188. Mol Plant-Microbe Interact. 2012;25(5):637-647.
26. Lessard P, O'Brien X, Currie D, Sinskey A. pB264, a small, mobilizable, temperature sensitive plasmid from *Rhodococcus*. BMC Microbiol. 2004;4(1):15.
27. Matsui T, Saeki H, Shinzato N, Matsuda H. Analysis of the 7.6-kb cryptic plasmid pNC500 from *Rhodococcus rhodochrous* B-276 and construction of *Rhodococcus*–*E. coli* shuttle vector. Appl Microbiol Biotechnol. 2007;74(1):169-175.
28. Na K-s, Nagayasu K, Kuroda A, Takiguchi N, Ikeda T, Ohtake H, Kato J. Development of a genetic transformation system for benzene-tolerant *Rhodococcus opacus* strains. J Biosci Bioeng. 2005;99(4):408-414.
29. Nakashima N, Tamura T. Isolation and characterization of a rolling-circle-type plasmid from *Rhodococcus erythropolis* and application of the plasmid to multiple-recombinant-protein expression. Appl Environ Microbiol. 2004;70(9):5557-5568.
30. De Mot R, Nagy I, De Schrijver A, Pattanapipitpaisal P, Schoofs G, Vanderleyden J. Structural analysis of the 6 kb cryptic plasmid pFAJ2600 from *Rhodococcus erythropolis* NI86/21 and construction of *Escherichia coli-Rhodococcus* shuttle vectors. Microbiology 1997;143(10):3137-3147.
31. Stecker C, Johann A, Herzberg C, Averhoff B, Gottschalk G. Complete nucleotide sequence and genetic organization of the 210-kilobase linear plasmid of *Rhodococcus erythropolis* BD2. J Bacteriol. 2003;185(17):5269-5274.
32. Brady D, Beeton A, Zeevaart J, Kgaje C, van Rantwijk F, Sheldon RA. Characterisation of nitrilase and nitrile hydratase biocatalytic systems. Appl Microbiol Biotechnol. 2004;64(1):76-85.
33. Kinfe HH, Chhiba V, Frederick J, Bode ML, Mathiba K, Steenkamp PA, Brady D. Enantioselective hydrolysis of b-hydroxy nitriles using the whole cell biocatalyst *Rhodococcus rhodochrous* ATCC BAA-870. J Mol Catal B: Enzym. 2009;59(4):231-236.
34. Chhiba V, Bode ML, Mathiba K, Kwezi W, Brady D. Enantioselective biocatalytic hydrolysis of β-aminonitriles to β-amino-amides using *Rhodococcus rhodochrous* ATCC BAA-870. J Mol Catal B: Enzym. 2012;76(0):68-74.
35. Pawar SV, Yadav GD. Enantioselective enzymatic hydrolysis of *rac*-mandelonitrile to *R*-mandelamide by nitrile hydratase immobilized on poly(vinyl alcohol)/chitosan–glutaraldehyde support. Ind Eng Chem Res. 2014;53(19):7986-7991.
36. Chhiba-Govindjee VP, Mathiba K, van der Westhuyzen CW, Steenkamp P, Rashamuse JK, Stoychev S, Bode ML, Brady D. Dimethylformamide is a novel nitrilase inducer in *Rhodococcus rhodochrous*. Appl Microbiol Biotechnol. 2018;102(23):10055-10065.
37. Chen J, Zheng R-C, Zheng Y-G, Shen Y-C. Microbial transformation of nitriles to high-value acids or amides. In: Zhong J-J, Bai F-W, Zhang W, editors. Biotechnology in China I. vol. 113: Springer Berlin / Heidelberg; 2009. p.33-77.

38. Rodríguez JR. Understanding nitrile-degrading enzymes: classification, biocatalytic nature and current applications. Rev Latinoam Biotecnol Ambient Algal. 2014;5(1):1-18.
39. Martínková L, Stolz A, Rantwijk Fv, D'Antona N, Brady D, Otten LG. Nitrile converting enzymes involved in natural and synthetic cascade reactions. In: Riva S, Fessner W-D, editors. Cascade Biocatalysis. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA; 2014. p.249-270.
40. Chhiba V, Bode M, Mathiba K, Brady D. Enzymatic stereoselective synthesis of β-amino acids. In: Riva S, Fessner W-D, editors. Cascade Biocatalysis. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA; 2014. p.297-314.
41. Bigey F, Janbon G, Arnaud A, Galzy P. Sizing of the *Rhodococcus* sp. R312 genome by pulsed-field gel electrophoresis. Localization of genes involved in nitrile degradation. Antonie van Leeuwenhoek. 1995;68(2):173-179.
42. Stoddard SF, Schmidt TM, Hein R, Roller BRK, Smith BJ. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nuc Acids Res. 2014;43(D1):D593-D598.
43. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000;7(1-2):203-214.
44. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP, Lopez R. Analysis Tool Web Services from the EMBL-EBI. Nuc Acids Res. 2013;41(W1):W597-W600.
45. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, Park YM, Buso N, Lopez R. The EMBL-EBI bioinformatics web and programmatic tools framework. Nuc Acids Res. 2015;43(W1):W580-W584.
46. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7(1).
47. Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinform. 2013;14(1):60.
48. Meier-Kolthoff JP, Hahnke RL, Petersen J, Scheuner C, Michael V, Fiebig A, Rohde C, Rohde M, Fartmann B, Goodwin LA *et al*. Complete genome sequence of DSM 30083^T, the type strain (U5/41^T) of *Escherichia coli*, and a proposal for delineating subspecies in microbial taxonomy. Stand Genomic Sci. 2014;9(1):2.
49. Meier-Kolthoff JP, Klenk H-P, Göker M. Taxonomic use of DNA G+C content and DNA–DNA hybridization in the genomic age. Int J Syst Evol Microbiol. 2014;64(2):352-356.
50. Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, Lu P, Szafron D, Greiner R, Wishart DS. BASys: a web server for automated bacterial genome annotation. Nucl Acids Res. 2005;33(suppl 2):W455-W459.
51. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nuc Acids Res. 2000;28(1):33-36.

52. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997;278(5338):631-637.
53. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nuc Acids Res. 2015;43(D1):D261-D269.
54. Lowe TM, Eddy SR. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nuc Acids Res. 1997;25(5):955-964.
55. Chan PP, Lin B, Lowe TM. tRNAscan-SE 2.0. 2019.
56. Gardy JL, Brinkman FSL. Methods for predicting bacterial protein subcellular localization. Nat Rev Microbiol. 2006;4:741.
57. Pao SS, Paulsen IT, Saier MH. Major facilitator superfamily. Microbiol Mol Biol Rev. 1998;62(1):1-34.
58. Walmsley AR, Barrett MP, Bringaud F, Gould GW. Sugar transporters from bacteria, parasites and mammals: structure–activity relationships. Trends Biochem Sci. 1998;23(12):476-481.
59. Bassegoda A, Pastor FIJ, Diaz P. *Rhodococcus* sp. Strain CR-53 LipR, the first member of a new bacterial lipase family (Family X) displaying an unusual Y-type oxyanion hole, similar to the *Candida antarctica* lipase clan. Appl Environ Microbiol. 2012;78(6):1724-1732.
60. Zhang Y, Pan J, Luan Z-J, Xu G-C, Park S, Xu J-H. Cloning and characterization of a novel esterase from *Rhodococcus* sp. for highly enantioselective synthesis of a chiral cilastatin precursor. Appl Environ Microbiol. 2014;80(23):7348-7355.
61. Jiménez JI, Miñambres B, García JL, Díaz E. Genomic insights in the metabolism of aromatic compounds in *Pseudomonas*. In: Ramos J-L, editors. Pseudomonas. Springer US; 2004. p.425-462.
62. Yamada H, Kobayashi M. Nitrile hydratase and its application to industrial production of acrylamide. Biosci, Biotechnol, Biochem. 1996;60(9):1391-1400.
63. Nagasawa T, Takeuchi K, Yamada H. Occurrence of a cobalt-induced and cobalt-containing nitrile hydratase in *Rhodococcus rhodochrous* J1. Biochem Biophys Res Commun. 1988;155(2):1008-1016.
64. Thomas SM, DiCosimo R, Nagarajan V. Biocatalysis: applications and potentials for the chemical industry. Trends Biotechnol. 2002;20(6):238-242.
65. Ziemert N, Alanjary M, Weber T. The evolution of genome mining in microbes – a review. Nat Prod Rep. 2016;33(8):988-1005.
66. Crits-Christoph A, Diamond S, Butterfield CN, Thomas BC, Banfield JF. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. Nature. 2018;558(7710):440-444.
67. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nuc Acids Res. 2011;39(suppl_2):W339-W346.
68. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, Suarez Duran HG, de los Santos Emmanuel LC, Kim HU, Nave M *et al*. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. Nuc Acids Res. 2017;45(W1):W36-W41.

69. Yu Q, Schaub P, Ghisla S, Al-Babili S, Krieger-Liszkay A, Beyer P. The lycopene cyclase CrtY from *Pantoea ananatis* (formerly *Erwinia uredovora*) catalyzes an FAD_{red}-dependent non-redox reaction. J Biol Chem. 2010;285(16):12109-12120.
70. Oldfield E, Lin F-Y. Terpene biosynthesis: Modularity rules. Angew Chem, Int Ed. 2012;51(5):1124-1137.
71. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, Rocha EPC, Vergnaud G, Gautheret D, Pourcel C. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nuc Acids Res. 2018;46(W1):W246-W251.
72. Bertelli C, Laird MR, Williams KP, Group SFURC, Lau BY, Hoad G, Winsor GL, Brinkman FS. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nuc Acids Res. 2017;45(W1):W30-W35.
73. Szabó M, Kiss J, Olasz F. Functional organization of the inverted repeats of IS30. J Bacteriol. 2010;192(13):3414-3423.
74. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W *et al*. Environmental genome shotgun sequencing of the Sargasso sea. Science. 2004;304(5667):66-74.
75. Hughes Martiny JB, Field D. Ecological perspectives on the sequenced genome collection. Ecol Lett. 2005;8(12):1334-1345.
76. Warren A, Archuleta J, Feng W-c, Setubal J. Missing genes in the annotation of prokaryotic genomes. BMC Bioinf. 2010;11(1):131.
77. Galperin MY, Koonin EV. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. Nuc Acids Res. 2004;32(18):5452-5463.
78. Roberts RJ. Identifying protein function - A call for community action. PLoS Biol. 2004;2(3):e42.
79. Frishman D. Protein annotation at genomic scale: The current status. Chem Rev. 2007;107(8):3448-3466.
80. Mills CL, Beuning PJ, Ondrechen MJ. Biochemical functional predictions for protein structures of unknown or uncertain function. Comput Struct Biotechnol J. 2015;13:182-191.
81. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nuc Acids Res. 1999;27(23):4636-4641.
82. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nuc Acids Res. 2013;41(D1):D36-D42.
83. Harrison PW, Lower RPJ, Kim NKD, Young JPW. Introducing the bacterial 'chromid': not a chromosome, not a plasmid. Trends Microbiol. 2010;18(4):141-148.
84. Jiménez JI, Miñambres B, García JL, Díaz E. Genomic analysis of the aromatic catabolic pathways from *Pseudomonas putida* KT2440. Environ Microbiol. 2002;4(12):824-841.

85. Zampolli J, Zeaiter Z, Di Canito A, Di Gennaro P. Genome analysis and -omics approaches provide new insights into the biodegradation potential of *Rhodococcus*. Appl Microbiol Biotechnol. 2019;103(3):1069-1080.
86. Riebel A, de Gonzalo G, Fraaije MW. Expanding the biocatalytic toolbox of flavoprotein monooxygenases from *Rhodococcus jostii* RHA1. J Mol Catal B, Enzym. 2013;88:20-25.
87. Summers BD, Omar M, Ronson TO, Cartwright J, Lloyd M, Grogan G. *E. coli* cells expressing the Baeyer–Villiger monooxygenase 'MO14' (*ro*03437) from *Rhodococcus jostii* RHA1 catalyse the gram-scale resolution of a bicyclic ketone in a fermentor. Org Biomol Chem. 2015;13(6):1897-1903.
88. Van der Werf MJ. Purification and characterization of a Baeyer–Villiger mono-oxygenase from *Rhodococcus erythropolis* DCL14 involved in three different monocyclic monoterpene degradation pathways. Biochem J. 2000;347(3):693-701.
89. Rosłoniec KZ, Wilbrink MH, Capyk JK, Mohn WW, Ostendorf M, Van Der Geize R, Dijkhuizen L, Eltis LD. Cytochrome P450 125 (CYP125) catalyses C26-hydroxylation to initiate sterol side-chain degradation in *Rhodococcus jostii* RHA1. Mol Microbiol. 2009;74(5):1031-1043.
90. Grogan G. Cytochromes P450: exploiting diversity and enabling application as biocatalysts. Curr Opin Chem Biol. 2011;15(2):241-248.
91. Xiong F, Shuai J-J, Peng R-H, Tian Y-S, Zhao W, Yao Q-H, Xiong A-S. Expression, purification and functional characterization of a recombinant 2,3-dihydroxybiphenyl-1,2-dioxygenase from *Rhodococcus rhodochrous*. Mol Biol Rep. 2010;38(7):4303-4308.
92. Kuyukina MS, Ivshina IB, Serebrennikova MK, Krivoruchko AV, Korshunova IO, Peshkur TA, Cunningham CJ. Oilfield wastewater biotreatment in a fluidized-bed bioreactor using co-immobilized *Rhodococcus* cultures. J Environ Chem Eng. 2017;5(1):1252-1260.
93. Li S, Chaulagain MR, Knauff AR, Podust LM, Montgomery J, Sherman DH. Selective oxidation of carbolide C–H bonds by an engineered macrolide P450 mono-oxygenase. Proc Natl Acad Sci USA. 2009;106(44):18463-18468.
94. Bugg TDH, Ahmad M, Hardiman EM, Singh R. The emerging role for bacteria in lignin degradation and bio-product formation. Curr Opin Biotechnol. 2011;22(3):394-400.
95. Whyte LG, Smits THM, Labbé D, Witholt B, Greer CW, van Beilen JB. Gene cloning and characterization of multiple alkane hydroxylase systems in *Rhodococcus* strains Q15 and NRRL B-16531. Appl Environ Microbiol. 2002;68(12):5933-5942.
96. Leipold F, Rudroff F, Mihovilovic MD, Bornscheuer UT. The steroid monooxygenase from *Rhodococcus rhodochrous*; a versatile biocatalyst. Tetrahedron: Asymm. 2013;24(24):1620-1624.
97. Toda H, Ohuchi T, Imae R, Itoh N. Microbial production of aliphatic (*S*)-epoxyalkanes by using *Rhodococcus* sp. strain ST-10 styrene monooxygenase expressed in organic-solvent-tolerant *Kocuria rhizophila* DC2201. Appl Environ Microbiol. 2015;81(6):1919-1925.
98. Vignali E, Tonin F, Pollegioni L, Rosini E. Characterization and use of a bacterial lignin peroxidase with an improved manganese-oxidative activity. Appl Microbiol Biotechnol. 2018;102(24):10579-10588.

99. van Beilen JB, Smits THM, Whyte LG, Schorcht S, Röthlisberger M, Plaggemeier T, Engesser K-H, Witholt B. Alkane hydroxylase homologues in Gram-positive strains. Environ Microbiol. 2002;4(11):676-682.
100. Banerjee A. Stereoselective microbial Baeyer-Villiger oxidations. In: Patel RN, editors. Stereoselective Biocatalysis. CRC Press; 2000. p.867-876.
101. Kostichka K, Thomas SM, Gibson KJ, Nagarajan V, Cheng Q. Cloning and characterization of a gene cluster for cyclododecanone oxidation in *Rhodococcus ruber* SC1. J Bacteriol. 2001;183(21):6478-6486.
102. Schumacher DJ, Fakoussa MR. Degradation of alicyclic molecules by *Rhodococcus ruber* CD4. Appl Microbiol Biotechnol. 1999;52(1):85-90.
103. Fink MJ, Rudroff F, Mihovilovic MD. Baeyer–Villiger monooxygenases in aroma compound synthesis. Bioorg Med Chem Lett. 2011;21(20):6135-6138.
104. Galinski EA, Pfeiffer H-P, Trüper HG. 1,4,5,6-Tetrahydro-2-methyl-4-pyrimidinecarboxylic acid. A novel cyclic amino acid from halophilic phototrophic bacteria of the genus *Ectothiorhodospira*. Eur J Biochem. 1985;149(1):135-139.
105. Roberts MF. Organic compatible solutes of halotolerant and halophilic microorganisms. Saline Syst. 2005;1(1):5.
106. García-Estepa R, Argandoña M, Reina-Bueno M, Capote N, Iglesias-Guerra F, Nieto JJ, Vargas C. The *ectD* gene, which is involved in the synthesis of the compatible solute hydroxyectoine, is essential for thermoprotection of the halophilic bacterium *Chromohalobacter salexigens*. J Bacteriol. 2006;188(11):3774-3784.
107. Pastor JM, Salvador M, Argandoña M, Bernal V, Reina-Bueno M, Csonka LN, Iborra JL, Vargas C, Nieto JJ, Cánovas M. Ectoines in cell stress protection: Uses and biotechnological production. Biotechnol Adv. 2010;28(6):782-801.
108. Frings E, Sauer T, Galinski EA. Production of hydroxyectoine: high cell-density cultivation and osmotic downshock of *Marinococcus* strain M52. J Biotechnol. 1995;43(1):53-61.
109. Schiraldi C, Maresca C, Catapano A, Galinski EA, De Rosa M. High-yield cultivation of *Marinococcus* M52 for production and recovery of hydroxyectoine. Res Microbiol. 2006;157(7):693-699.
110. Tetali SD. Terpenes and isoprenoids: a wealth of compounds for global use. Planta. 2019;249(1):1-8.
111. Yamada Y, Kuzuyama T, Komatsu M, Shin-ya K, Omura S, Cane DE, Ikeda H. Terpene synthases are widely distributed in bacteria. Proc Natl Acad Sci USA. 2015;112(3):857-862.
112. Dickschat JS. Bacterial terpene cyclases. Nat Prod Rep. 2016;33(1):87-110.
113. Bosello M, Robbel L, Linne U, Xie X, Marahiel MA. Biosynthesis of the siderophore rhodochelin requires the coordinated expression of three independent gene clusters in *Rhodococcus jostii* RHA1. J Am Chem Soc. 2011;133(12):4587-4595.
114. Wang H, Fewer DP, Holm L, Rouhiainen L, Sivonen K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. Proc Natl Acad Sci

USA. 2014;111(25):9259-9264.

115. Brandão PFB, Bull AT. Nitrile hydrolysing activities of deep-sea and terrestrial mycolate actinomycetes. Antonie Van Leeuwenhoek. 2003;84(2):89-98.
116. Brady D, Dube N, Petersen R. Green chemistry: highly selective biocatalytic hydrolysis of nitrile compounds. S Afr J Sci. 2006;102(7-8):339-344.
117. Velankar H, Clarke KG, Preez Rd, Cowan DA, Burton SG. Developments in nitrile and amide biotransformation processes. Trends Biotechnol. 2010;28(11):561-569.
118. O'Reilly C, Turner PD. The nitrilase family of CN hydrolysing enzymes - a comparative study. J Appl Microbiol. 2003;95(6):1161-1174.
119. Gürtler V, Mayall BC, Seviour R. Can whole genome analysis refine the taxonomy of the genus *Rhodococcus*? FEMS Microbiol Rev. 2004;28(3):377-403.
120. Okamoto S, Eltis LD. Purification and characterization of a novel nitrile hydratase from *Rhodococcus* sp. RHA1. Mol Microbiol. 2007;65(3):828-838.
121. Coffey L, Owens E, Tambling K, O'Neill D, O'Connor L, O'Reilly C. Real-time PCR detection of Fe-type nitrile hydratase genes from environmental isolates suggests horizontal gene transfer between multiple genera. Antonie van Leeuwenhoek. 2010;98(4):455-463.
122. D'Antona N, Nicolosi G, Morrone R, Kubác D, Kaplan O, Martínková L. Synthesis of novel cyano-cyclitols and their stereoselective biotransformation catalyzed by *Rhodococcus erythropolis* A4. Tetrahedron: Asymmetry. 2010;21(6):695-702.
123. Precigou S, Goulas P, Duran R. Rapid and specific identification of nitrile hydratase (NHase)-encoding genes in soil samples by polymerase chain reaction. FEMS Microbiol Lett. 2001;204(1):155-161.
124. Brandão PFB, Clapp JP, Bull AT. Diversity of nitrile hydratase and amidase enzyme genes in *Rhodococcus erythropolis* recovered from geographically distinct habitats. Appl Environ Microbiol. 2003;69(10):5754-5766.
125. Kobayashi M, Nishiyama M, Nagasawa T, Horinouchi S, Beppu T, Yamada H. Cloning, nucleotide sequence and expression in *Escherichia coli* of two cobalt-containing nitrile hydratase genes from *Rhodococcus rhodochrous* J1. Biochim Biophys Acta, Gene Struct Expression. 1991;1129(1):23-33.
126. Shen Y, Wang M, Li X, Zhang J, Sun H, Luo J. Highly efficient synthesis of 5-cyanovaleramide by *Rhodococcus ruber* CGMCC3090 resting cells. J Chem Technol Biotechnol. 2012;87(10):1396-1400.
127. Kubáč D, Čejková A, Masák J, Jirků V, Lemaire M, Gallienne E, Bolte J, Stloukal R, Martínková L. Biotransformation of nitriles by *Rhodococcus equi* A4 immobilized in LentiKats®. J Mol Catal B: Enzym. 2006;39(1–4):59-61.
128. Vejvoda V, Šveda O, Kaplan O, Přikrylová V, Elišáková V, Himl M, Kubáč D, Pelantová H, Kuzma M, Křen V *et al*. Biotransformation of heterocyclic dinitriles by *Rhodococcus erythropolis* and fungal nitrilases. Biotechnol Lett. 2007;29(7):1119-1124.
129. Kobayashi M, Komeda H, Yanaka N, Nagasawa T, Yamada H. Nitrilase from *Rhodococcus rhodochrous* J1. Sequencing and overexpression of the gene and identification of an essential

cysteine residue. J Biol Chem. 1992;267(29):20746-20751.

130. Luo H, Fan L, Chang Y, Ma J, Yu H, Shen Z. Gene cloning, overexpression, and characterization of the nitrilase from *Rhodococcus rhodochrous* tg1-A6 in *E. coli*. Appl Biochem Biotechnol. 2010;160(2):393-400.
131. Yeom S-J, Kim H-J, Lee J-K, Kim D-E, Oh D-K. An amino acid at position 142 in nitrilase from *Rhodococcus rhodochrous* ATCC 33278 determines the substrate specificity for aliphatic and aromatic nitriles. Biochem J. 2008;415(3):401-407.
132. Thuku RN, Brady D, Benedik MJ, Sewell BT. Microbial nitrilases: versatile, spiral forming, industrial enzymes. J Appl Microbiol. 2009;106(3):703-727.
133. Cobzaru C, Ganas P, Mihasan M, Schleberger P, Brandsch R. Homologous gene clusters of nicotine catabolism, including a new ω-amidase for α-ketoglutaramate, in species of three genera of Gram-positive bacteria. Res Microbiol. 2011;162(3):285-291.
134. Webster NA, Ramsden DK, Hughes J. Comparative characterisation of two *Rhodococcus* species as potential biocatalysts for ammonium acrylate production. Biotechnol Lett.23(2):95-101.
135. Kamal A, Kumar MS, Kumar CG, Shaik TB. Bioconversion of acrylonitrile to acrylic acid by *Rhodococcus ruber* strain AKSH-84. J Microbiol Biotechnol. 2011;21(1):37-42.
136. Coffey L, Clarke A, Duggan P, Tambling K, Horgan S, Dowling D, O'Reilly C. Isolation of identical nitrilase genes from multiple bacterial strains and real-time PCR detection of the genes from soils provides evidence of horizontal gene transfer. Arch Microbiol. 2009;191(10):761-771.
137. Kobayashi M, Nagasawa T, Yamada H. Nitrilase of *Rhodococcus rhodochrous* J1. Purification and characterization. Eur J Biochem. 1989;182(2):349-356.
138. Kobayashi M, Yanaka N, Nagasawa T, Yamada H. Purification and characterization of a novel nitrilase of *Rhodococcus rhodochrous* K22 that acts on aliphatic nitriles. J Bacteriol. 1990;172(9):4807-4815.
139. Pathak A, Chauhan A, Blom J, Indest KJ, Jung CM, Stothard P, Bera G, Green SJ, Ogram A. Comparative genomics and metabolic analysis reveals peculiar characteristics of *Rhodococcus opacus* strain M213 particularly for naphthalene degradation. PLoS ONE. 2016;11(8):e0161032.
140. Grissa I, Pourcel C, Vergnaud G. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nuc Acids Res. 2007;35(suppl_2):W52-W57.
141. Zhao H, Tian K, Qiu Q, Wang Y, Zhang H, Ma S, Jin S, Huo H. Genome analysis of *Rhodococcus* sp. DSSKP-R-001: A highly effective b-estradiol-degrading bacterium. Int J Genomics. 2018;2018:11.
142. Di Gennaro P, Rescalli E, Galli E, Sello G, Bestetti G. Characterization of *Rhodococcus opacus* R7, a strain able to degrade naphthalene and o-xylene isolated from a polycyclic aromatic hydrocarbon-contaminated soil. Res Microbiol. 2001;152(7):641-651.
143. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. PHASTER: a better, faster version of the PHAST phage search tool. Nuc Acids Res. 2016;44(W1):W16-W21.

144. Zheng G-W, Xu J-H. New opportunities for biocatalysis: driving the synthesis of chiral chemicals. Curr Opin Biotechnol. 2011;22(6):784-792.
145. Panke S, Wubbolts M. Advances in biocatalytic synthesis of pharmaceutical intermediates. Curr Opin Chem Biol. 2005;9(2):188-194.
146. Jemli S, Ayadi-Zouari D, Hlima HB, Bejar S. Biocatalysts: application and engineering for industrial purposes. Crit Rev Biotechnol. 2016;36(2):246-258.
147. Pollard DJ, Woodley JM. Biocatalysis for pharmaceutical intermediates: the future is now. Trends Biotechnol. 2007;25(2):66-73.
148. Yam KC, Okamoto S, Roberts JN, Eltis LD. Adventures in *Rhodococcus* - from steroids to explosives. Can J Microbiol. 2011;57(3):155-168.
149. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res. 2017;27(5):722-736.
150. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26(5):589-595.
151. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK *et al*. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE. 2014;9(11):e112963.
152. Aziz R, Bartels D, Best A, DeJongh M, Disz T, Edwards R, Formsma K, Gerdes S, Glass E, Kubal M *et al*. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. 2008;9(1):75.
153. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD *et al*. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep. 2015;5:8365.
154. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M *et al*. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nuc Acids Res. 2014;42(D1):D206-D214.
155. Kwasiborski A, Mondy S, Chong T-M, Chan K-G, Beury-Cirou A, Faure D. Core genome and plasmidome of the quorum-quenching bacterium *Rhodococcus erythropolis*. Genetica. 2015;143(2):253-261.
156. Novikov AD, Lavrov KV, Kasianov AS, Gerasimova TV, Yanenko AS. Draft genome sequence of *Rhodococcus* sp. strain M8, which can degrade a broad range of nitriles. Genome Announc. 2018;6(6):e01526-01517.
157. Yamaguchi T, Asano Y. Draft genome sequence of an aldoxime degrader, *Rhodococcus* sp. strain YH3-3. Genome Announc. 2016;4(3):e00406-00416.

Tables

Table 1: Fully sequenced[1] and well described *Rhodococcus* species ranked by completion date

Organism	Date Completed[2]	Group	Reference	Chromosome (Mbp)	Plasmid (Mbp)	Total Size, Mbp	G + C %	Protein coding genes
R. rhodochrous ATCC BAA-870	2018	This study	This paper	5.37	0.53	5.9	65	7548*
R. erythropolis R138	19-03-2013	Centre National de la Recherche Scientifique, Institut des Sciences du Vegetal, France	NZ_CP007255 [155]	6.2	477,915; 91,729	6.8	62	6130
R. opacus PD630[3]	26-11-2012	National Laboratory of Macromolecules, Chinese Academy of Sciences, Beijing	NZ_CP003949 [18]	8.38	9 plasmids	9.17	67	8947
R. opacus PD630[3]	10-11-2011	Massachusetts Institute of Technology and The Broad Institute	GCF_000234335 [17]	-	-	9.27	67	7910
R. hoagii 103S[4]	21-10-2009	IREC (International *Rhodococcus equi* Genome Consortium)	NC_014659 [21]	5.04	None determined	5.04	69	4540
R. jostii RHA1	24-07-2006	Genome British Columbia, Vancouver	NC_008268 [10]	7.8	1,123,075; 442,536; 332,361	9.7	67	8690
R. erythropolis PR4	31-03-2005	Sequencing Center: National Institute of Technology and Evaluation, Japan	NC_012490 [14]	6.5	271,577; 104,014; 3,637	6.9	62	6321

[1] All sequences are completed and fully assembled, except GCF_000234335, which consists of 282 contigs.

[2] Date completed refers to genome sequence completion/submission to database; plasmids may have been completed at another time. Total genome size comprises the chromosome and the plasmid sequence. Genome information of strains other than BAA-870 is obtained from the NCBI database.

[3] Two separate references, therefore 2 entries.

[4] *R. equi* is renamed to *R. hoagii*

* Based on BASys annotation.

Table 2. Comparison of nitrile converting enzymes in different *Rhodococcus* species

Number of enzymes on the chromosome. If multiple enzymes are present on different genomic elements, the location is mentioned: chromosome (chr) or plasmids (pl).

Organism	Nitrilase	Nitrile Hydratase	NHase regulators	Amidase	Amidase Regulators	NCBI Assembly Reference
R. rhodochrous ATCC BAA-870	2 (pl)	1 (pl)	4 (pl)	7 (chr) 2 (pl)	2 (pl)	this study
R. erythropolis PR4	-	1	4	12	1	GCF_000010105 [14]
R. erythropolis SK121	-	1	-	2	-	GCF_000174835 (no reference)
R. hoagii 103S	-	-	-	11	-	GCF_000196695 [21]
R. hoagii ATCC 33707	-	-	-	11	-	GCF_000164155 (no reference)
R. jostii RHA1	1	1	-	14 (chr) 1 (pl)	-	GCF_000014565 [10]
R. opacus B4	-	-	-	13	-	GCF_000010805 (no reference)
R. opacus PD630	-	2	-	13	2	GCF_000599555 GCF_000234335 [17, 18]
Rhodococcus sp. M8	-	2	1	9	-	GCF_001890475 [156]
Rhodococcus sp. YH3-3	1	2	1	13	1	GCF_001653035 [157]

Additional File Legends

SuppInfo Frederick et al BAA-870 genome.

Supplementary tables. Table S1: All sequenced *Rhodococcus* strains (353) according to the NCBI database (accessed 13/03/2019). Table S2: All complete sequenced *Rhodococcus* species ranked by release date according to the NCBI Genome database (accessed 11/03/2019). Table S3: Whole genome distance statistics between *Rhodococcus rhodochrous* ATCC BAA-870 and two closely matched strains. Table S4: *Rhodococcus rhodochrous* ATCC BAA-870 protein function breakdown based on BASys annotation COG classifications

Figures

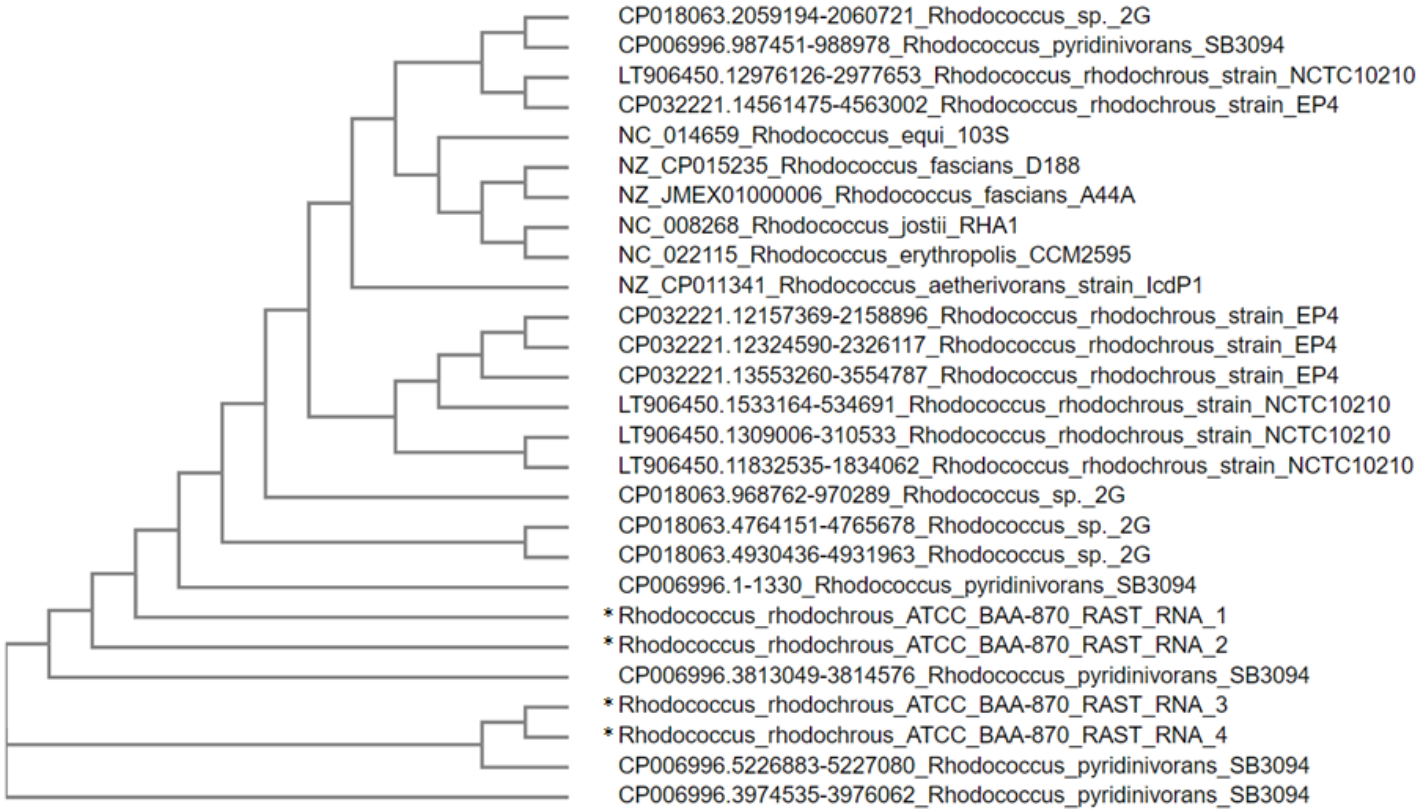

Figure 1

Phylogenetic tree created using rhodococcal 16S rRNA ClustalW sequence alignments. Neighbour joining, phylogenetic cladogram created using Phylogeny in ClustalW and and ClustalO multiple sequence alignment of R. rhodochrous ATCC BAA-870 16S rRNA genes and other closely matched genes from rhodococcal species. R. rhodochrous ATCC BAA-870 contains four copies of the 16S rRNA gene (labelled RNA_1 to RNA_4) and are indicated with an asterisk. For clarity, only closely matched BLAST results with greater than 95.5% sequence identity and those with complete 16S rRNA gene sequences, or from complete genomes, are considered. Additionally, 16S rRNA gene sequences (obtained from the NCBI gene database) from R. jostii RHA1, R. fascians A44A and D188, R. equi 103S, R. erythropolis CCM2595, and R. aetherivorans strain IcdP1 are included for comparison. Strain names are preceded by their NCBI accession number, as well as sequence position if there are multiple copies of the 16S rRNA gene in the same species.

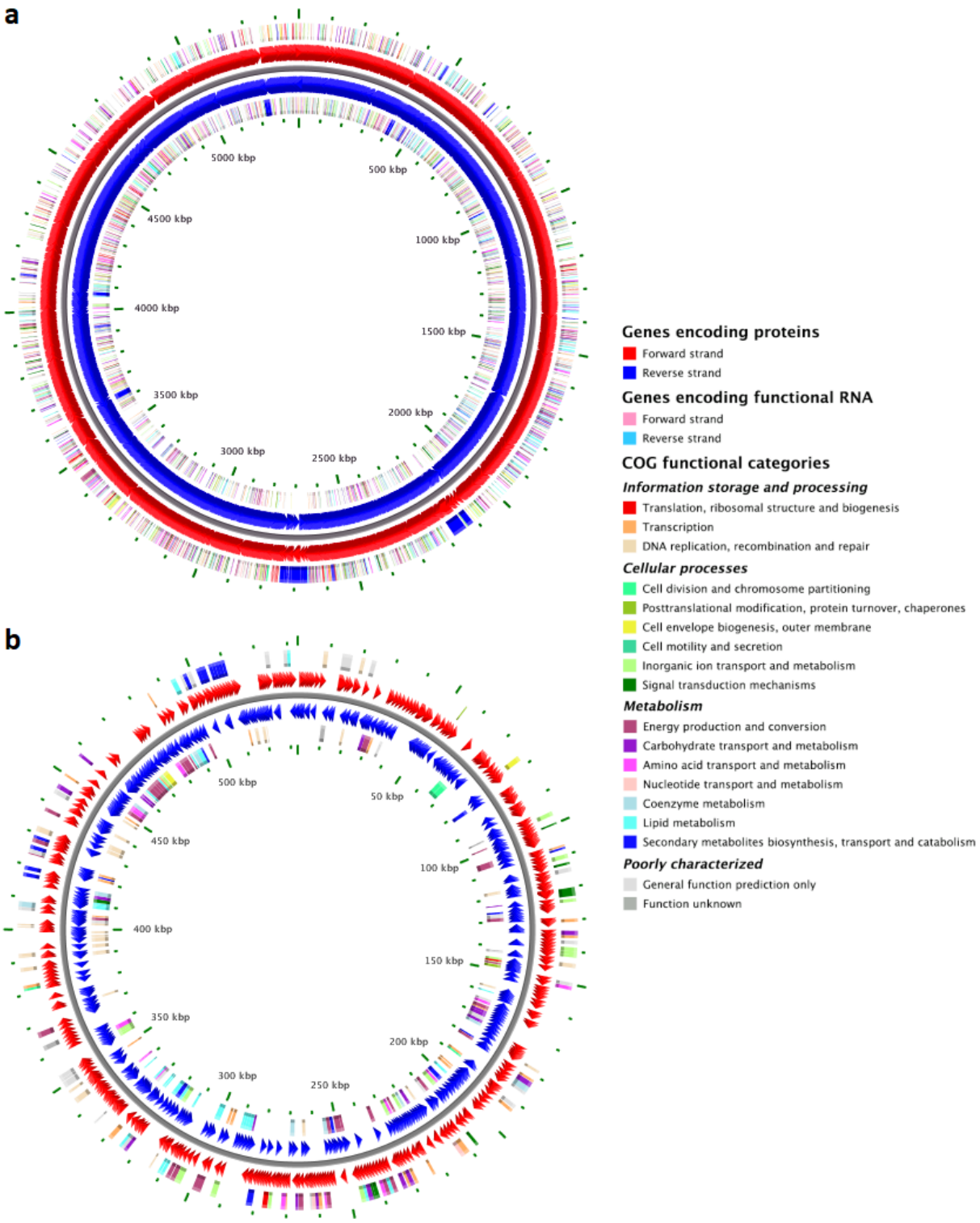

Figure 2

BASys bacterial annotation summary view of the Rhodococcus rhodochrous ATCC BAA-870 genome. BASys visual representation of (a) the 5,370,537 bp chromosome, with a breakdown of the 6871 genes encoded, and (b) the 533,288 bp linear plasmid, with a breakdown of the 677 genes encoded. Different colours indicate different subsystems for catabolic and anabolic routes.

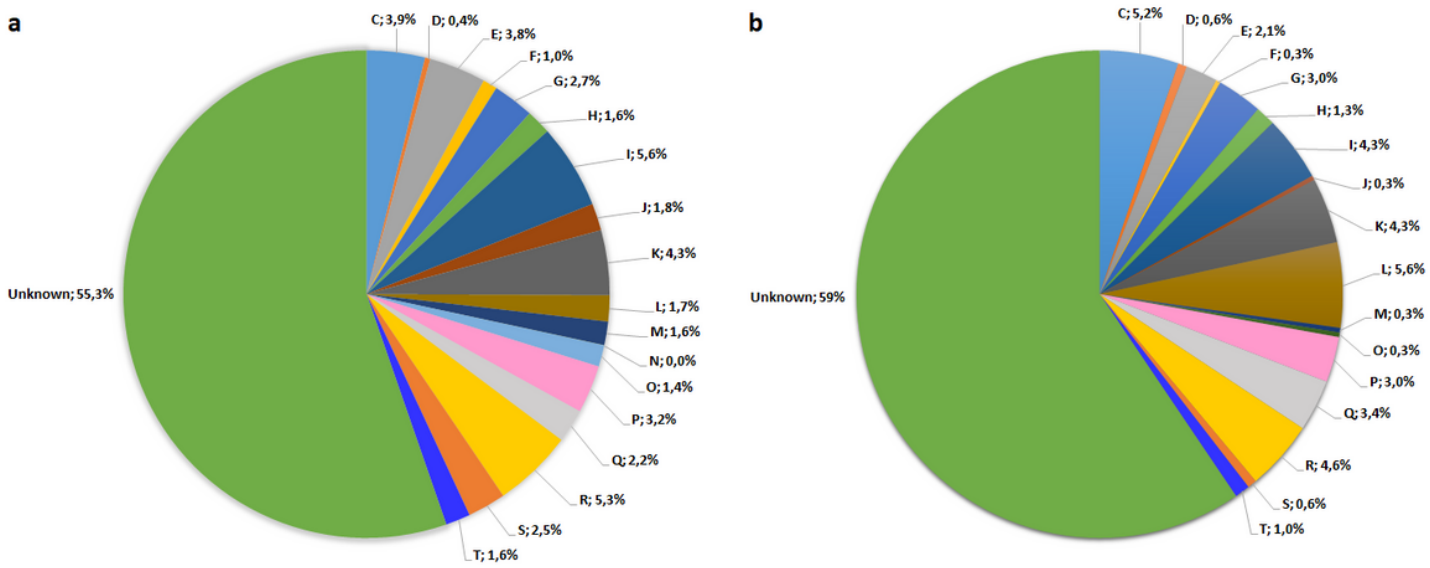

Figure 3

Protein function breakdown of Rhodococcus rhodochrous ATCC BAA-870 based on BASys annotation COG classifications. Unknown proteins form the majority of proteins in the BASys annotated genome, and make up 55 and 59 % respectively of genes in the (a) chromosome and (b) plasmid. For simplicity, functional categories less than 0.02% are not included in the graphic. Letters refer to COG functional categories, with one-letter abbreviations: C - Energy production and conversion; D - Cell division and chromosome partitioning; E - Amino acid transport and metabolism; F - Nucleotide transport and metabolism; G - Carbohydrate transport and metabolism; H - Coenzyme metabolism; I - Lipid metabolism; J - Translation, ribosomal structure and biogenesis; K - Transcription; L - DNA replication, recombination and repair; M - Cell envelope biogenesis, outer membrane; N - Secretion, motility and chemotaxis; O - Posttranslational modification, protein turnover, chaperones; P - Inorganic ion transport and metabolism; Q - Secondary metabolites biosynthesis, transport and catabolism; R - General function prediction only; S - COG of unknown function; T - Signal transduction mechanisms.

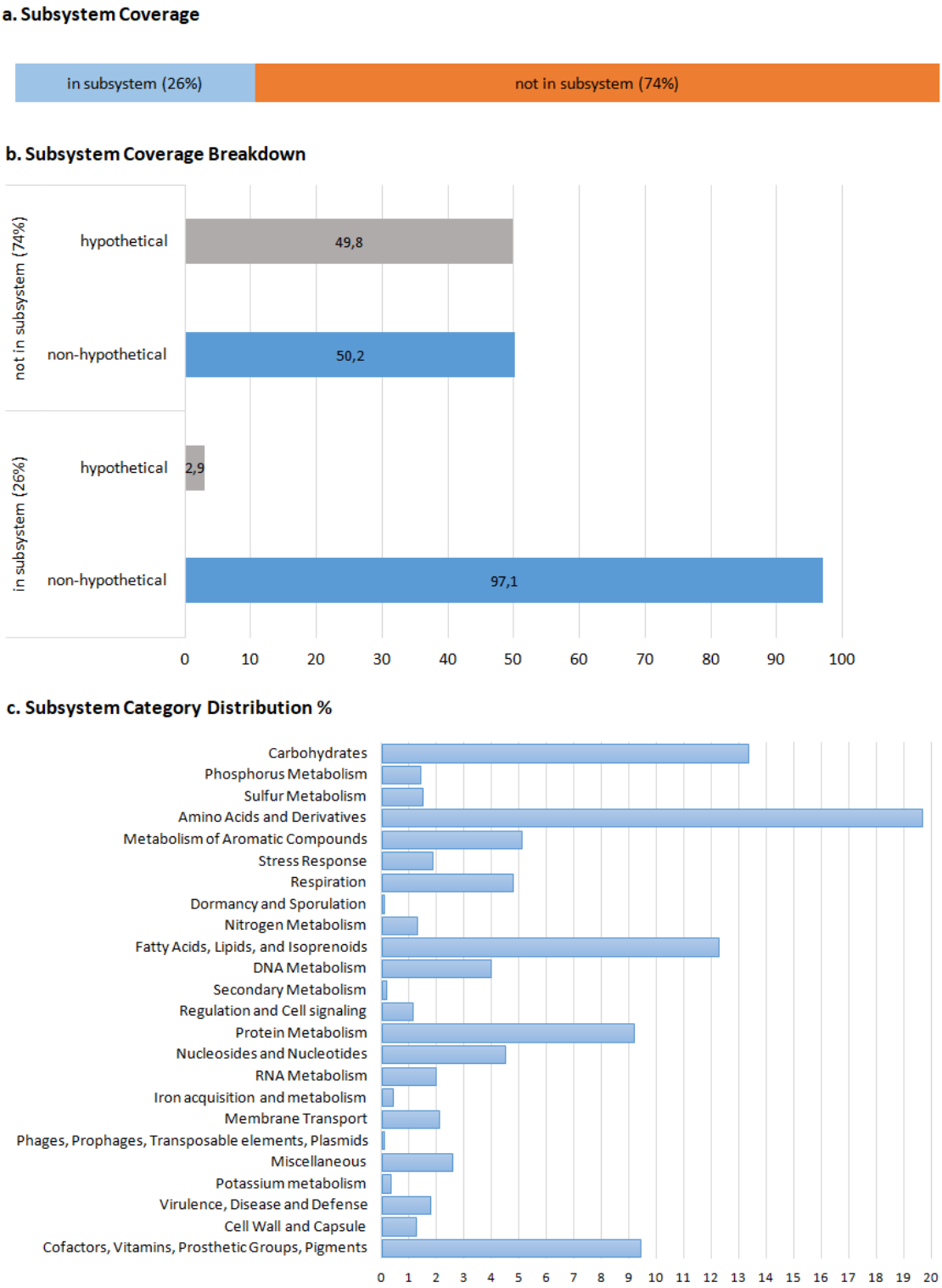

Figure 4

RAST annotation summary of the Rhodococcus rhodochrous ATCC BAA-870 genome. RAST annotation results show (a) the subsystem coverage, (b) the subsystem coverage breakdown, and (c) organisation of the subsystems by cellular process as a percentage showing the distribution of annotations across defined structural and functional subsystem roles. RAST uses a subsystem approach, in which annotations are assigned to groups with similar functional or structural roles. For R. rhodochrous ATCC

BAA-870, 26% of annotated genes belong to an identified functional role, or subsystem. The coverage breakdown shows the percentage of hypothetical and non-hypothetical annotations for genes assigned to subsystems and those for which a known functional role is not assigned (i.e. those not in the subsystem).