Genome sequence of the reddish-pigmented
Rubellimicrobium thermophilum type strain (DSM 16684\textsuperscript{T}), a member of the Roseobacter clade

Anne Fiebig\textsuperscript{1}, Thomas Riedel\textsuperscript{2,3§}, Sabine Gronow\textsuperscript{1}, Jörn Petersen\textsuperscript{1}, Hans-Peter Klenk\textsuperscript{1*}, Markus Göker\textsuperscript{1}

\textsuperscript{1} Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany
\textsuperscript{2} UPMC Université Paris 6, UMR 7621, Observatoire Océanologique, Banyuls-sur-Mer, France
\textsuperscript{3} CNRS, UMR 7621, LOMIC, Observatoire Océanologique, Banyuls-sur-Mer, France

*Corresponding author: Hans-Peter Klenk

§Former address: HZI – Helmholtz Centre for Infection Research, Braunschweig, Germany

Keywords: rod-shaped, reddish-pigmented, thermophile, chemoheterotrophic, prophage-like structures, Rhodobacteraceae, Roseobacter clade, Alphaproteobacteria

Rubellimicrobium thermophilum Denner et al. 2006 is the type species of the genus Rubellimicrobium, a representative of the Roseobacter clade within the Rhodobacteraceae. Members of this clade were shown to be abundant especially in coastal and polar waters, but were also found in microbial mats and sediments. They are metabolically versatile and form a physiologically heterogeneous group within the Alphaproteobacteria. Strain C-Ivk-R2A-2\textsuperscript{T} was isolated from colored deposits in a pulp dryer; however, its natural habitat is so far unknown. Here we describe the features of this organism, together with the draft genome sequence and annotation and novel aspects of its phenotype. The 3,161,245 bp long genome contains 3,243 protein-coding and 45 RNA genes.

Introduction

Strain C-Ivk-R2A-2\textsuperscript{T} (= DSM 16684 = CCUG 51817 = HAMBI 2421) is the type strain of the species Rubellimicrobium thermophilum [1]. The genus name Rubellimicrobium was derived from the Neo-Latin adjective 'rubellus', red or reddish, and the Neo-Latin noun 'microbium', microbe, referring to its reddish pigmentation. The species epithet was derived from the Greek noun 'thrmê', heat, as well as from the Neo-Latin adjective 'philus –a –um', friend/loving, referring to its growth temperature [1]. C-Ivk-R2A-2\textsuperscript{T} was isolated from colored deposits in a pulp dryer in Finland, so the natural habitat is so far unknown [1]. At the time of writing, PubMed records did not indicate any follow-up research with strain C-Ivk-R2A-2\textsuperscript{T} after the initial description and valid publication of the new species Rubellimicrobium thermophilum [1]. Here we present a summary classification and a set of features for R. thermophilum C-Ivk-R2A-2\textsuperscript{T}, together with the description of the genomic sequencing and annotation. We also describe novel aspects of its phenotype.

Features of the organism

16S rRNA gene analysis

The single genomic 16S rRNA gene sequence of R. thermophilum DSM 16684\textsuperscript{T} was compared using NCBI BLAST [2,3] under default settings (e.g., considering only the high-scoring segment pairs (HSPs) from the best 250 hits) with the most recent release of the Greengenes database [4] and the relative frequencies of taxa and keywords (reduced to their stem [5]) were determined, weighted by BLAST scores. The most frequently occurring genera were Rubellimicrobium (26.9%), Oceanicola (18.5%), Rhodobacter (12.4%), Methylarcula (10.4%) and Loktanella (10.1%) (37 hits in total). Regarding the five hits to sequences from members of the species, the average identity within HSPs was 99.9%, whereas the average coverage by HSPs was 99.2%. Among all other species, the one yielding the highest score was 'Pararubellimicrobium aerilata' (EU338486), which corresponded to an identity of 94.2% and an HSP coverage of 97.8%. (Note that the Greengenes database uses the INSDC (=
EMBL/NCBI/DDBJ) annotation, which is not an authoritative source for nomenclature or classification.) The highest-scoring environmental sequence was AJ489269 (Greengenes short name 'food Echinamoeba thermarum clone'), which showed an identity of 99.9% and an HSP coverage of 99.1%. The most frequently occurring keywords within the labels of all environmental samples which yielded hits were 'skin' (10.1%), 'fossa' (6.0%), 'poplit' (3.6%), 'forearm, volar' (3.6%) and 'water' (2.5%) (213 hits in total). The most frequently occurring keywords within the labels of those environmental samples which yielded hits of a higher score than the highest scoring species were 'biofilm' (18.2%), 'echinamoeba, food, thermarum' (9.1%) and 'color, machin, moder, paper, paper-machin, thermophil' (9.1%) (2 hits in total). Figure 1 shows the phylogenetic neighborhood of R. thermophilum in a 16S rRNA sequence based tree. The sequence of the single 16S rRNA gene copy in the genome does not differ from the previously published 16S rDNA sequence (AJ844281).

Figure 1 Phylogenetic tree highlighting the position of R. thermophilum relative to the type strains of the type species of the other genera within the family Rhodobacteraceae. The tree was inferred from 1,330 aligned characters [6,7] of the 16S rRNA gene sequence under the maximum likelihood (ML) criterion [8]. Rooting was done initially using the midpoint method [9] and then checked for its agreement with the current classification (Table 1). The branches are scaled in terms of the expected number of substitutions per site. Numbers adjacent to the branches are support values from 650 ML bootstrap replicates [10] (left) and from 1,000 maximum-parsimony bootstrap replicates [11] (right) if larger than 60%. Lineages with type strain genome sequencing projects registered in GOLD [12] are labeled with one asterisk, those also listed as 'Complete and Published' with two asterisks [13].

http://standardsingenomics.org
Morphology and physiology

Cells of strain C-Ivk-R2A-2\(^{T}\) stain Gram-negative are rod shaped and 0.6-0.8 µm in width and 2.0-4.0 µm in length (Figure 2) [1]. Cells are motile and possess one to three polar flagella [1]. C-Ivk-R2A-2\(^{T}\) is moderately thermophilic and grows over a temperature range of 28-56°C with an optimum between 45°C and 52°C, whereas no growth occurs at room temperature or at temperatures higher than 57°C. Colonies grown on Reasoner’s 2A agar (R2A) at 45°C for 2 days are translucent, entire, convex and smooth. Cells are red-pigmented (carotenoids); the pigment absorbance spectrum reveals three distinct peaks: one major peak at 495 nm and to others at 465 nm and 525 nm [1]. Cells are strictly aerobic. Growth does not occur under anaerobic conditions whether or not the cultures are grown in the dark or the light. Bacteriochlorophyll \(a\) is not synthesized. Cells are cytochrome c-oxidase positive, weakly positive for catalase as well as urease-positive. Nitrate is not reduced [1]. Intracellular inclusion bodies containing polyphosphate and polyhydroxyalkanoates are produced [1].

The cells of strain C-Ivk-R2A-2\(^{T}\) assimilate the following compounds: L-arabinose, \(p\)-arbutin, D-cellbiose, D-fructose, D-galactose, gluconate, D-glucose, D-mannose, D-maltose, \(\alpha\)-D-melibiose, D-rhamnose, D-ribose, sucrose, salicin, D-trehalose, D-xylose, adonitol, \(\text{myo}\)-inositol, maltitol, D-mannitol, D-sorbitol, acetate, 4-aminobutyrate, glutarate, DL-3-hydroxybutyrate, DL-lactate, L-malate, oxoglutarate, pyruvate, L-alanine, L-ornithine and L-proline. Cells do not produce acid from D-glucose, lactose, sucrose, L-arabinose, L-rhamnose, maltose, D-xylose, cellobiose, D-mannitol, dulcitol, salicin, adonitol, \(\text{myo}\)-inositol, sorbitol, raffinose, trehalose, methyl \(\alpha\)-D-glucoside, erythritol, melibiose, D-arabitol or D-mannose [1]. The strain does not assimilate the following compounds: \(N\)-acetyl-D-glucosamine, putrescine, propionate, \(cis\) and \(trans\)-aconitate, adipate, azelate, citrate, fumarate, itaconate, mesaconate, suberate, \(\beta\)-alanine, L-aspartate, L-histidine, L-leucine, L-phenylalanine, L-serine, L-tryptophan, 3- and 4-hydroxybenzoate and phenylacetate [1]. The following compounds are hydrolyzed by strain C-Ivk-R2A-2\(^{T}\): \(p\)-nitrophenyl (pNP) \(\alpha\)-D-glucopyranoside, pNP \(\beta\)-D-glucopyranoside, bis-pNP phosphatase, pNP phenylphosphonate and L-alanine \(p\)-nitroanilide (pNA), whereas aesculin, pNP \(\beta\)-D-galactopyranoside, pNP \(\beta\)-D-glucuronide, pNP phosphorylcholine, \(2\)-deoxythymidine-5\(^{\prime}\)-pNP phosphate, L-glutamate-\(\gamma\)-3-carboxy pNA, L-proline pNA, Tween 80, starch and casein are not hydrolyzed [1].

Strain C-Ivk-R2A-2\(^{T}\) was also found to be susceptible to ampicillin, chloramphenicol, colistin sulfate, gentamicin, kanamycin, lincomycin, neomycin, nitrofurantoin, penicillin G, polymyxin B, streptomycin, tetracycline and vancomycin [1].
The physiology of *R. thermophilum* DSM 16684T was investigated in this study using Generation-III microplates in an OmniLog phenotyping device (BIOLOG Inc., Hayward, CA, USA). The microplates were inoculated at 28°C and 37°C, respectively, with a cell suspension at a cell density of 95-96% turbidity and dye IF-A. Further additives were vitamin, micronutrient and sea-salt solutions. The plates were sealed with parafilm to avoid a loss of fluid. The exported measurement data were further analyzed with the omple package for R [23,24], using its functionality for statistically estimating parameters from the respiration curves such as the maximum height, and automatically translating these values into negative and positive reactions.

At 28°C, the strain was positive for D-turanose, pH 6, 1% NaCl, 4% NaCl, D-galactose, 3-O-methyl-D-glucose, D-fucose, L-fucose, L-rhamnose, 1% sodium lactate, myo-inositol, rifampicin SV, L-aspartic acid, L-glutamic acid, L-histidine, L-serine, D-glucuronic acid, quinic acid, L-lactic acid, citric acid, α-keto-glutaric acid, D-malic acid, L-malic acid, nalidixic acid, potassium tellurite, aceotocetic acid and sodium formate. The strain was negative for dextrin, D-maltose, D-trehalose, D-cellobiose, β-gentiobiose, sucrose, stachyose, pH 5, D-raffinose, α-D-lactose, D-melibiose, β-methyl-D-galactoside, D-salicyl, N-acetyl-D-glucosamine, N-acetyl-β-D-mannosamine, N-acetyl-D-galactosamine, N-acetyl-neuraminic acid, 8% NaCl, D-glucose, D-mannose, D-fructose, inosine, fusidic acid, D-serine, D-sorbitol, D-mannitol, D-arabitol, glycero1, D-glucose-6-phosphate, D-fructose-6-phosphate, D-aspartic acid, D-serine, trehalomycin, minocycline, gelatin, glycolyl-L-proline, L-alanine, L-arginine, L-pyrogulamic acid, glycolyl-D-proline, L-pyroglutamic acid, lincomycin, guanidine hydrochloride, niaprole, D-galacturonic acid, L-galactonic acid-γ-lactone, mucic acid, D-saccharic acid, vancomycin, tetrazolium violet, tetrazolium blue, p-hydroxy-phenylacetic acid, D-glucuronic acid, D-glucaric acid, D-saccharic acid, vancomycin, tetrazolium violet, tetrazolium blue, p-hydroxy-phenylacetic acid, methyl pyruvate, D-lactic acid methyl ester, bromo-succinic acid, lithium chloride, tween 40, γ-amino-n-butyric acid, α-hydroxy-butyric acid, β-hydroxy-butyric acid, α-keto-butyric acid, propionic acid, acetic acid, aztreonam, butyric acid and sodium bromate.

At 37°C, the strain was positive for D-maltose, D-trehalose, D-cellobiose, β-gentiobiose, sucrose, D-turanose, stachyose, pH 6, D-raffinose, D-melibiose, β-methyl-D-galactoside, D-salicyl, 1% NaCl, 4% NaCl, D-glucose, D-mannose, D-fructose, D-galactose, 3-O-methyl-D-glucose, D-fucose, L-rhamnose, L-sorbose, L-rhamnose, inosine, 1% sodium lactate, D-sorbitol, D-mannitol, D-arabitol, myo-inositol, glycerol, rifampicin SV, L-alanine, L-arginine, L-aspartic acid, L-glutamic acid, L-histidine, L-serine, pectin, D-glucuronic acid, D-glucaric acid, glucuronamide, quinic acid, methyl pyruvate, L-lactic acid, citric acid, α-keto-glutaric acid, D-malic acid, L-malic acid, nalidixic acid, potassium tellurite, tween 40, γ-amino-n-butyric acid, β-hydroxy-butyric acid, propionic acid, acetic acid and sodium formate. No reactions could be observed for dextrin, pH 5, α-D-lactose, N-acetyl-D-glucosamine, N-acetyl-β-D-mannosamine, N-acetyl-D-galactosamine, N-acetyl-neuraminic acid, 8% NaCl, fusidic acid, D-serine, D-glucose-6-phosphate, D-fructose-6-phosphate, D-aspartic acid, D-serine, trehalomycin, minocycline, gelatin, glycolyl-L-proline, L-pyroglutamic acid, lincomycin, guanidine hydrochloride, niaprole, D-galacturonic acid, L-galactonic acid-γ-lactone, mucic acid, D-saccharic acid, vancomycin, tetrazolium violet, tetrazolium blue, p-hydroxy-phenylacetic acid, D-lactic acid methyl ester, bromo-succinic acid, lithium chloride, α-hydroxy-butyric acid, α-keto-butyric acid, acetoacetic acid, aztreonam, butyric acid and sodium bromate.

According to [1], *R. thermophilum* is able to metabolize a wide range of carbon sources. This observation is not fully confirmed by the OmniLog measurements at 28°C. For instance, more than eleven sugars were not metabolized under the given cultivation conditions in the Generation-III microplates. This is apparently caused by distinct cultivation conditions, because the behavior is in high agreement with [1] if a temperature of 37°C is chosen, which is closer to the reported optimum temperature [1]. Particularly the optimal growth temperature of 45°C highly differs from the one that had to be used in the OmniLog assays (28°C). Conversely, in contrast to [1] the OmniLog measurements yielded positive reactions for citrate, L-histidine and L-serine at 28°C and additionally for propionate at 37°C. This may be due to the higher sensitivity of respiratory measurements compared to growth measurements [24,25].

http://standardsingenomics.org
Chemotaxonomy

The principal cellular fatty acids of strain C-Ivk-R2A-2T are C19:0 cyclo ω7c (43.9 %), C16:0 (22.3 %), C18:0 (22.0 %), C18:1 ω7c (4.5 %), C10:0 3-OH (1.2 %), C18:1 11-methyl (0.9 %), C20:2 ω6c ω9c (0.7 %), C17:0 cyclo (0.5 %)C17:0 (0.4 %) and summed feature 2 containing C16:1iso I and/or C14:0 3-OH (1.2 %). Two unknown fatty acids are identified by their equivalent chain length (ECL): ECL 11.799 (2.3 %) as well as ECL 17.322 (0.7 %) [1].

Additionally, ubiquinone Q-10 is the predominant respiratory lipoquinone, but ubiquinone Q-9 was also detected in minor amounts [1]. The polyamine pattern is characterized by the major compounds spermidine (11.5 µmol/g dry weight), sym-homospermidine (9.7 µmol/g dry weight) and putrescine (8.9 µmol/g dry weight) [1]. Minor polyamine pattern compounds are spermine (1.9 µmol/g dry weight), sym-norspermidine (0.4 µmol/g dry weight), cadaverine (0.1 µmol/g dry weight) and diaminopropane in trace amounts [1]. Interestingly, this polyamine composition is different from other previously reported members of the family Rhodobacteraceae. Whereas members of Paracoccus, Rhodobacter, Rhodovulum and Roseomonas mainly contain spermidine and putrescine, the polyamine compound sym-homospermidine is not detectable in these representatives [1].

Table 1. Classification and general features of R. thermophilum C-Ivk-R2A-2T according to the MIGS recommendations [14].

| MIGS ID  | Property                  | Term                                      | Evidence code |
|----------|---------------------------|-------------------------------------------|---------------|
| MIGS-7   | MIGS-7 Current classification | Domain Bacteria                           | TAS [15]      |
|          |                            | Phylum Proteobacteria                      | TAS [16]      |
|          |                            | Class Alphaproteobacteria                 | TAS [17]      |
|          |                            | Order Rhodobacterales                     | TAS [18,19]   |
|          |                            | Family Rhodobacteraceae                  | TAS [20]      |
|          |                            | Genus Rubellimicrobium                   | TAS [1]       |
|          |                            | Species Rubellimicrobium thermophilum     | TAS [1]       |
|          |                            | Type strain C-Ivk-R2A-2T                  | TAS [1]       |
|          | Gram stain                | negative                                  | TAS [1]       |
|          | Cell shape                | rod-shaped                                | TAS [1]       |
|          | Motility                  | motile                                    | TAS [1]       |
|          | Sporulation               | non-sporulating                           | TAS [1]       |
|          | Temperature range          | thermophile (28°C – 56°C)                | TAS [1]       |
|          | Optimum temperature       | 45-52°C                                  | TAS [1]       |
|          | Salinity                  | stenohaline                              | NAS           |
| MIGS-22  | Relationship to oxygen    | aerobic                                   | TAS [1]       |
|          | Carbon source             | mono- and polysaccharides                | TAS [1]       |
| MIGS-6   | Habitat                   | not reported                              |               |
| MIGS-6.2 | pH                        | not reported                              |               |
| MIGS-15  | Biotic relationship       | free living                              | NAS           |
|          | Biosafety level           | 1                                         | TAS [21]      |
| MIGS-23.1| Isolation                | colored deposits in a pulp dryer         | TAS [1]       |
| MIGS-4   | Geographic location       | Finland                                   | TAS [1]       |
| MIGS-4.1 | Latitude                  | not reported                              |               |
| MIGS-4.2 | Longitude                 | not reported                              |               |
| MIGS-4.3 | Depth                     | not reported                              |               |

Evidence codes - TAS: Traceable Author Statement (i.e., a direct report exists in the literature); NAS: Nontraceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). Evidence codes are from the Gene Ontology project [22].
**Genome sequencing and annotation**

**Genome project history**
The genome was sequenced within the project “Ecology, Physiology and Molecular Biology of the *Roseobacter* clade: Towards a Systems Biology Understanding of a Globally Important Clade of Marine Bacteria” funded by the German Research Council (DFG). The strain was chosen for genome sequencing according the *Genomic Encyclopedia of Bacteria and Archaea* (GEBA) criteria [26,27]. Project information is stored at the Genomes On-Line Database [12]. The Whole Genome Shotgun (WGS) sequence is deposited in Genbank and the Integrated Microbial Genomes database (IMG) [28]. A summary of the project information is shown in Table 2.

**Growth conditions and DNA isolation**
A culture of DSM 16684T was grown aerobically in DSMZ medium 830 (R2A medium) [29] at 45°C. Genomic DNA was isolated using Jetflex Genomic DNA Purification Kit (GENOMED 600100) following the standard protocol provided by the manufacturer but modified by an incubation time of 60 min, the incubation on ice over night on a shaker, the use of additional 50 µl proteinase K, and the addition of 100 µl protein precipitation buffer. DNA is available from DSMZ through the DNA Bank Network [30].

**Genome sequencing and assembly**
The genome was sequenced using a combination of Illumina and 454 libraries (Table 2). Illumina sequencing was performed on a GA IIx platform with 150 cycles. The paired-end library contained inserts of 456 nt length on average. To correct sequencing errors and improve quality of the reads, clipping was performed using fastq-mcf [31] and quake [32]. The remaining 4,190,250 reads with an average length of 106 nt were assembled using Velvet [33]. To gain information on the contig arrangement an additional 454 run was performed. The paired-end jumping library of 3 kb insert size was sequenced on a 1/8 lane. Pyrosequencing resulted in 115,925 reads, with an average read length of 451 nt, assembled with Newbler (Roche Diagnostics) into a draft assembly comprising 36 scaffolds. Both draft assemblies (Illumina and 454 sequences) were fractionated into artificial Sanger reads of 1000 nt in length plus 75 nt overlap on each site. These artificial reads served as an input for the phred/phrap/consed package [34]. By manual editing the number of contigs was reduced to 44 organized in ten scaffolds. The combined sequences provided a 203 × coverage of the genome.

**Table 2. Genome sequencing project information**

| MIGS ID | Property                  | Term                                                      |
|---------|---------------------------|-----------------------------------------------------------|
| MIGS-31 | Finishing quality         | Non-contiguous finished                                   |
| MIGS-28 | Libraries used            | Two genomic libraries: one Illumina PE library (456 bp insert size), one 454 PE library (3kb insert size) |
| MIGS-29 | Sequencing platforms      | Illumina GAIIx, 454 GS-FLX + Titanium (Roche)             |
| MIGS-31.2 | sequencing coverage     | 203 ×                                                     |
| MIGS-30 | Assemblers                | Velvet version 1.1.36, Newbler version 2.3, consed 20.0   |
| MIGS-32 | Gene calling method       | Prodigal 1.4                                              |
| INSDC ID |                           | AAOLV000000000                                          |
| GenBank Date of Release | July 31, 2013 |                                                             |
| GOLD ID |                           | Gi11864                                                   |
| NCBI project ID | 178147          |                                                             |
| Database: IMG | 2521172529     |                                                             |
| MIGS-13 | Source material identifier| DSM 16684                                                 |
| Project relevance | Tree of Life, biodiversity |                                                             |
Rubellimicrobium thermophilum DSM 16684T

**Genome annotation**

Genes were identified using Prodigal [35] as part of the JGI genome annotation pipeline [36]. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) non-redundant database, UniProt, TIGR-Fam, Pfam, PRIAM, KEGG, COG, and InterPro databases. Identification of RNA genes were carried out by using HMMER 3.0rc1 [37] (rRNAs) and tRNAscan-SE 1.23 [38] (tRNAs). Other non-coding genes were predicted using INFERNAL 1.0.2 [39] Additional gene prediction analysis and functional annotation was performed within the Integrated Microbial Genomes - Expert Review (IMG-ER) platform [40]. CRISPR elements were detected using CRT [41] and PILER-CR [42].

**Genome properties**

The genome statistics are provided in Table 3 and Figure 3. The genome has a total length of 3,161,245 bp and a G+C content of 69.1%. Of the 3,288 genes predicted, 3,243 were protein-coding genes, and 45 RNAs. The majority of the protein-coding genes (80.4%) were assigned a putative function while the remaining ones were annotated as hypothetical proteins. The distribution of genes into COGs functional categories is presented in Table 4.

| Table 3. Genome Statistics | Value | % of Total |
|----------------------------|-------|------------|
| Genome size (bp)           | 3,161,245 | 100.00    |
| DNA coding region (bp)     | 2,813,333 | 88.99     |
| DNA G+C content (bp)       | 2,185,501 | 69.13     |
| Number of scaffolds         | 10     |            |
| Total genes                | 3,288  | 100.00     |
| RNA genes                  | 45     | 1.37       |
| rRNA operons               | 1      |            |
| tRNA genes                 | 37     | 1.13       |
| Protein-coding genes       | 3,243  | 98.63      |
| Genes with function prediction (proteins) | 2,645 | 80.44 |
| Genes in paralog clusters  | 2,688  | 81.75      |
| Genes assigned to COGs     | 2,599  | 79.05      |
| Genes assigned Pfam domains| 2,689  | 81.78      |
| Genes with signal peptides | 235    | 7.15       |
| Genes with transmembrane helices | 664  | 20.19     |
| CRISPR repeats             | 1      |            |
Insights into the genome

The ten scaffolds of the draft genome sequence of strain C-Ivk-R2A-2T were screened with BLAST for the presence of the four abundant plasmid replicases from the Rhodobacterales, representing DnaA-like, RepABC-, RepA- and RepB-type replicons [43]. None of these typical extrachromosomal elements was detected.

Prophage-like structures have been found in many bacteria and they are known to drive the diversity of bacteria by facilitating lateral gene transfer [44]. Genome analysis of strain DSM 16684T further revealed the presence of several genes encoding proteins associated with prophages (ruthe_00218 to 00220, ruthe_00605, ruthe_00607 to 00610, ruthe_00612, ruthe_00614, ruthe_00617, ruthe_00618, ruthe_00620, ruthe_2061, ruthe_2066, ruthe_02072, ruthe_02185, ruthe_02480, ruthe_02482 to 02484, ruthe_02495, ruthe_02499, ruthe_02502, ruthe_02792, ruthe_02972, ruthe_02974, ruthe_02976, ruthe_02977, ruthe_02984, ruthe_02988, and ruthe_02991 to 03295).

The soxB gene (ruthe_01788) encodes a component of the thiosulfate-oxidizing Sox enzyme complex, which is known to be part of the genomes of various groups of bacteria [45]. Several other genes involved in this process were also detected (e.g. ruthe_01784, ruthe_01785 and ruthe_01786).

Genome analysis of strain R. thermophilum DSM 16684T further revealed the presence of several genes encoding proteins associated with the utilization of urease (ruthe_02149 to 02151, ruthe_02153 to 02156). Several genes encoding proteins involved in the transport of Fe3+-siderophores and Fe3+-hydroxamate via ABC-transporters were also detected (e.g. ruthe_03167 to 03172).

Additionally, several gene sequences associated with CRISPRs (ruthe_02227 to 02230, ruthe_02232 to 02234, ruthe_02250, ruthe_02251, ruthe_02253 and ruthe_02255), cytochrome c oxidase activity (ruthe_00413 to 00417), cytochrome cbb3 oxidase activity (ruthe_01647 to 01654) as well as cytochrome bd-I ubiquinol oxidase activity (ruthe_01776, ruthe_01777) were found.

Additional gene sequences of interest encode a predicted ring-cleavage extradiol dioxygenase (ruthe_00477), which indicates a possible degradation of aromatic compounds. A sensor of blue light using FAD (BLUF, ruthe_01818) was also found, indicating possible blue-light dependent signal transduction.
Table 4. Number of genes associated with the general COG functional categories

| Code | Value | Percentage | Description                                                                 |
|------|-------|------------|-----------------------------------------------------------------------------|
| J    | 149   | 5.2        | Translation, ribosomal structure and biogenesis                             |
| A    | 3     | 0.1        | RNA processing and modification                                              |
| K    | 162   | 5.7        | Transcription                                                               |
| L    | 117   | 4.1        | Replication, recombination and repair                                        |
| B    | 3     | 0.1        | Chromatin structure and dynamics                                             |
| D    | 25    | 0.9        | Cell cycle control, cell division, chromosome partitioning                  |
| Y    | 0     | 0.0        | Nuclear structure                                                           |
| V    | 30    | 1.1        | Defense mechanisms                                                          |
| T    | 86    | 3.0        | Signal transduction mechanisms                                               |
| M    | 165   | 5.8        | Cell wall/membrane/envelope biogenesis                                      |
| N    | 32    | 1.1        | Cell motility                                                               |
| Z    | 1     | 0.0        | Cytoskeleton                                                                |
| W    | 0     | 0.0        | Extracellular structures                                                    |
| U    | 51    | 1.8        | Intracellular trafficking and secretion, and vesicular transport             |
| O    | 118   | 4.1        | Posttranslational modification, protein turnover, chaperones                |
| C    | 170   | 5.9        | Energy production and conversion                                            |
| G    | 306   | 10.7       | Carbohydrate transport and metabolism                                        |
| E    | 334   | 11.7       | Amino acid transport and metabolism                                          |
| F    | 74    | 2.6        | Nucleotide transport and metabolism                                         |
| H    | 128   | 4.5        | Coenzyme transport and metabolism                                           |
| I    | 99    | 3.5        | Lipid transport and metabolism                                              |
| P    | 145   | 5.1        | Inorganic ion transport and metabolism                                       |
| Q    | 85    | 3.0        | Secondary metabolites biosynthesis, transport and catabolism                |
| R    | 331   | 11.5       | General function prediction only                                             |
| S    | 254   | 8.9        | Function unknown                                                            |
| -    | 689   | 21.0       | Not in COGs                                                                 |

Acknowledgements

The authors gratefully acknowledge the assistance of Iljana Schröder for growing *R. thermophilum* DSM 16684T cultures and Meike Döppner for DNA extraction and quality control (both at the DSMZ). This work was performed under the auspices of the German Research Foundation (DFG) Transregio-SFB 51 Roseobacter grant.

References

1. Denner EBM, Kolari M, Hoornstra D, Tsigko I, Kämpfer P, Busse HJ, Salkinoja-Salonen M. *Rubellimicrobium thermophilum* gen. nov., sp. nov., a red-pigmented, moderately thermophilic bacterium isolated from coloured slime deposits in paper machines. *Int J Syst Evol Microbiol* 2006; 56:1355-1362. PubMed [http://dx.doi.org/10.1099/ijs.0.63751-0](http://dx.doi.org/10.1099/ijs.0.63751-0)

2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; 215:403-410. PubMed

3. Korf I, Yandell M, Bedell J. BLAST, O’Reilly, Sebastopol, 2003.

4. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a Chimera-Checked 16S
5. Porter MF. An algorithm for suffix stripping. Program: electronic library and information systems 1980; 14:130-137.

6. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. Bioinformatics 2002; 18:452-464. PubMed http://dx.doi.org/10.1093/bioinformatics/18.3.452

7. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 2000; 17:540-552. PubMed http://dx.doi.org/10.1093/oxfordjournals.molbev.a026334

8. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML web-servers. Syst Biol 2008; 57:758-771. PubMed http://dx.doi.org/10.1080/10635150802429642

9. Hess PN, De Moraes Russo CA. An empirical test of the midpoint rooting method. Biol J Linn Soc Lond 2007; 92:669-674. http://dx.doi.org/10.1111/j.1095-8312.2007.00864.x

10. Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A. How many bootstrap replicates are necessary? Lect Notes Comput Sci 2009; 5541:184-200. http://dx.doi.org/10.1007/978-3-642-02008-7_13

11. Swofford DL. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.0 b10. Sinauer Associates, Sunderland, 2002.

12. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 2012; 40:D571-D579. PubMed http://dx.doi.org/10.1093/nar/gkr1100

13. Wagner-Döbler I, Ballhausen B, Berger M, Brinkhoff T, Buchholz I, Bunk B, Cypionka H, Daniel R, Drepper T, Gerds T, et al. The complete genome sequence of the algal symbiont Dinoroseobacter shibae – a hitchhiker’s guide to life in the sea. ISME J 2010; 4:61-77. PubMed http://dx.doi.org/10.1038/ismej.2009.94

14. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol 2008; 26:541-547. PubMed http://dx.doi.org/10.1038/nbt1360

15. Woese CR, Kandler O, Weслиis ML. Towards a natural system of organisms. Proposal for the domains Archaea and Bacteria. Proc Natl Acad Sci USA 1990; 87:4576-4579. PubMed http://dx.doi.org/10.1073/pnas.87.12.4576

16. Garrity GM, Bell JA, Lilburn T. Phylum XIV. Proteobacteria phyl nov. In: Brenner DJ, Krieg NR, Stanley JT, Garrity GM (eds), Bergey’s Manual of Systematic Bacteriology, second edition. Vol. 2 (The Proteobacteria), part B (The Gammaproteobacteria), Springer, New York, 2005, p. 1.

17. Garrity GM, Bell JA, Lilburn T. Class I. Alphaproteobacteria class. nov. In: Brenner DJ, Krieg NR, Stanley JT, Garrity GM (eds), Bergey’s Manual of Systematic Bacteriology, second edition. Vol. 2 (The Proteobacteria), part C (The Alpha-, Beta-, Delta-, and Epsilonproteobacteria), Springer, New York, 2005, p. 161.

18. Validation List No. 107. List of new names and new combinations previously effectively, but not validly, published. Int J Syst Evol Microbiol 2006; 56:1-6. PubMed http://dx.doi.org/10.1099/ijsv.0.64188-0

19. Garrity GM, Bell JA, Lilburn T. Order III. Rhodobacterales ord. nov. In: Brenner DJ, Krieg NR, Staley JT, Garrity GM (eds), Bergey’s Manual of Systematic Bacteriology, second edition. Vol. 2 (The Proteobacteria), part C (The Alpha-, Beta-, Delta-, and Epsilonproteobacteria), Springer, New York, 2005, p. 161.

20. Garrity GM, Bell JA, Lilburn T. Family I. Rhodobacteraceae fam. nov. In: Brenner DJ, Krieg NR, Staley JT, Garrity GM (eds), Bergey’s Manual of Systematic Bacteriology, second edition. Vol. 2 (The Proteobacteria), part C (The Alpha-, Beta-, Delta-, and Epsilonproteobacteria), Springer, New York, 2005, p. 161.

21. BAuA. Classification of Bacteria and Archaea in risk groups. TRBA 2010; 466:93.

22. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25:25-29. PubMed http://dx.doi.org/10.1038/75556

23. Vaas LAI, Sikorski J, Michael V, Goker M, Klenk HP. Visualization and curve-parameter estimation strategies for efficient exploration of phenotype microarray kinetics. PLoS ONE 2012; 7:e34846.

http://standardsingenomics.org
Rubellimicrobium thermophilum DSM 16684T

PubMed
http://dx.doi.org/10.1371/journal.pone.0034846

24. Vaas LA, Sikorski J, Hofner B, Fiebig A, Buddruhs N, Klenk HP, Gökler M. ompt: an R package for analyzing OmniLog® phenoypotype microarray date. Bioinformatics 2013; 29:1823-1824. PubMed http://dx.doi.org/10.1093/bioinformatics/btt291

25. Bochner BR. Global phenotypic characterization of bacteria. FEMS Microbiol Rev 2009; 33:191-205. PubMed http://dx.doi.org/10.1111/j.1574-6976.2008.00149.x

26. Gökler M, Klenk HP. Phylogenom-driven target selection for genome-sequencing (and other) projects. Stand Genomic Sci 2013; 8:360-374. http://dx.doi.org/10.4056/sigs.3446951

27. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, et al. A phylogenom-driven Genomic Encyclopaedia of Bacteria and Archaea. Nature 2009; 462:1056-1060. PubMed http://dx.doi.org/10.1038/nature08656

28. Markowitz VM, Ivanova NN, Chen IMA, Chu K, Kyrpides NC. IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics 2009; 25:2271-2278. PubMed http://dx.doi.org/10.1093/bioinformatics/btp393

29. List of growth media used at the DSMZ: http://www.dsmz.de/catalogues/cataloque-microorganisms/culture-technology/list-of-media-for-microorganisms.html

30. Gemeinholzer B, Dröge G, Zetzsche H, Haszprunar G, Klenk HP, Güntsch A, Berendsohn WG, Wägele JW. The DNA Bank Network: the start from a German initiative. Biopreserv Biobank 2011; 9:51-55. http://dx.doi.org/10.1089/bio.2010.0029

31. Aronesty E. ea-utils: Command-line tools for processing biological sequencing data. 2011; http://code.google.com/p/ea-utils.

32. Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. Genome Biol 2010; 11:R116. PubMed http://dx.doi.org/10.1186/gb-2010-11-11-r116

33. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008; 18:821-829. PubMed http://dx.doi.org/10.1101/gr.074922.107

34. Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. Genome Res 1998; 8:195-202. PubMed http://dx.doi.org/10.1101/gr.8.3.195

35. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 2010; 11:119. PubMed http://dx.doi.org/10.1186/1471-2105-11-119

36. Mavromatis K, Ivanova NN, Chen IM, Szeto E, Markowitz VM, Kyrpides NC. The DOE-JGI Standard operating procedure for the annotations of microbial genomes. Stand Genomic Sci 2009; 1:63-67. PubMed http://dx.doi.org/10.4056/sigs.632

37. Finn DR, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. Web Server Issue 2011; 39:W29-W37.

38. Lowe TM, Eddy SR. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. Nucleic Acids Res 1997; 25:955-964.

39. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: Inference of RNA alignments. Bioinformatics 2009; 25:1335-1337. PubMed http://dx.doi.org/10.1093/bioinformatics/btp157

40. Markowitz VM, Ivanova NN, Chen IMA, Chu K, Kyrpides NC. IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics 2009; 25:2271-2278. PubMed http://dx.doi.org/10.1093/bioinformatics/btp393

41. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics 2007; 8:209. PubMed http://dx.doi.org/10.1186/1471-2105-8-209

42. Edgar RC. PILER-CR: Fast and accurate identification of CRISPR repeats. BMC Bioinformatics 2007; 8:18. PubMed http://dx.doi.org/10.1186/1471-2105-8-18

43. Petersen J, Frank O, Gökler M, Pradella S. Extrachromosomal, extraordinary and essential--the plasmids of the Roseobacter clade. Appl Microbiol Biotechnol 2013; 97:2805-2815. PubMed http://dx.doi.org/10.1007/s00253-013-4746-8

44. Ochman H, Lawrence J, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature 2000; 405:299-304. PubMed http://dx.doi.org/10.1038/35012500

45. Friedrich CG, Bardischewsky F, Rother D, Quentmeier A, Fischer J. Prokaryotic sulfur oxidation. Curr Opin Microbiol 2005; 8:253-259. PubMed http://dx.doi.org/10.1016/j.mib.2005.04.005