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

Joni Frederick1,2,3, Fritha Hennessy1, Uli Horn4, Pilar de la Torre Cortés5, Marcel van den Broek5, Ulrich Strych6,7, Richard Willson6,8, Charles A. Hefer9,10, Jean-Marc G. Daran5, Trevor Sewell2, Linda G. Otten11*, and Dean Brady1,12*

**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, monoxygenases, 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, 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 (γ-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 (Additional file 1: 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) [18] which degrades long alkanes [19]. 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_ASIJ01000002) [20] also has many monooxygenases, 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 [16]. 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) [15], 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 [21], which also resulted in the realisation that sequencing provides additional means to traditional ways of determining

### Table 1

| Organism          | Date Completed | Group                          | Reference                                      | Chromosome (Mbp) | Plasmid (Mbp) | Total Size, Mbp | G+C % | Protein coding genes |
|-------------------|----------------|--------------------------------|------------------------------------------------|------------------|---------------|------------------|-------|----------------------|
| *R. rhodochrous*  | 2018           | This study                     | This paper                                     | 5.37             | 0.53          | 5.9              | 66    | 7548                 |
| ATCC BAA-870      |                |                                |                                                |                  |               |                  |       |                      |
| *R. erythropolis* | 19-03-2013     | Centre National de la Recherche Scientifique, Institut des Sciences du Vegetal, France | NZ_CP007255 [14] | 6.2             | 477,915; 91,729 | 6.8              | 62    | 6130                 |
| R138              |                |                                |                                                |                  |               |                  |       |                      |
| *R. opacus*       | 26-11-2012     | National Laboratory of Macromolecules, Chinese Academy of Sciences, Beijing | NZ_CP003949 [15] | 8.38            | 9 plasmids    | 9.17             | 67    | 8947                 |
| PD630*            |                |                                |                                                |                  |               |                  |       |                      |
| *R. opacus*       | 10-11-2011     | Massachusetts Institute of Technology and The Broad Institute | GCF_000234335 [16] | –               | –             | 9.27             | 67    | 7910                 |
| PD630*            |                |                                |                                                |                  |               |                  |       |                      |
| *R. haagii*       | 21-10-2009     | IREC (International Rhodococcus *equi* Genome Consortium) | NC_014659 [17] | 5.04            | None determined | 5.04             | 69    | 4540                 |
| 1035*             |                |                                |                                                |                  |               |                  |       |                      |
| *R. jostii*       | 24-07-2006     | Genome British Columbia, Vancouver | NC_008268 [10] | 7.8             | 1,123,075; 442,536; 332,361 | 9.7              | 67    | 8690                 |
| RHA1              |                |                                |                                                |                  |               |                  |       |                      |
| *R. erythropolis* | 31-03-2005     | Sequencing Center: National Institute of Technology and Evaluation, Japan | NC_012490 [18] | 6.5             | 271,577; 104,014; 3637 | 6.9              | 62    | 6321                 |
| PR4               |                |                                |                                                |                  |               |                  |       |                      |

*All sequences are completed and fully assembled, except GCF_000234335, which consists of 282 contigs

*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

*Two separate references, therefore 2 entries

*R. equi is renamed to R. haagii

*Based on BASys annotation
speciation in the very diverse genus of *Rhodococcus* [22]. 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) [17]. 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 [23], strain PAM1593 [24] and 96 strains isolated from Normandy (France) [25]. As many important traits are often located on (easily transferable) plasmids, numerous rhodococcal plasmid sequences have been submitted to the NCBI (Additional file 1: Table S2). More elaborate research has been published on the virulence plasmid pFiD188 from *R. fascians* D188 [26], pB264, a cryptic plasmid from *Rhodococcus* sp. B264–1 [27], pNC500 from *R. rhodochrous* B-276 [28], and several plasmids from *R. opacus* B4 [29] and PD630 [15]. *R. erythropolis* harbours many plasmids besides the three from strain PR4, including pRE8424 from strain DSM8424 [30], pFAJ2600 from NL86/21 [31] and pBD2 from strain BD2 [32]. 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 [33], 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 [33–36]. 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 [37]. Biocatalytic nitrile hydrolysis affords valuable applications in industry, including production of solvents, extractants, pharmaceuticals, drug intermediates, and pesticides [38–41]. 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 [42]. 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 to 5 copies, with an average of 4 [43]. 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 [44]. 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 [45–47] (Fig. 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 [48–50], 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 (Additional file 1: 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 [51]. 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 (Fig. 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 (Fig. 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 [52]. As each COG consists of individual proteins or groups of paralogs from at least 3 lineages, it corresponds to an ancient conserved domain [53, 54]. 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 (Fig. 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
Fig. 2 (See legend on next page.)

Genes encoding proteins
- Red: Forward strand
- Blue: Reverse strand

Genes encoding functional RNA
- Pink: Forward strand
- Green: Reverse strand

COG functional categories

Information storage and processing
- Translation, ribosomal structure and biogenesis
- Transcription
- DNA replication, recombination and repair

Cellular processes
- Cell division and chromosome partitioning
- Posttranslational modification, protein turnover, chaperones
- Cell envelope biogenesis, outer membrane
- Cell motility and secretion
- Inorganic ion transport and metabolism
- Signal transduction mechanisms

Metabolism
- Energy production and conversion
- Carbohydrate transport and metabolism
- Amino acid transport and metabolism
- Nucleotide transport and metabolism
- Coenzyme metabolism
- Lipid metabolism
- Secondary metabolites biosynthesis, transport and catabolism

Poorly characterized
- General function prediction only
- Function unknown
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 functional 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 (Fig. 4), as well as the COG classification breakdowns provided by BASys (Fig. 3 and Additional file 1: 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^Met_. Additional analysis of
**Fig. 4** (See legend on next page.)

### a. Subsystem Coverage

- In subsystem (26%)
- Not in subsystem (74%)

### b. Subsystem Coverage Breakdown

- Hypothetical: 49.8%
- Non-hypothetical: 50.2%

### c. Subsystem Category Distribution %

- Carbohydrates
- Phosphorus Metabolism
- Sulfur Metabolism
- Amino Acids and Derivatives
- Metabolism of Aromatic Compounds
- Stress Response
- Respiration
- Dormancy and Sporulation
- Nitrogen Metabolism
- Fatty Acids, Lipids, and Isoprenoids
- DNA Metabolism
- Secondary Metabolism
- Regulation and Cell signaling
- Protein Metabolism
- Nucleosides and Nucleotides
- RNA Metabolism
- Iron Acquisition and metabolism
- Membrane Transport
- Phages, Prophages, Transposable elements, Plasmids
- Miscellaneous
- Potassium metabolism
- Virulence, Disease and Defense
- Cell Wall and Capsule
- Cofactors, Vitamins, Prosthetic Groups, Pigments

---

*Fig. 4* (See legend on next page.)
the genome sequence using the tRNA finding tool tRNAscan-SE v. 2.0 [55, 56] 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 [57], 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 [51]. 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 chemotactic 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 [58, 59]. 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 [60, 61] are, however, present in this strain.

Aromatic catabolism and oxidoreductases
As deduced from the better characterized pseudomonads [62], 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 peroxyreductins. Furthermore, there are 71 monooxy-genase 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 monoxygenases (EC 1.14.13.16) and a phenol monooxygenase (EC 1.14.13.7) on the plasmid, a
methane monoxygenase (EC 1.14.13.25), two alkane monoxygenases (EC 1.14.15.3) and five phenylacetone monoxygenases (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 [63], and both Mitsubishi Rayon Co., Ltd. (Japan) and Semmin (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 [64]. Both processes rely on rhodococcal nitrile hydratase activity [65].

As *R. rhodochrous* ATCC BAA-870 was isolated from a nitrile enrichment culture [33], 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 [37]. 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 [66], and the secondary metabolism of diverse soil-dwelling microbes remains relatively unexplored despite their huge biosynthetic potential [67]. 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) [68, 69]. Biosynthetic gene clusters identified in *R. rhodochrous* BAA-870 include ectoine (1,4,5,6-tetrahydro-2-methyl-4-pyrimidinocarboxylic 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 phytone saturase and several phytene 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 [70], farnesyl diphosphate synthase, farnesyl transferase, geranylgeranyl pyrophosphate synthetases and digeranylgeranylgeranylpyrophosphate reductase. Farnesyl diphosphate synthase and geranylgeranyl pyrophosphate synthases are potential anticancer and anti-infective drug targets [71]. 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 dimers. 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 [72]. 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 *ruvC* 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) [73] 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 [74]. 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 [75, 76]. 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 (Additional file 1: 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 [77]. 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 [77]. 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 [78–81]. 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 (Fig. 2) and RAST (Fig. 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 [82]. 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 [83] and show a large variation in genome size between 4 and 10 Mbp (Additional file 1: 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. haogii genomes are the smallest at ~ 5 Mbp. All rhodococcal genomes have a high GC content, ranging from 62 to 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 [84].

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 [85]. 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 [86]. 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 [87], which includes monoxygenases capable of catalysing Baeyer–Villiger oxidations wherein a ketone is converted to an ester [88, 89].

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 [90–93]. Cytochrome P450 oxygenase is often found fused with a reductase, as in Rhodococcus sp. NCIMB 9784 [94]. 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 [95].

The oxygenases found in rhodococci include multiple alkane monoxygenases (genes alkB1–alkB4) [96], steroid monoxygenase [97], styrene monoxygenase [98], peroxidase [99] and alkane hydroxylase homologs [100]. R. rhodochrous ATCC BAA-870 has 87 oxygenase genes while the PCB degrading R. jostii RHA1 has 203 oxygenases, including 19 cyclohexanone monoxygenases (EC 1.14.13.22), implying that of the two, strain BAA-870 is less adept at oxidative catabolism. Rhodococcal cyclohexanone monoxygenases 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) [65]. Chiral lactones can also be used as intermediates in the production of prostaglandins [101]. The same oxidative pathway can be used to biotransform cyclohexadecanone to lauryl lactone or 12-hydroxydodecanoic acid [102, 103]. Cyclohexadecanone monoxygenase 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 [104]. Similar to R. jostii RHA1, R.
rhodochrous ATCC BAA-870 encodes several monoxygenases. 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 [105, 106], and hydroxyectoine has been shown to confer heat stress protection in vivo [107]. Ectoines provide a variety of useful biotechnological and biomedical applications [108], and strains engineered for improved ectoine synthesis have been used for the industrial production of hydroxyectoine as a solute and enzyme stabiliser [109, 110]. 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 [71, 111]. However, it is only recently that bacterial terpenoids have been considered as a possible source of new natural product wealth [112, 113], 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 terpene 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 [112]. 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 [114]. 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 [115].

### Nitrile conversion

Many rhodococci can hydrolyse a wide range of nitriles [116–119]. 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).

| Organism                        | Nitrilase | Nitrile Hydratase | Ntase regulators | Amidase | Amidase Regulators | NCBI Assembly Reference    |
|---------------------------------|-----------|-------------------|-----------------|---------|-------------------|---------------------------|
| *R. rhodochrous* ATCC BAA-870   | 2 (pl)    | 1 (pl)            | 4 (pl)          | 7 (chr) | 2 (pl)            | this study                |
| *R. erythropolis* PR4           | –         | 1                 | 4               | 12      | 1                 | GCF_000010105 [18]        |
| *R. erythropolis* SK121         | –         | 1                 | –               | 2       | –                 | GCF_000174835 (no reference) |
| *R. hoagii* 103S                | –         | –                 | –               | 11      | –                 | GCF_000196695 [17]        |
| *R. hoagii* ATCC 33707          | –         | –                 | –               | 11      | –                 | GCF_000164155 (no reference) |
| *R. jostii* RHA1                | 1         | 1                 | –               | 14 (chr)| –                 | GCF_000014565 [10]        |
| *R. opacus* B4                  | –         | –                 | –               | 13      | –                 | GCF_000010805 (no reference) |
| *R. opacus* PD630               | –         | 2                 | –               | 13      | 2                 | GCF_000599555 GCF_000234335 [15, 16] |
| *Rhodococcus* sp. MB            | –         | 2                 | 1               | 9       | –                 | GCF_001890475 [120]       |
| *Rhodococcus* sp. YH3–3         | 1         | 2                 | 1               | 13      | 1                 | GCF_001653035 [121]       |

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

---

**Table 2** Comparison of nitrile converting enzymes in different *Rhodococcus* species.
ATCC BAA-870 contains several nitrile converting enzymes which is in line with previous activity assays using this *Rhodococcus* strain [34, 35]. 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 [37]. 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. (Fig. 1) [122]. *R. jostii* RHA1 expresses a nitrile hydratase (an acetonitrile hydratase) and utilises nitriles such as acetonitrile, acrylonitrile, propionitrile and butyronitrile [123], while *R. opacus* expresses nitrile hydrolysis activity [116]. *R. erythropolis* PR4 expresses a Fe-type nitrile hydratase [124], and *R. erythropolis* strains are well known for expressing this enzyme [116, 125, 126] as part of a nitrile metabolism gene cluster [122]. This enzyme has been repeatedly determined in this species from isolated diverse locations [127], expressing broad substrate profiles, including acetonitrile, propionitrile, acrylonitrile, butyronitrile, succinonitrile, valeronitrile, isovaleronitrile and benzonitrile [116].

The nitrile hydratase enzymes of *R. rhodochrous* have to date been shown to be of the Co-type [6, 126, 128], 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-toluenesulphonitrile and 3-hydroxy-3-phenylpropionitrile [33]. *R. ruber* CGMCC3090 and other strains express nitrile hydratases [116, 129] while the nitrile hydrolysis activity of *R. haangi* [116] is also attributed to a nitrile hydratase [130].

The alternative nitrile hydrolysis enzyme, nitrilase, is also common in rhodococci (Table 2), including *R. erythropolis* [131], *R. rhodochrous* [132–135], *R. opacus* B4 [136] and *R. ruber* [137, 138]. The nitrilase from *R. ruber* can hydrolyse acetonitrile, acrylonitrile, succinonitrile, fumaronitrile, adiponitrile, 2-cyanopropidine, 3-cyanopropyridine, indole-3-acetonitrile and mandelonitrile [138]. The nitrilases from multiple *R. erythropolis* strains were active towards phenylacetonitrile [139]. *R. rhodochrous* nitrilase substrates include (among many others) benzonitrile for *R. rhodochrous* J1 [140] and crotononitrile and acrylonitrile for *R. rhodochrous* K22 [141]. *R. rhodochrous* ATCC BAA-870 expresses an enantioselective aliphatic nitrilase encoded on the plasmid, which is induced by dimethylformamide [37].

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 [142], identified using the CRISPR-Finder tool [143]. 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 [144]. 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 [145]), 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 [146] 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 [147–149]. The generally large and complex genomes of Rhodococcus species afford a wide range of genes attributed to 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 [150] are present on the genome: proteases, lipases, esterases, reductases, nitrilase/cyanohydrolase/nitrile hydratases and amidases, transaminase, epoxide hydrolase, mono-oxygenases 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 [151] 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 5 kb 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: genomsize = 6 m) [152] 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 [153], to the assembly and the resulting Binary Alignment Map file was processed by Pilon [154] 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 [51]. BASys annotates based on microbial ab initio gene prediction using GLIMMER [82]. The genome sequence was also run on the RAST (Rapid Annotation using Subsystem Technology) server using the default RASTtk annotation pipeline for comparison [155, 156]. RAST annotation uses the manually curated SEED database to infer gene annotations based on protein functional roles within families [157]. 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.

Supplementary information
Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-019-6405-7.

**Additional file 1:** SuppInfo Frederick et al. BAA-870 genome. 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.

**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

**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.

**Authors’ contributions**
JP: 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. PTC and US: genome sequencing and data acquisition. RW, CAH and J-MGD: interpretation of sequence data. TS: academic supervisor of principal investigator. LGO: 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.

**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.

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

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare that they have no competing interests.

**Author details**
1Protein Technologies, CSIR Biosciences, Meiring Naude Road, Brummeria, Pretoria, South Africa. 2Electron Microscope Unit, University of Cape Town, Rondebosch 7701, South Africa. 3Present Address: LadHyx, UMR CNRS 7646, École Polytechnique, 91128 Palaiseau, France. 4Meraka, CSIR, Meiring Naude Road, Brummeria 0091, South Africa. 5Industrial Microbiology, Department of Biotechnology, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, The Netherlands. 6Biology and Biochemistry, University of Houston, 4800 Calhoun Road, Houston, TX 77204, USA. 7Present Address: Department of Pediatrics, Section of Tropical Medicine, Baylor College of Medicine, 1102 Bates Avenue, Houston, TX 77030, USA. 8Chemical and Biomolecular Engineering, University of Houston, 4800 Calhoun Road, Houston, TX 77204, USA. 9Bioinformatics and Computational Biology Unit, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0002, South Africa. 10Present Address: AgResearch Limited, Lincoln Research Centre, Private Bag 4749, Christchurch 8140, New Zealand. 11Biocatalysis, Department of Biotechnology, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, The Netherlands. 12Molecular Sciences Institute, School of Chemistry, University of the Witwatersrand, PO, Wits 2050, South Africa.

**Received:** 16 April 2019 **Accepted:** 16 December 2019

**References**
1. van der Geize R, Dijkhuizen L. Harnessing the catabolic diversity of *rhodococci* for environmental and biotechnological applications.Curr Opin Microbiol. 2004;7(2):255–61.
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, Eltsi LD. Catabolism of aromatic compounds and steroids by *Rhodococcus*. In: Alvarez HM, editor. Biology of Rhodococcus. Berlin: Springer Berlin Heidelberg; 2010. p. 133–69.
McWilliam H, Li W, Uludag M, Siquizato S, Park YM, Buso N, Cowley AP, Lopez R. Analysis tool web services from the EMBL-EBI. Nuc Acids Res. 2013; 41(W1):W597–600.

Li W, Cowley A, Uludag M, Gur T, McWilliam H, Siquizato S, Park YM, Buso N, Lopez R. The EMBL-EBI bioinformatics web and programmatic tools framework. Nuc Acids Res. 2015;43(W1):W580–4.

Severes F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7(539):1-6.

Meier-Kolthoff JP, Aich AF, Kleink H-P, Goker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinform. 2013;14(1):60.

Meier-Kolthoff JP, Hahnke BL, Petersen J, Scheuerer C, Michael V, Fiebig A, Rohde C, Rohde M, Fatmann B, Goodwin LA, et al. Complete genome sequence of DSM 30083', the type strain (U5/41) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy. Stand Genomic Sci. 2014;9(1):2.

Gardy JL, Brinkman FSL. Methods for predicting bacterial protein subcellular localization in sequence data. Nuc Acids Res. 2011;39(suppl_2):W445–9.

Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein classification: the COG database. Syst Biol. 2000;49(4):646–658.

Lowe TM, Eddy SR. RNAstructure: an algorithm and web server for RNA secondary structure prediction and folding. Nuc Acids Res. 2004;32(1):W485–9.

Galperin MY, Koonin EV. The COG database: a tool for genome-scale analysis of protein families and evolution. Nuc Acids Res. 2002;30(13):3296–3302.

Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nuc Acids Res. 2013;41(D1):D261–9.

Lowry OM, Rosebrough NJ, Farr AL, Randall RJ. Protein measurement with the Folin phenol reagent. J Biol Chem. 1951;193:265–275.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nuc Acids Res. 2013;41(D1):D36–42.

van Dommeela GH, Stothard P, Brinkman FSL. Methods for predicting bacterial protein subcellular localization in sequence data. Nuc Acids Res. 2011;39(suppl_2):W445–9.

Galperin MY, Koonin EV. Conserved hypothetical protein families: prioritization of targets for experimental study. Nuc Acids Res. 2004;32(18):5452–63.

Roberts RJ. Identifying protein function - a call for community action. PLoS Biol. 2004;2(3):e42.

Frishman D. Protein annotation at genomic scale: the current status. Chem Rev. 2007;107(8):3448–66.

Mills CL, Beunig PJ, Oudrechen MJ. Biochemical functional predictions for protein structures of unknown or uncertain function. Comput Struct Biotechnol J. 2015;13:182–91.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial genome annotation with G conservation. Nuc Acids Res. 1999;27(5):54–6.

Tatusov RL, Koonin EV. Identification of protein families and the evolution of the functional genomics database. Methods Enzymol. 2000;323:15–38.

Galperin MY, Koonin EV. Conserved hypothetical proteins: prioritization of targets for experimental study. Nuc Acids Res. 2004;32(18):5452–63.

Roberts RJ. Identifying protein function - a call for community action. PLoS Biol. 2004;2(3):e42.

Frishman D. Protein annotation at genomic scale: the current status. Chem Rev. 2007;107(8):3448–66.

Mills CL, Beunig PJ, Oudrechen MJ. Biochemical functional predictions for protein structures of unknown or uncertain function. Comput Struct Biotechnol J. 2015;13:182–91.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial genome annotation with G conservation. Nuc Acids Res. 1999;27(5):54–6.

Tatusov RL, Koonin EV. Identification of protein families and the evolution of the functional genomics database. Methods Enzymol. 2000;323:15–38.

Galperin MY, Koonin EV. Conserved hypothetical proteins: prioritization of targets for experimental study. Nuc Acids Res. 2004;32(18):5452–63.

Roberts RJ. Identifying protein function - a call for community action. PLoS Biol. 2004;2(3):e42.

Frishman D. Protein annotation at genomic scale: the current status. Chem Rev. 2007;107(8):3448–66.

Mills CL, Beunig PJ, Oudrechen MJ. Biochemical functional predictions for protein structures of unknown or uncertain function. Comput Struct Biotechnol J. 2015;13:182–91.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial genome annotation with G conservation. Nuc Acids Res. 1999;27(5):54–6.

Tatusov RL, Koonin EV. Identification of protein families and the evolution of the functional genomics database. Methods Enzymol. 2000;323:15–38.

Galperin MY, Koonin EV. Conserved hypothetical proteins: prioritization of targets for experimental study. Nuc Acids Res. 2004;32(18):5452–63.

Roberts RJ. Identifying protein function - a call for community action. PLoS Biol. 2004;2(3):e42.

Frishman D. Protein annotation at genomic scale: the current status. Chem Rev. 2007;107(8):3448–66.
110. Schiraldi C, Maresca C, Catapano A, Galinski EA, De Rosa M. High-yield production of hydroxyectoine: high cell-density cultivation of Marinococcus Biotechnol. 1995;43(1):53
97. Leipold F, Rudroff F, Mihovilovic MD, Bornscheuer UT. The steroid mycolate actinomycetes. Antonie Van Leeuwenhoek. 2003;84(2):89
98. 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 DC42201. Appl Environ Microbiol. 2015;81(6):1919–25.
99. Fignani E, Tonin F, Pollegioni L, Rosini E. Characterization and use of a bacterial ligin peroxidase with an improved manganese-oxidative activity. Appl Microbiol Biotechnol. 2018;102(24):1620–4.
100. van Beilen JB, Smits THM, Whyte LG, Schorcht S, Röthlisberger M, Plaggemeier T, Engesser K-H, Witholt B. Alkane hydroxylase homologues in bacterial lignin peroxidase with an improved manganese-oxidative activity. FEMS Microbiol Lett. 2011;62(1):59–64.
101. Schumaier DJ, Falkoua MS. Degradation of allylic molecules by Rhodococcus ruber CDA. Appl Microbiol Biotechnol. 1999;52(1):85–90.
102. Fink MJ, Rudloff F, Mihovilovic MD. Baeyer-Villiger monooxygenases in aroma compound synthesis. Bioorg Med Chem Lett. 2011;21(20):6315–8.
103. Galinski EA, Pfeiffer H-P, Trüper HG. 1,4,5,6-Tetrahydro-2-methyl-4-(10R)-tetrahydropyridinecarboxylic acid. A novel cyclic α-cyano acid from halophilic phototrophic bacteria of the genus Ectothiorhodospira. Eur J Biochem. 1985;149(1):135–9.
104. Roberts MF. Organic compatible solutes of halotolerant and halophilic microorganisms. Saline Syst. 2005;1(1).
105. García-Estepa R, Argandoña M, Reina-Bueno M, Capote N, Iglesias-Sánchez MR. Purification and characterization of a gene cluster for cyclododecanone oxidation in Rhodococcus ruber RHA1. Mol Microbiol. 2007;65(3):828–38.
106. Coffey L, Ovens E, Tambling K, O’Neill D, 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):455–63.
107. Yamaguchi T, Asano Y. Draft genome sequence of Rhodococcus sp. strain M0, which can degrade a broad range of nitriles. Genome Announc. 2018;6(6):e01526–17.
108. Brackets D, Dube N, Petersen R. Green chemistry: highly selective biocatalytic transformations catalyzed by Rhodococcus erythropolis A4. Tetrahedron Asymmetry. 2010;21(6):695–702.
109. Velankar H, Clarke KG, Preez R, Cowan DA, Burton SG. Developments in the taxonomy of the genus Rhodococcus. FEMS Microbiol Rev. 2004;28(3):377–403.
110. 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 U S A. 2014;111(25):9259–64.
111. Dickschat JS. Bacterial terpene cyclases. Nat Prod Rep. 2011;28(7):799–822.
112. Yamada Y, Kuzuyama T, Komatsu M, Shin-ya K, Omura S, Cane DE, Ikeda H. Purification and characterization of a novel nitrile hydroxylase from Rhodococcus sp. strain RJ1. Mol Microbiol. 2009;75(5):382–98.
113. Seifert DH. Selective oxidation of carbolide C6(3+S)-amine by using Rhodococcus sp. strain ST-10 styrene monooxygenase expressed in organic-solvent-tolerant Kocuria rhizophila DC42201. Appl Environ Microbiol. 2015;81(6):1919–25.
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 U S A. 2014;111(25):9259–64.
115. Brandão PFB, Clapp JP, Bull AT. Diversity of nitrile hydratase and their stereoselective biotransformation catalyzed by Rhodococcus erythropolis A4. Tetrahedron Asymmetry. 2010;21(6):695–702.
116. Kamal A, Kumar MS, Kumar CG, Shaik TB. Bioconversion of acrylonitrile to γ-lactam by a Rhodococcus sp. strain YH3-3. Appl Environ Microbiol. 1999;65(1):85–90.
117. Garcia-Estepa R, Argandoña M, Reina-Bueno M, Capote N, Iglesias-Sánchez MR. Purification and characterization of a gene cluster for cyclododecanone oxidation in Rhodococcus ruber RHA1. Mol Microbiol. 2007;65(3):828–38.
118. Coffey L, Ovens E, Tambling K, O’Neill D, 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):455–63.
119. D’Antoni N, Nicolosi G, Mornone R, Kubac D, Kaplan O, Martinková L. Synthesis of novel cyano-cyclitols and their stereoselective biotransformation catalyzed by Rhodococcus erythropolis A4. Tetrahedron Asymmetry. 2010;21(6):695–702.
120. Novikov AD, Lavrov KV, Kasianov AS, Gerasimova TV, Yanenko AS. Draft genome sequence of Rhodococcus sp. strain M0, which can degrade a broad range of nitriles. Genome Announc. 2018;6(6):e01526–17.
121. Yamaguchi T, Asano Y. Draft genome sequence of an aldolase degrader, Rhodococcus sp. strain YH3-3. Genome Announc. 2016;4(3):e00406–16.
122. Güttler V, Mayall BC, Seviour R. Can whole genome analysis refine the taxonomy of the genus Rhodococcus? FEMS Microbiol Rev. 2004;28(3):377–403.
123. Okamoto S, Eis利 LD. Purification and characterization of a novel nitrile hydrazine from Rhodococcus sp. strain RJ1. Mol Microbiol. 2007;65(3):828–38.
124. Coffey L, Ovens E, Tambling K, O’Neill D, 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):455–63.
125. D’Antoni N, Nicolosi G, Mornone R, Kubac D, Kaplan O, Martinková L. Synthesis of novel cyano-cyclitols and their stereoselective biotransformation catalyzed by Rhodococcus erythropolis A4. Tetrahedron Asymmetry. 2010;21(6):695–702.
126. Sheng Y, Wang M, Li X, Zhang J, Sun H, Liu J. Highly efficient synthesis of 5-cyanovaleramide by Rhodococcus ruber CGMCXC3090 resting cells. J Chem Technol Biotechnol. 2012;87(10):1396–400.
127. Brandão PFB, Clapp JP, Bull AT. Diversity of nitrile hydratase and amidase enzymes in Rhodococcus erythropolis recovered from geographically distinct habitats. Appl Environ Microbiol. 2003;69(10):5754–66.
141. 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):4807–15.

142. Pathak A, Chauhan A, Blom J, Indest KJ, Jung CM, Stothard P, Bera G, Green SI, 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.

143. Grissa I, Poulcel C, Vergnaud G. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nuc Acids Res. 2007;35(suppl_2):W52–7.

144. 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 β-estradiol-degrading bacterium. Int J Genomics. 2018:2018:11.

145. 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):641–51.

146. Arndt D, Grant JR, Marcu A, Saged T, Pon A, Liang Y, Wishart DS. PHASTER: a better, faster version of the PHAST phage search tool. Nuc Acids Res. 2016;44(W1):W16–21.

147. Zheng G-W, Xu J-H. New opportunities for biocatalysis: driving the synthesis of chiral chemical. Curr Opin Biotechnol. 2011;22(6):784–92.

148. Panke S, Wubbolts M. Advances in biocatalytic synthesis of pharmaceutical intermediates. Curr Opin Chem Biol. 2005;9(2):188–94.

149. Jemli S, Ayadi-Zouari D, Hlima HB, Bejar S. Biocatalysts: application and engineering for industrial purposes. Crit Rev Biotechnol. 2016;36(2):246–58.

150. Pollard DJ, Woodley JM. Biocatalysis for pharmaceutical intermediates: the future is now. Trends Biotechnol. 2007;25(2):66–73.

151. Yam KC, Okamoto S, Roberts JN, Eltis LD. Adventures in *Rhodococcus* - from steroids to explosives. Can J Microbiol. 2011;57(3):155–68.

152. 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):722–36.

153. Li H, Durbin R. Fast and accurate long-read alignment with burrows–wheeler transform. Bioinformatics. 2010;26(5):589–95.

154. 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.

155. Aziz R, Bartels D, Best A, Delongh M, Disz T, Edwards R, Formisso K, Gerdes S, Glass E, Kubal M, et al. The RAST server: rapid annotations using subsystems technology. BMC Genomics. 2008;9(1):75.

156. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, et al. RASTK: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep. 2015;5:8365.

157. 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):D206–14.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.