Marine Genomics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Marine Genomics

journal homepage: www.elsevier.com/locate/margen

Method paper

Draft genome sequence of *Shimia marina* CECT 7688^T

Lidia Rodrigo-Torres, María J. Pujalte, David R. Arahal *

Departamento de Microbiología y Ecología and Colección Española de Cultivos Tipo (CECT), Universidad de Valencia, Valencia, Spain

ARTICLE INFO

Article history:
Received 30 November 2015
Received in revised form 27 January 2016
Accepted 27 January 2016
Available online xxxx

Keywords:
Rhodobacteraceae
Strictly aerobic
Moderately halophilic
Motile
Biofilm formation

ABSTRACT

Shimia marina is a member of the family *Rhodobacteraceae* described in 2006. Strain CL-TA03^T (=CECT 7688^T) was isolated from a biofilm formed on an acrylic slide submerged in surface water in a coastal fish farm in Tongyeong, Korea. Here we report the draft genome sequence and annotation of *S. marina* CECT 7688^T which is composed by 4,001,860 bp arranged in 45 scaffolds with a G + C content of 57.4%, 3878 protein coding genes, 40 tRNA genes, 4 rRNA genes and 1 repeat region. An overview of annotated genes revealed diverse genes encoding for exopolysaccharide and capsular biosynthesis enzymes, secondary metabolite biosynthesis enzymes, multiple antibiotic and metal resistance and the ability for degrading aromatic compounds.

1. Introduction

Shimia is a genus of the family *Rhodobacteraceae*, order *Rhodobacterales* within the class *Alphaproteobacteria*. It also belongs to the so-called *Roseobacter* group, which contains several of the more ubiquitous and predominant components of marine bacteria and plays an important role in biochemical global processes (Pujalte et al., 2014). At the time of writing, 376 type strains of species belonging to the family *Rhodobacteraceae* were present in NCBI Taxonomy Database, 87 of them had an assembly at NCBI Assembly Database but only a member of the genus *Shimia* (*Shimia* sp. SK013) had an assembly available.

Shimia marina, the type species of the genus, was described by Choi and Cho (2006), and since four other species with validly published name have been described in this genus: *Shimia isoporae* (Chen et al., 2011), *Shimia haliotis* (Hyun et al., 2013), *Shimia biformata* (Hameed et al., 2013) and *Shimia sagamensis* (Nogi et al., 2015).

S. marina CL-TA03^T was isolated from a biofilm in a coastal fish farm in Tongyeong, Korea. The 16S rRNA gene sequence similarity between *S. marina* and other *Shimia* species (calculated in www.ezbiocloud.net/eztaxon) is as follows: 97.99% to *S. isoporae*, 97.83% to *S. haliotis*, 97.77% to *S. sagamensis* and 97.15% to *S. biformata*. Other closely related species are *Thalasobius aestuarii* (96.53%), *Ruegeria faecimaris* (96.49%) and *Nautella italica* (96.41%).

The type strains of the five species of genus *Shimia* were isolated from different environments, such as reef-building coral, an abalone gut, surface water and marine sediments, showing the diverse niches they are able to colonize (Choi and Cho, 2006; Chen et al., 2011; Hyun et al., 2013; Hameed et al., 2013; Nogi et al., 2015). Metagenomic studies about coral white diseases (Godwin et al., 2012; Séré et al., 2013) placed *S. marina* as a bacterium possibly related to that pathogenicity, but this condition has not been proved experimentally.

This strain was selected for genomic *de novo* sequencing to provide a reference genome for taxonomic purposes and to study the potential biological abilities of the organism. Here we present the genome properties of *S. marina* CECT 7688^T together with its annotation results. As a biofilm former and potential coral pathogen, phenotype related candidate genes were searched represented. The source of *S. marina* CECT 7688^T and its sequencing information are shown in Table 1.

2. Data description

Genome of *S. marina* CECT 7688^T was sequenced at Central Service of Support to Experimental Research (SCSIE) of the University of Valencia (Valencia, Spain) using an Illumina Miseq technology with 2 × 250 bp paired-end reads.

Reads were assembled using Velvet 1.0.0 de novo assembler (Zerbino and Birney, 2008) included as an application in BaseSpace Genomics Cloud Computing (https//:basespace.illumina.com).

The draft genome was annotated using Prokka (Seemann, 2014) an open source software tool, within Galaxy Orione Server (Cuccuru et al., 2014), and using RAST v2.0 (Rapid Annotation using Subsystem Technology) (Aziz et al., 2008). Prokka annotation was used for further analysis using the following web tools: Pfam domains were predicted with NCBI Batch CD-Search Tool (Marchler-Bauer et al., 2015) using default

Abbreviations: CPBP, Capsule polysaccharide biosynthesis protein; Kdo, 3-Deoxy-D-manno-octulosonic acid; EPS, Exopolysaccharide; TOMM, Thiazole-oxazole modified microcin; CRISPR, Clustered regularly interspaced short palindromic repeats; CHOS, Chitooligosaccharides.

* Corresponding author.
E-mail addresses: lidia.rodrigo@uv.es (L. Rodrigo-Torres), maria.j.pujalte@uv.es (M.J. Pujalte), arahal@uv.es (D.R. Arahal).

http://dx.doi.org/10.1016/j.margen.2016.01.006
1874-7787/

Table 1
General features of *S. marina* CECT 7688^T genome sequencing project.

Feature	Value
Organism	*Shimia marina* CECT 7688^T
INSDC accession number	CYPW01000001–CYPW01000045
Investigation type	Bacteria
Project name	Whole genome sequence of *Shimia marina* CECT 7688^T
Geographic location (latitude and longitude)	34,8429° N 128,4485° E
Geographic location (country and/or sea region)	Korea
Collection date	2002
Environment (biome)	Sea water
Environment (feature)	Acrylic slide submerged in surface sea water on a coastal fish farm
Environment (material)	Biofilm
Environment package	Water
Depth	0 m
Isolation and growth condition	http://dx.doi.org/10.1099/ijs.0.64235-0
Reference for biomaterial	http://dx.doi.org/10.1099/ijs.0.64235-0
Propagation	Periodic transfer
Sequencing method/technology	Illumina MiSeq platform
Ploidy	PATO: 0001375
Number of replicons	1
Estimated size	4.0 Mb
Assembly method	Velvet 1.0.0
Finishing strategy	Draft, 35×, 45 scaffolds
Assembly coverage (final genome coverage)	19×

Table 2
Genomic features of *S. marina* CECT 7688^T (annotation by Prokka).

Feature				Value
Genome size (bp)				4,001,860
DNA coding (bp)				3,662,396
DNA G + C (bp)				2,297,068
DNA scaffolds				45
Total genes				3922
Protein coding genes				3878
RNA genes				44
Genes with function prediction				3012
Genes assigned to COGs				3009
	Category	Value	%age	Description
	J	164	4.2	Translation, ribosomal structure and biogenesis
	A	0	0.0	RNA processing and modification
	K	243	6.2	Transcription
	L	129	3.3	Replication, recombination and repair
	B	3	0.1	Chromatin structure and dynamics
	D	32	0.8	Cell cycle control, cell division, chromosome partitioning
	V	41	1.0	Defense mechanisms
	T	184	4.7	Signal transduction mechanisms
	M	201	5.1	Cell wall/membrane biogenesis
	N	48	1.2	Cell motility
	U	83	2.1	Intracellular trafficking and secretion
	O	151	3.9	Posttranslational modification, protein turnover, chaperones
	C	218	5.6	Energy production and conversion
	G	152	3.9	Carbohydrate transport and metabolism
	E	387	9.9	Amino acid transport and metabolism
	F	84	2.1	Nucleotide transport and metabolism
	H	145	3.7	Coenzyme transport and metabolism
	I	155	4.0	Lipid transport and metabolism
	P	190	4.8	Inorganic ion transport and metabolism
	Q	106	2.7	Secondary metabolite biosynthesis, transport and catabolism
	R	408	10.4	General function prediction only
	S	300	7.6	Function unknown
	–	913	23.3	Not in COGs
Genes with Pfam domains				2474
Genes with signal peptides				381
Genes with transmembrane helices				889
CRISPR repeats				1

parameters. Signal peptides were searched using SignalP 4.1 Server (Petersen et al., 2011). Transmembrane helix domains were predicted through TMHMM server v.2.0 (Krogh et al., 2001). CRISPR repeats were examined by CRISPRFinder (Grissa et al., 2007). Protein coding genes were analyzed for COG functional annotation using WebMGA server (Wu et al., 2011). Secondary metabolites were predicted using antiSMASH 2.0 (Medema et al., 2011). Graphical map was obtained using CG view Server (Grant and Stothard, 2008) with gbk format and modified gff format annotation files.

The sequencing experiment produced 552,734 reads comprising 137,470,766 bp, with 35× coverage. Trimming or filtering was not considered necessary, so all reads were used for the assembly that resulted into 45 scaffolds containing 4,001,860 bp and a G + C content of 57.4%. Prokka predicted 3878 protein coding genes, 4 rRNA genes, 40 tRNA genes and a repeat region. 910 protein coding genes were assigned to hypothetical proteins (23.5% of total protein coding genes). RAST predicted 3901 protein coding genes, 3 rRNA genes, 44 tRNA. Genomic features and COG categories are summarized in Table 2. A graphical map with annotated genes and COG categories is shown in Fig. 1.

Capsule and exopolysaccharide synthesis genes play an important role in biofilm formation together with other genes (Enos-Berlage et al., 2005). The ability of *S. marina* to produce biofilms was reported (Choi and Cho, 2006), thus, capsule and exopolysaccharides synthesis and regulation genes were explored in the draft genome. Seven genes coding for capsule polysaccharide biosynthesis proteins (CPBP) were found along the genome, some of them were accompanied by Vi polysaccharide export inner protein VexB, D genes and Wza polysaccharide export protein gene. Five *kps* genes are also annotated in this draft genome, three *kpsT* genes coding for capsular polysaccharide transport system ATP-binding proteins, two *kpsM* coding for capsular polysaccharide transport system permease proteins and one *kpsU* coding for 3-deoxy-manno-octulosonate-cytidyl transferase. These genes are positioned closely to CPBP coding genes and *vexB*, *D* coding genes. The genome also harbors *kds* genes coding for 2-dehydro-3-deoxyphosphooctonate aldolase (*kdsA*, two copies), 3-deoxy-manno-octulosonate-cytidyl transferase (*kdsB*) and arabinose-5-phosphate isomerase (*kdsD*, four copies) and also three copies of *waa*A gene, encoding a Kdo transferase. *kds* genes are involved in Kdo synthesis, the linker between capsular polysaccharides and membranes. *kps* and *kds* genes are the main genes involved in capsular biosynthesis and export (Willis and Whitfield, 2013). RAST predicted 11 genes related to exopolysaccharide biosynthesis (also present in the annotation by Prokka): five glycosyl transferases, group 2 family protein coding genes, two glycosyl transferases, group 1 family protein genes, and three genes also involved in exopolysaccharide (EPS) biosynthesis (an EPS transport protein, an EPS domain protein and undecaprenyl-phosphate galactose phosphotransferase). Two genes were also predicted coding for glucans biosynthesis protein G (*mdoG* gene) and glucans biosynthesis glycosyl transferase H (*opgH* gene) both with signal peptides suggesting an extracellular polysaccharide biosynthesis activity. Besides, genes for alginate biosynthesis (*algA*,*C*,*D*) and cellulose biosynthesis (*bcsA*,*C*) and a *exoI* gene involved in succinoglycan biosynthesis were also annotated in this draft genome. All these findings could explain this strain ability to produce biofilms and colonize surfaces.

Previous studies reported the cell ability of *S. marina* CECT 7688^T to join in small chains (Choi and Cho, 2006). Here we report two genes coding for morphology and auto-aggregation proteins and a gene coding for an H-type lectin domain protein involved in self-non-self-recognition of cells through binding with carbohydrates in the genome, this lectin is sometimes found in association with C-term domain of coagulation factor F5/8, so can be consider an hemagglutinin. Ability to cell invasion or extracellular hydrolases production was also examined. Two invasion proteins B involved in pathogenesis were found. Extracellular proteases were also annotated as a putative Zn-dependent protease with a signal peptide together with an extracellular serine protease