Available online at www.sciencedirect.com
ScienceDirect
Genomics 91 (2008) 335–346
GENOMICS
www.elsevier.com/locate/ygeno

Evolution in the laboratory: The genome of *Halobacterium salinarum* strain R1 compared to that of strain NRC-1 ☆

F. Pfeiffer [a], S.C. Schuster [a,1], A. Broicher [a], M. Falb [a,2], P. Palm [a], K. Rodewald [a], A. Ruepp [b,3], J. Soppa [c], J. Tittor [a], D. Oesterhelt [a,*]

[a] *Department of Membrane Biochemistry, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany*
[b] *Department of Molecular Structural Biology, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany*
[c] *Institute for Molecular Biosciences, Goethe University, Frankfurt, Germany*

Received 2 August 2007; accepted 2 January 2008
Available online 3 March 2008

Abstract

We report the sequence of the *Halobacterium salinarum* strain R1 chromosome and its four megaplasmids. Our set of protein-coding genes is supported by extensive proteomic and sequence homology data. The structures of the plasmids, which show three large-scale duplications (adding up to 100 kb), were unequivocally confirmed by cosmid analysis. The chromosome of strain R1 is completely colinear and virtually identical to that of strain NRC-1. Correlation of the plasmid sequences revealed 210 kb of sequence that occurs only in strain R1. The remaining 350 kb shows virtual sequence identity in the two strains. Nevertheless, the number and overall structure of the plasmids are largely incompatible. Also, 20% of the protein sequences differ despite the near identity at the DNA sequence level. Finally, we report genome-wide mobility data for insertion sequences from which we conclude that strains R1 and NRC-1 originate from the same natural isolate. This exemplifies evolution in the laboratory.

Keywords: Archaea; Comparative genomics; Genome sequence; Halophilicity; *Halobacterium salinarum;* Insertion sequence

Halobacterium salinarum has been intensively studied during the past decades, through which our understanding of various biological processes such as energy metabolism, environmental response, gene regulation, and the archaeal cell cycle has been greatly increased (for recent reviews see [1,2]). A microorganism corresponding to the description of *Hbt. salinarum* was isolated from salted fish more than 80 years ago. Since then many haloarchaeal species have been isolated, which, after considerable renaming, are currently grouped into 25 genera. Several years ago, it was decided that the species *Halobacterium salinarum, Halobacterium halobium,* and *Halobacterium cutirubrum* are so similar that they should be regarded as strains of one species named *Halobacterium salinarum* [3]. *Hbt. salinarum* shows very high genetic variability [4,5] that was attributed to the large number of insertion sequences (ISH elements) (for review see [6]).

The active and successful research of several laboratories has led to the initiation of two independent genome sequencing initiatives, one for *Halobacterium* sp. NRC-1, the other for *Hbt. salinarum* strain R1. When the sequence of *Hbt.* sp. NRC-1 genome appeared [7], the chromosome of *Hbt. salinarum* strain R1, which is reported here, was complete and being annotated. The assembly of the smaller replicons had not been finished at that time due to major problems caused by large-scale duplications and the high number of insertion elements.

☆ Sequence data from this article have been deposited with the DDBJ/EMBL/GenBank Data Libraries under Accession Nos. AM774415 (chromosome), AM774416 (pHS1), AM774417 (pHS2), AM774418 (pHS3), and AM774419 (pHS4).

* Corresponding author. Fax: +49 89 8578 3557.
E-mail address: oesterhe@biochem.mpg.de (D. Oesterhelt).
[1] Current address: Center for Comparative Genomics and Bioinformatics, Center for Infectious Disease Dynamics, Penn State University, University Park, PA 16802, USA.
[2] Current address: Sanofi-Aventis Deutschland GmbH, Industriepark Hoechst, 65926 Frankfurt am Main, Germany.
[3] Current address: Institut für Bioinformatik, GSF-Forschungszentrum für Umwelt und Gesundheit, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany.

0888-7543/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.01.001

Genome sequences from different strains of the same organism may vary significantly; for example, there is a 5–7% sequence deviation between the two *Helicobacter pylori* strains J99 and 26695 at the amino acid level [8]. In contrast, the number of sequence differences for *Hbt. salinarum* strains R1 and NRC-1 was vanishingly small, although NRC-1 had been published as if it were a distinct species. However, since then it has been reclassified as a strain of *Hbt. salinarum* [9]. It was also evident that the protein-coding gene sets differed considerably, although they were derived from nearly identical DNA sequences. This inconsistency is due to the high error rate of automatic gene finder programs for GC-rich genomes, especially with respect to start codon selection [10–13]. The correctness of the protein-coding gene set can be increased by experimental or bioinformatic analysis, for example, integration of proteomic data or evaluation of sequence homology data.

The genome of *Hbt. salinarum* strain R1 has been publicly available since 2002 through the HaloLex Web portal (www.halolex.mpg.de). In this publication, we report the sequences of the chromosome and the four plasmids of *Hbt. salinarum* R1, including a high-quality annotation of protein-coding genes that is well supported by proteomic experiments and sequence homology data. Comparison of strains R1 and NRC-1 revealed genome-scale data on sequence differences and ISH element mobility. From these results, we conclude that both strains originate from the same natural isolate and have since diverged in the laboratory.

Results and discussion

The genome of Halobacterium salinarum strain R1

The genome of *Hbt. salinarum* strain R1 (DSM 671) comprises a single major chromosome of 2 Mb with a very high GC content of 68.0% and four megaplasmids (pHS1 to pHS4) that have a total of 667,814 bp and a lower GC content of 58.8% (Table 1). The genome contains 2878 protein-coding genes (Supplementary Table S1).

Table 1
Basic characteristics of the replicons from *Halobacterium salinarum* strain R1 (DSM 671)

	Chromosome	pHS1	pHS2	pHS3	pHS4
Length (bp)	2,000,962	147,625	194,963	284,332	40,894
GC content	68.0%	57.4%	58.6%	59.8%	57.9%
% coding (protein + RNA)	91.5%	83.6%	80.4%	84.8%	82.9%
Encoded proteins	2,132	172	230	305	39
Average protein length (amino acids)	284	239	226	263	289
Encoded stable RNAs	52	—	—	—	—

	Plasmids (total)	Genome (total)
Length (bp)	667,814	2,668,776
GC content	58.8	—
Encoded proteins	746	2,878

The major chromosome

The chromosome is densely packed with 2132 protein-coding genes and genes for 52 stable RNAs. Together, these cover 91.5% of the chromosomal sequence.

The replication origin is delineated by a 31-bp inverted repeat that is flanked on one side by a Cdc6 homolog (*orc7,* OE4380F) [14]. On the other side the repeat is flanked by a set of three genes (OE4377R, OE4376R, OE4374R) that are also found adjacent to the replication origin in *Natronomonas pharaonis* [10], *Haloquadratum walsbyi* [15], and *Haloarcula marismortui* [16].These genes have no known function, but the positional conservation observed in all halophiles may indicate an involvement of the three proteins in the replication process.

The chromosome contains a 60-kb insertion with plasmid-like characteristics: (a) a reduced GC content of 56% and (b) a reduced proteomic protein identification rate [17]. This insertion corresponds to the previously described "AT-rich island" [18].

Plasmid pHS3

Plasmid pHS3 is 284 kb long and codes for a number of essential and important proteins, most of them in or adjacent to a 67-kb region (Fig. 1A) with chromosome-like features (increased GC content of 65%, increased proteomic identification ratio) [17]. The most prominent proteins are (a) the only arginine-tRNA ligase of *Hbt. salinarum* (argS); (b) the two subunits of aspartate carbamoyltransferase (pyrBI), which catalyzes the first step of pyrimidine biosynthesis; (c) all enzymes of the arginine deiminase pathway for arginine fermentation (arc operon) [19], including the arginine/ornithine antiporter (OE5204R, unpublished data); and (d) the only catalase (perA) which is involved in protection against oxidative stress. Thus, because pHS3 encodes essential proteins it may be considered a second chromosome rather than a plasmid.

Plasmids pHS1, pHS2, and pHS4

The three plasmids pHS1, pHS2, and pHS4 are related to each other through their large-scale duplications (Fig. 1A). Regions that are labeled by the same letter show (near) sequence identity. The regions are listed in Supplementary Table S2.

The 147-kb plasmid pHS1 corresponds to the previously described plasmid pHH1 [20] and carries a high number of ISH elements. Only one-third (48 kb) of the pHS1 sequence is specific for this plasmid (regions B, G, L, M). The other two-thirds (99 kb) represent three large-scale duplications (Fig. 1A): a perfect 61.8-kb duplication of pHS2 (regions C, D, F), a perfect 30.0-kb duplication of pHS4 (region K) adjacent to an imperfect duplication with 98.5% sequence identity over 7.3 kb (region H).

Plasmid pHS2 is 195 kb long, of which 61.8 kb are duplicated on pHS1 and the remaining 133 kb are specific to pHS2.

Plasmid pHS4 with 41 kb was not detected until the late stages of genome assembly since 92% of it represents sequences duplicated on pHS1, while only 8% of the sequence (3.4 kb, region Y) is specific to pHS4. There is a perfect 30.0-kb duplication (region K) and an adjacent imperfect duplication of 7316 matching bases with only 1.5% sequence difference (region H). An additional difference is the presence of two ISH elements, which occur only on pHS1 in region H.