The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2

Qunxin She[a,b], Rama K. Singh[b,c], Fabrice Confalonieri[b,d], Yvan Zivanovic[b,d], Ghislaine Allard[e], Mariana J. Awayez[a], Christina C.-Y. Chan-Weiher[e], Ib Groth Clausen[f], Bruce A. Curtis[c], Anick De Moors[e], Gael Erauso[g], Cynthia Fletcher[c], Paul M. K. Gordon[c], Ineke Heikamp-de Jong[g], Alex C. Jeffries[c], Catherine J. Kozera[c], Nadine Medina[d], Xu Peng[a], Hoa Phan Thi-Ngoc[a], Peter Redder[a], Margaret E. Schenk[h], Cynthia Theriault[c], Niels Tolstrup[f], Robert L. Charlebois[e], W. Ford Doolittle[h], Michel Duguet[d], Terry Gaasterland[i], Roger A. Garrett[a,j], Mark A. Ragan[c,k], Christoph W. Sensen[c,l], and John Van der Oost[g]

[a]Microbial Genome Group, Institute of Molecular Biology, University of Copenhagen, Sølvgade 83H, DK-1307 Copenhagen, Denmark; [c]National Research Council of Canada, Institute for Marine Biosciences, 1411 Oxford Street, Halifax, NS, Canada B3H 3Z1; [d]Université Paris-Sud, Institut de Génétique et Microbiologie, 15, Rue Georges Clemenceau, Bâtiment 400, FR-91405 Orsay Cedex, France; [e]University of Ottawa, Department of Biology, 30 Marie Curie, Ottawa, ON, Canada K1N 6N5; [f]Novozymes, Novo Alle, DK-2880 Bagsværd, Denmark; [g]Wageningen University, Laboratory of Microbiology, Hesselink van Suchtelenweg 4, NL-6703 CT, Wageningen, The Netherlands; [h]Dalhousie University, Department of Biochemistry, Sir Charles Tupper Medical Building, Halifax, NS, Canada B3H 4H7; [i]The Rockefeller University, 1230 York Avenue, New York, NY 10021; and [k]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Qld 4072, Australia

Communicated by Carl R. Woese, University of Illinois at Urbana–Champaign, Urbana, IL, May 4, 2001 (received for review February 15, 2001)

The genome of the crenarchaeon *Sulfolobus solfataricus P2* contains 2,992,245 bp on a single chromosome and encodes 2,977 proteins and many RNAs. One-third of the encoded proteins have no detectable homologs in other sequenced genomes. Moreover, 40% appear to be archaeal-specific, and only 12% and 2.3% are shared exclusively with bacteria and eukarya, respectively. The genome shows a high level of plasticity with 200 diverse insertion sequence elements, many putative nonautonomous mobile elements, and evidence of integrase-mediated insertion events. There are also long clusters of regularly spaced tandem repeats. Different transfer systems are used for the uptake of inorganic and organic solutes, and a wealth of intracellular and extracellular proteases, sugar, and sulfur metabolizing enzymes are encoded, as well as enzymes of the central metabolic pathways and motility proteins. The major metabolic electron carrier is not NADH as in bacteria and eukarya but probably ferredoxin. The essential components required for DNA replication, DNA repair and recombination, the cell cycle, transcriptional initiation and translation, but not DNA folding, show a strong eukaryal character with many archaeal-specific features. The results illustrate major differences between crenarchaea and euryarchaea, especially for their DNA replication mechanism and cell cycle processes and their translational apparatus.

CELL BIOLOGY

Sulfolobus solfataricus is an aerobic crenarchaeon that grows optimally at 80°C and pH 2- to 4-metabolizing sulfur (1). It is the most widely studied organism of the crenarchaeal branch of the Archaea and is a model for research on mechanisms of DNA replication, the cell cycle, chromosomal integration, transcription, RNA processing, and translation (2). Several extrachromosomal elements of *Sulfolobus* have been characterized, including conjugative plasmids, novel plasmids, and four virus families (reviewed in ref. 3). In addition, first-generation vectors and knockout mutants have been produced for genetic studies (reviewed in ref. 4).

Genome sequencing was a joint Canadian–European Union project, and the emerging sequences have facilitated many studies on *Sulfolobus* cell biology (http://www-archbac.u-psud.fr/projects/sulfolobus/).

Genome Sequencing, Organization, and Annotation

The genome was cloned and mapped by using cosmid, λ, and bacterial artificial chromosome libraries and sequenced (5–8). Some regions were checked by generating PCR fragments and sequencing them from both ends. Around 30,000 reads were produced for the complete sequence, and about 8,000 were custom primer walking reactions. The overall average coverage was over 5-fold. Differences between clones arose mainly because insertion sequence (IS) elements moved during cell culture. Genes were identified and functions assigned essentially as described (9). The final sequence corresponds to a single chromosome of 2,992,245 bp, within the original estimate of 3 (±0.1) Mb (10). Three thousand thirty-two genes were identified. About 11% of the genome consists of putatively mobile elements.

Many ORFs appear to be *Sulfolobus*- and/or archaea-specific. Seven hundred forty-three ORFs are exclusive to *S. solfataricus* (BLASTP $e < 10^{-5}$) and 1,602 ORFs yield matches among euryarchaea; 193 of the latter yield no match outside archaea. 1,030 ORFs produce matches in the *Aeropyrum pernix* genome (at $e < 10^{-10}$), and 45 of these match *A. pernix*, exclusively, in the GenBank/European Molecular Biology Laboratory database. Three hundred fifty seven *Sulfolobus* ORFs produce a match among bacteria but not eukarya, whereas 67 match eukarya but not bacteria at inclusion and exclusion thresholds $e = 10^{-10}$ and 10^{-5}, respectively (11, 12). An additional 701 ORFs give a match in both domains.

Fifty-two major gene families were identified, ranging in size from 2 to 26 members ($e < 10^{-5}$) (13). The largest family is involved in fatty acid biosynthesis, especially genes encoding acetyl-CoA synthetases. Others include alcohol and other dehydrogenases (17 members) and ATP-binding subunits of ABC transporters (19 members). Gene order most closely resembles that of other archaea. If we define conserved ORF clusters as supersets of pairs of neighboring ORFs (separated by no more than two other initiation codons) that match neighboring ORFs in a second genome, then at BLASTP $e < 10^{-5}$ the *Sulfolobus* genome shares with 8 other archaeal genomes 57–140 (mean 87) clusters of 2.67–3.20 (mean 2.85) ORFs, whereas it shares 7–63 (mean 28) clusters of 2.06–2.91 (mean 2.45) ORFs with 28 bacterial genomes, and only 4 clusters of 2 ORFs with the yeast genome. One hundred forty clusters are shared with the cren-

Abbreviations: IS element, insertion sequence element; LCTR, large clusters of 20-nt tandem repeat sequences.

Data deposition: The sequence reported in this paper has been deposited in the European Molecular Biology Laboratory/GenBank database (accession no. AE006641).

[b]Q.S., R.K.S., F.C., and Y.Z. contributed equally to this work.

[j]To whom reprint requests should be addressed. E-mail: garrett@mermaid.molbio.ku.dk.

[l]Present address: University of Calgary, Department of Biochemistry and Molecular Biology, 3330 Hospital Drive N.W., Calgary, AB, Canada T2N 4N1.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

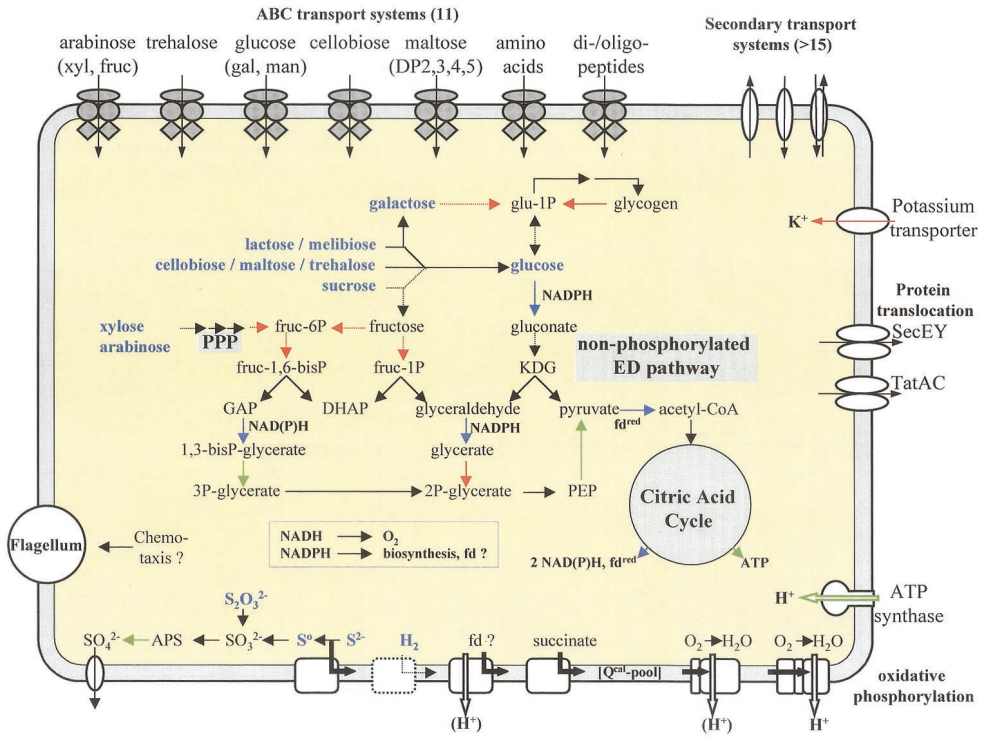

Fig. 1. Overview of metabolism and transport in *S. solfataricus*. Pathways for energy production and carbohydrate catabolism are shown, and extracellular enzymes that hydrolyze polymers (proteases, glycosyl hydrolases) are not shown. Arrows denote the following reactions: chemical conversion (black), energy consuming (red), energy yielding (green), redox (blue), respiratory electron transfer (solid black), proton export (black, open), and import (green, open). Conversions that were anticipated but for which no gene was detected are shown as broken arrows. Eleven operons encoding ABC transport systems are present, and those with established substrate specificity are depicted (xyl, xylose; fruc, fructose; glu, glucose; gal, galactose; man, mannose. DP, degree of polymerization) (17). At least 15 secondary transporters (permeases) are present (symport and antiport). Carbohydrates that are imported and/or support growth of *S. solfataricus* are in blue. All but one of the genes encoding enzymes of the nonphosphorylated Entner–Doudoroff (ED) pathway was identified (KDG, 2-keto-3-deoxygluconate; GAP, glyceraldehyde-3P; DHAP, dihydroxyacetone-P; PEP, phosphoenolpyruvate). Only two genes involved in the pentose phosphate pathway (PPP) were identified (see text). All citric acid cycle enzymes are encoded. Several components of the aerobic respiratory network are identified that are involved in (*i*) reduction of the caldariella-quinone (Q^{cal}) pool: a putative ferredoxin dehydrogenase (see text), succinate dehydrogenase, and (*ii*) oxidation of the Q^{cal} pool: SoxABCD and SoxM terminal oxidases; an ATP synthase converts the proton gradient into ATP; alternative electron donors (in blue) are hydrogen (H_2) and sulfide (S^{2-}), reducing the Q^{cal} pool via hydrogenase and sulfide reductase, respectively. Elemental sulfur and thiosulphate are completely converted to sulfate (APS, adenylylsulfate); some flagellar components are present (see text). Both Sec/signal recognition particle-type and Tat-type protein translocation systems are present.

archaeon *A. pernix*, more than with any euryarchaeon (range 57–104).

Annotation was performed by searching sequence databases of genomes and metabolic pathway enzymes at http://www-archbac.u-psud.fr/projects/sulfolobus/in combination with Magpie (14) and the genome is in GenBank/European Molecular Biology Laboratory (accession no. AE006641).

Physiology and Growth

S. solfataricus grows optimally at pH 2–4 but maintains its cytoplasmic pH at about 6.5 by generating a large pH gradient across the cytoplasmic membrane (15). Typically, for acidophiles, the overall proton motive force is reduced by an unusual charge distribution across its membrane. Thus, it is positive inside (16), at least partly because of active uptake of potassium ions by a predicted Trk-like potassium transporter (Sso1757). The large pH gradient is exploited to generate energy via ATP synthase (Fig. 1). Eight ATPase subunits are encoded within the juxtapositioned transcriptional units *atpI* and *atpFEABDGK* (Sso0559–0567) (7). They are A type (A_0A_1) and differ strongly from V- and F-type ATPases common to bacteria, eukarya, and some euryarchaea.

The proton gradient is considered to drive the uptake of inorganic and organic solutes, including sugars and peptides, via "secondary transport systems." At least 15 of the necessary proteins are encoded. High-affinity uptake via ABC-type transporters is also common: 11 operons encode a membrane-anchored extracytoplasmic-binding protein, one or two membrane-embedded permeases, and one or two cytoplasmic ATPases (ref. 17; Fig. 1). This probably reflects periodic low concentrations of organic substrates, such as carbohydrates and peptides, in natural habitats. No phosphotransferase system-type transporters were found as for other archaea.

Proteolytic growth of *S. solfataricus*, as for the euryarchaeon *Thermoplasma acidophilum* (18), proceeds via concerted action of extracellular (Sso2045, 1141, 2551) and intracellular proteasome subunits—Sso0277, 0738, and 0766, tricorn protease, Sso2098 and tricorn cofactors—Sso3115, 2154, 2675) proteases. Although *S. solfataricus* grows on starch, no genes encoding extracellular α-amylases/pullulanases were found, suggesting that starch dissociates at pH 3 and 80°C. Three putative extracellular cellulases (Sso1354, 1949, 2534) may facilitate heterotrophic growth on more stable β-linked glucan polymers. Two gene clusters encode enzymes responsible for intracellular synthesis and degradation of trehalose (Sso2093–2095) and glycogen (Sso0987–0991). Cytoplasmic conversion of oligo- and disaccharides to monosaccharides is catalyzed by glycosyl hydrolases, including β-glycosidase (Sso3019), α-fucosidase (Sso3060), β-xylosidase (Sso3032), α-xylosidase (Sso3022), and β-glucuronidase (Sso3036/11867), the genes of which are partly clustered.

Metabolic Pathways

The central metabolic pathways are a glycolytic pathway, a pentose phosphate pathway, and the citric acid cycle (Fig. 1). Conversion of glucose to pyruvate via the nonphosphorylating Entner–Doudoroff pathway produces no net energy (19). Genes for most enzymes, except gluconate dehydratase, are present (Sso3204, 3197, 3194, 0666, 0913, 0981). Conversion of pentose substrates (xylose, arabinose) is predicted to proceed via the pentose phosphate pathway, or a variant thereof. However, only genes encoding ribose-5-P isomerase (Sso0978) and transketolase (Sso0297 and 0299) are assigned. In contrast, all citric acid cycle genes are present (Sso1077, 1095, 2182, 2356 to 2359, 2482, 2483, 2585, 2589, 2815, 2816, 2863).

It is striking that NAD+ is used rarely as an electron acceptor in some central metabolic redox reactions. Both glucose dehydrogenase and glyceraldehyde dehydrogenase are reported to reduce NADP+ specifically. Moreover, glyceraldehyde-3-phosphate dehydrogenase, isocitrate dehydrogenase, and malate dehydrogenase show a dual cofactor specificity, with a slight preference for NADP+. Oxidative conversion of pyruvate and 2-oxoglutarate proceeds via ferredoxin-dependent oxidoreductases (19).

Minimal reduction of NAD+ may be because of cofactor reoxidation systems, which maintain an intracellular redox balance. Moreover, the respiratory NADH dehydrogenase, an essential component of many bacterial and mitochondrial respiratory chains (Complex I), appears to be absent from archaea. The activity reported for *Sulfolobus acidocaldarius* (20) probably derives from a cytoplasmic flavin-containing NADH oxidase (Sso1900, Sso2025). Genes encoding an NADH dehydrogenase core (NuoBCDHIL, *Escherichia coli* nomenclature; Sso0665, Sso0322–0329) are present, as in the aerobic archaea *A. pernix*, *T. acidophilum*, and *Halobacterium* NRC-1, but genes encoding the subunits involved in NADH binding and oxidation (NuoEFG) are absent. By analogy with related hydrogenase complexes (21), we infer that archaeal complexes oxidize ferredoxin, but that the released electrons pass to the quinone pool that constitutes the respiratory chain. An analogous quinone reductase complex, for which no electron donor was identified, occurs in *Helicobacter pylori* (22).

Apart from the suggested ferredoxin-dependent quinone reductase complex, *S. solfataricus* also possesses a classical succinate dehydrogenase (Sso2356–2359). Moreover, oxidation of sulfide (H_2S, FeS_2) and molecular hydrogen (Knallgas reaction) has been reported for *Sulfolobus* (19, 23). Although a sulfide reductase is encoded (Sso2261), there is no uptake hydrogenase. However, there is a gene cluster encoding a putative formate hydrogen lyase complex (Sso1020–1029) that could substitute for an uptake hydrogenase. Caldariella quinol is oxidized by two archaeal-specific cytochrome complexes, SoxABCD (Sso10828, 2656–2658) (24) and SoxEFGHIM (Sso2968–2973) (25).

Heterotrophic growth of *Sulfolobus* strains has been observed only in the presence of oxygen, and no alternative electron acceptors such as nitrate, DMSO, trimethylamine N-oxide, Fe^{3+}, or elemental sulfur can support anoxic growth (23). A gene for a unique quinol-oxidizing NO reductase (26) is truncated by an IS element (Sso1571/1573). Adjacent to this inactivated *nor* gene lies a gene cluster (Sso1577–1580) that encodes a predicted molybdopterin-containing oxidoreductase complex that may be a sulfite dehydrogenase. A respiratory-linked sulfide quinone reductase homolog (Sso2261) is also present. Subsequent oxidation of sulfur to sulfate is catalyzed by four cotranscribed enzymes (Sso2909–2912) (Fig. 1).

S. solfataricus can synthesize all 20 amino acids (23). The histidine operon, *hisCGABdFDEHI* (Sso0592–0600 and Sso6227), shows a novel organization. The *hisBpx* gene, found in some proteobacteria, is the only one absent (7). Highly conserved pyrimidine synthesis genes are concentrated in two operon-like structures within 5.5 kb (Sso0610–0615). Unusually, there is a close correlation between gene order and temporal position of the enzymes in the biosynthetic pathway. Only genes encoding carbamoylphosphate synthetase (Sso0640–0643) occur elsewhere, clustered with arginine biosynthesis genes (7, 27).

Motility and Protein Translocation

Movement of *S. solfataricus* is effected by clockwise, nonreversing flagella (23), and genes are present encoding methyl-accepting chemotaxis protein (Sso1469) and flagellar accessory proteins (orthologs of *Methanococcus jannaschii* FlaJIH; Sso2315, 2316, 2318, 2323). Archaeal flagellins, as well as binding proteins of *Sulfolobus* ABC transporters, are translocated by a transport system that recognizes type IV pilin-like signal peptides (28). Extracytoplasmic, cofactor-containing subunits of the respiratory systems are probably translocated via twin-arginine translocation (TAT). Indeed, *tatC* (Sso3108), and typical leader sequences, e.g., for the Rieske iron–sulfur protein (Sso2971), are present. In addition, components of the most common protein translocation system are encoded, including signal recognition particle (SRP, Sso0971) and SRP receptor protein (FtsY, Sso0348). The latter is adjacent to a gene encoding a SecE homolog (Sso5663), which together with SecY (Sso0695) constitutes the core of the Protein Secretion (Sec) pore, and a putative signal peptidase (Sso0916).

Chromatin and DNA-Binding Proteins

S. solfataricus and *A. pernix*, in contrast to euryarchaea (29), contain no proteins that share an ancestry with eukaryal histones. Nevertheless, a predicted protein (Sso0009, 327 aa) and two paralogs (Sso1117 and Sso0028) belong to a family of histone deacetylases (30), common to archaea and eukarya, which may influence transcription as in eukarya (31).

Other proteins usually present in eukaryotic chromatin, including high-mobility group (32), poly ADP-ribose polymerase (33), and chromatin assembly factor (CAF-1) (34), are absent from archaea. By contrast, a predicted structural maintenance of chromosomes-like protein (Sso2249; 864 aa), shares 46–48% similarity with archaeal proteins belonging to a large family of cohesins/condensins that also occurs in eukarya and bacteria (35).

No DNA-binding proteins of the bacterial HU or integration host factor type are encoded, but several small, basic, putative DNA-binding proteins are present, which constitute two families of about 7 and 10 kDa: Sso7d and Sso10d, respectively. Three paralogs of Sso7d are present (Sso10610, 9535, 9180). Orthologs of these proteins have been found only in the genus *Sulfolobus* including Ssh7a and 7b of *Sulfolobus shibatae* (36) and Sac7a, 7b, and 7d of *S. acidocaldarius* (37). Sac7d binds in the minor groove of DNA, causing sharp kinking (38). The 10-kDa protein (Sso0962) is exclusive to archaea (39). ORF Sso6877, which neighbors Sso0962 and shows 62% sequence similarity, may be a paralog. A single-strand-specific DNA-binding protein (SSB) (40) (Sso2364; 148 aa) may be required for DNA replication and repair.

DNA Replication and the Cell Cycle

Sulfolobus, like *Halobacterium* (41) but unlike *Pyrococcus* (42), encodes multiple CDC6 homologs (Sso0257, Sso0771, and Sso2184; Fig. 2*A*). In eukarya, CDC6 and minichromosome maintenance proteins interact with ORC/origin complexes, and one another, to regulate initiation of chromosome replication (43). Cumulative tetranucleotide skews of the *Sulfolobus* genome display complex overall shapes, which preclude a definite conclusion about the number and locations of the origin(s), although a single origin located near Sso0771 is favored by G + C and codon skew analyses (44). All archaeal genomes show sequence similarity to CDC6/CDC18 proteins and ORC1 of some eukarya, suggesting that the eukaryal CDC6 and ORC proteins are paralogs. Possibly, archaeal CDC6 homologs combine initiation functions of both eukaryal proteins. A further *Sulfolobus* CDC

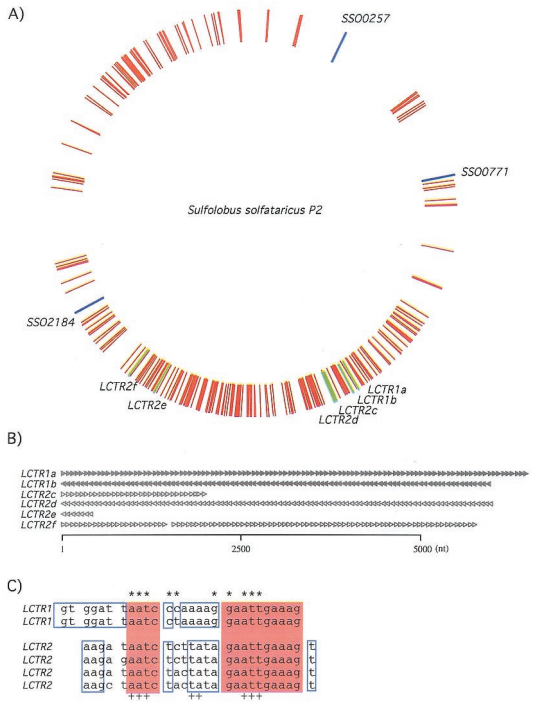

Fig. 2. (*A*) IS elements, predicted replication origins, and clusters of regular tandem repeats mapped on the *Sulfolobus* genome. Genes encoding three CDC6 homologs (Sso0257, Sso0771, and Sso2184) (blue lines); six loci of clustered tandem repeats (LCTRa–f) (green lines); IS elements (red lines) (see Table 1). (*B*) Loci of tandem repeat clusters and their orientation, indicated by arrowheads. Clustered tandem repeats of sequence family 1 (LCTRa and b) and 2 (LCTRc–f) are indicated by solid and hollow arrows, respectively. (*C*) Sequence characteristics of the repeat units. Those in loci LCTR1 and LCTR2 share two blocks of identical sequence (solid red boxes). Within each sequence family, there is a low level of sequence variability, whereas some regions are invariant (boxed in blue). (*) and (+) denote the positions involved in imperfect palindromic base pairing for tandem repeats of sequence families 1 and 2, respectively.

protein (Sso0083) has been implicated in cell division (45). The tRNA pseudouridine 55 synthase, which corresponds to centromere–microtubule-binding protein CBF5, is composed of two subunits in crenarchaea (Sso0393, Sso5761).

Three DNA polymerases, B1, B2, and B3 (Sso0552, Sso1459, and Sso0081, respectively), are encoded. Only B1 and B3 contain all exonuclease and polymerase motifs. Genes for DNA primases of bacterial- (Sso0079) and eukaryal-type (Sso1048) are present, which have euryarchaeal homologs, and one ATP-dependent DNA ligase (Sso0189). Also encoded are four topoisomerases: one ATP-independent type I, homologous to TopA (Sso0907), and two ATP-dependent reverse gyrases (Sso0420 and Sso0963). One of the latter is closely related to TopR of *S. acidocaldarius* (46), whereas the other is nearly identical to TopR of *S. shibatae* (47). The genome contains only one type II topoisomerase of the TopoVI family (Sso0968 and Sso0969) (48). No gyrase genes were detected, although *Sulfolobus* is novobiocin sensitive (49). Neither reverse gyrase nor TopoVI activity is inhibited by this drug (50, 51). However, a partially purified TopoII protein from *S. shibatae*, composed of two subunits (65 and 85 kDa) (52), could yield insight into the drug target.

Neither *ftsZ* nor *minD* is present in the genomes of *Sulfolobus* and *A. pernix* (53), suggesting that the bacterial, and possibly euryarchaeal, FtsZ-MinCDE system is absent from crenarchaea. There is also cytological evidence for a distinctive cell cycle mechanism in crenarchaea (54). A ParA homolog (Sso0034) is closely related to *Bacillus* Soj protein and many bacterial plasmid proteins that facilitate DNA partitioning. A similar archaeal Soj protein is encoded in *Sulfolobus* conjugative plasmid pNOB8 (55) and the *M. jannaschii* plasmid pURB800 (56). Both pNOB8 and the *Sulfolobus* chromosome contain large clusters of 20-nt tandem repeat sequences (LCTR) with a periodicity of 60–80 bp per unit (Fig. 2). Each unit contains a sequence with an imperfect 10-bp inverted repeat that could provide a binding site for a ParB (55, 57). About 420 copies of such repeats are concentrated at six loci, which fall into two related sequence families (Fig. 2*C*). Four clustered loci (LCTR1a and b and LCTR2c and d) contain 325 repeat elements, in different orientations (Fig. 2*B*) extending over 80 kb (Fig. 2*A*). The other two loci (LCTR2e and f) are clustered about 0.5 Mb away (Fig. 2*A*). Fewer copies of tandem repeats are present in the genomes of *A. pernix*, euryarchaea, and some pathogenic and hyperthermophilic bacteria.

DNA Repair and Recombination

Several genes are implicated in repair and recombination pathways, and two contribute to reversal lesion systems: *phr* encodes a photoreactivation enzyme, and the other encodes cysteine-*S*-methyltransferase (58). Endonucleases involved in base or nucleotide excision repair include two copies of endonucleases III, IV, and V, and a homolog of Rad2/Fen-1. The *uvrABC* system common to *Halobacterium* and *Methanobacterium thermoautotrophicum* is absent. Two homologous *mutT* genes encode enzymes involved in detoxification of the nucleotide pool, increasing fidelity of DNA replication, but no *mut*S/L mismatch repair genes were found. A putative bypass polymerase (Sso2448) homologous to the DinB SOS-induced polymerase IV of *E. coli* (59) is the first example of this gene in a thermophilic archaeon. A DinB homolog, claimed to derive from the closely related *S. solfataricus P1* (60), exhibits only 53% sequence identity with Sso2448.

A single-copy specific recombinase, Xer, is more similar in sequence to the *E. coli* XerD than to the archaeal XerC protein. Two RadA-related proteins are encoded in the genome. One is similar biochemically to Rad51 (Sso0250) (61), and the other (Sso0777) shows sequence similarity to bacterial RadA-like genes but appears unrelated to *Pyrococcus furiosus* RadB (62). Two Holliday junction resolvases, Hjc and Hje, have been described for *S. solfataricus* (63). Hjc is encoded by Sso0575, and we infer that Sso1176 encodes Hje. To generate ssDNA substrates for recombination, *Sulfolobus* may use Sso2249 and Sso2250, homologs of the Rad50 and Mre11 proteins, respectively, of *Saccharomyces cerevisiae*. About 15 helicases are encoded that may be involved in repair and recombination pathways, including two homologs of *S. cerevisiae* Rad25, and single homologs of Rad54 and Chl1 of the Rad3 family.

Transcription

The archaeal transcription initiation machinery is a simplified version of the eukaryal system (reviewed in ref. 64). The archaeal and eukaryal polymerases are closely related in subunit complexity and sequence (65, 66). Fourteen RNA polymerase subunits are encoded by 8 loci: *rpoDN* (Sso0071, 5140), *rpoHB-'B''AA''*(Sso5468, 0227, 3254, 0225, 0223), *rpoG* (Sso0277; ortholog of *S. acidocaldarius* RpoG) (67), *rpoL* (Sso5577), *rpoE'E''* (Sso0415, 5798), *rpoF* (Sso0751; ortholog of *M. thermoautotrophicum rpoF*) (68), *rpoK* (Sso6768) and *rpoP* (Sso5865; ortholog of *M. thermoautotrophicum rpoP*) (68). Apart from *rpoG, rpoK*, and *rpoP*, the *rpo* operons also encode ribosomal proteins. In addition, *rpoHB'B''A'A''* lie adjacent to genes encoding translation factors (IF2, EF1α) and a NusA-like transcription elongation factor. Downstream from *rpoL*, two ORFs (Sso5576 and Sso0291) show homology to the N- and C-terminal ends of *rpoM* of *S. acidocaldarius*. However, Sso0291 is also similar to a transcription elongation factor (TFS), of unknown function, from euryarchaeal genomes (69). A NusG-like transcription antiterminator is located in a distinct gene cluster (Sso0342–0353) coding for ribosomal pro-

teins, a translation factor (eIF6), and an FtsY-like signal recognition particle-receptor.

Initiation of eukaryal RNA polymerase II requires several transcription factors. Transcriptional initiation at archaeal promoters requires the TATA-binding protein (TBP) and a homolog of transcription factor IIB (TFB). Genes encoding TBP (Sso0951) and at least two TFB paralogs (Sso0446 and 0946) are present. In addition, an ortholog of the TFIIEα-subunit (Sso0266) is present in all archaea. Orthologs of the eukaryal TFIIA, TFIIEβ, TFIIF, and TFIIH appear to be absent from all archaea. Transcriptional regulation of this eukaryal-like system involves many types of bacterial-like regulators (64, 70, 71). At least 38 potential transcription regulators occur, many belonging to the families Lrp/AsnC (72), MarR, and AcrR. *Sulfolobus* encodes eukaryal-like regulatory factors (71), including a TBP-interacting protein (TIP49; Sso2450) and a multiprotein bridging factor (Sso0270).

Translational Apparatus

The translational machinery exhibits eukaryal and bacterial characteristics and some archaeal-specific features. For example, a Shine–Dalgarno sequence is found upstream of the genes inside operons but not, generally, for the first gene in an operon or isolated genes; this indicates that different mechanisms are used for translation initiation in *S. solfataricus* (73).

Forty-six unlinked tRNA genes [by tRNA SCAN-SE (74)] carry 43 different anticodons. Three $tRNA^{Met}$ genes with CAU anticodons produce initiator and elongator tRNAs, whereas the third may be modified to read $tRNA^{Ile}$-AUA. Rare codons include CGG (0.07%) and CGC (0.10%).

Eighteen tRNAs are predicted to contain single introns, and $tRNA^{Cys}$ has two. Fourteen introns occur in the *A. pernix* genome (53) and considerably less (2–5) in euryarchaeal genomes. Sixteen of the introns, including two within $tRNA^{Met}$ genes, are located between the +1 and +2 nucleotides of the 3′-end of the anticodon, as in eukaryotic tRNAs. Identical introns are located at the same positions within the D-loops of $tRNA^{Leu}$-CUC and $tRNA^{Leu}$-UUC, whereas another lies in the anticodon stem of $tRNA^{Cys}$. Previously, no archaeal intron was detected in the D-loop or at two different positions of the same tRNA (75). Aminoacyl–tRNA synthetases are encoded for every amino acid except asparagine and glutamine, which probably require amidotransferases for their synthesis. There are multiple subunits of Glu-tRNA amidotransferase (*gatABC*; Sso0957, 0765, and 2122; Sso0232, Sso6855).

Organization of rRNA genes is crenarchaeal in character with linked 16S and 23S rRNA genes that are not cotranscribed with tRNAs, 5S rRNA, or 7S RNA. Matches with the modification systems of tRNAs and rRNAs include the archaeal intron splicing endonuclease (Sso0439), tRNA nucleotidyl transferase (Sso1039) and *N*-6 methylase (Sso0749). Genes encoding homologues of fibrillarin (Sso0940) and Nop56 (Sso0939) are also present. Six putative snoRNA genes were identified by searching *S. acidocaldarius* snoRNAs against the genome (76). The predicted 2-*O*-ribose methylation sites on the RNAs are: U52 in 16S rRNA; G84, G2658, and G2731 in 23S rRNA; and U35 in $tRNA^{Gln}$-UUG.

Sixty-five ribosomal proteins are present, 28 small subunit and 37 large subunit. Most occur in four operon-like clusters (7). Sequence similarities are generally higher with eukaryal than with bacterial proteins, as for other archaea. Homologs of the eukaryal-specific S25 (Sso0425) and S26 (Sso6179) are encoded that also occur in *A. pernix* (53) but not in euryarchaea. Genes encoding translational initiation factors IF-1 (SUI1), eIF1A, bacterial-type IF-2, eIF2α, β, and γ subunits, eIF2B α subunit, eIF4, eIF5A, and eIF6, as well as elongation factors EF-Tu, EF-G/2, and EF-1β subunit, are present.

S. solfataricus produces the thermosome chaperonin TF55 and not the bacterial heat-shock protein 70 chaperonin. Three subunits, TF55α, β, and γ, are encoded, whereas only one (TF55α) or two (TF55α, β) subunits occur in other archaea. Three subunits correlate with *Sulfolobus* chaperonins, forming nine-membered rings rather than the eight-membered rings of other archaeal chaperonins (77). *Sulfolobus* chaperonins constitute up to 4% of total cellular protein and can generate a filamentous structure resembling a eukaryotic cell skeleton *in vitro* (78).

Table 1. Insertion elements in the *S. solfataricus* genome

Name	Size, bp	Inverted repeat, bp	Direct repeat, bp	Family	Number of full-length copies	Number of partial copies
ISC774	774	15	0	ND	2	3
ISC1043	1,043	14	0	ISL3	4	2
ISC1048	1,048	23	2	IS630/T c1	11	1
ISC1058	1,058	19	9	IS5	14	1
ISC1078	1,084	19	2	IS630/T c1	8	2
ISC1160	1,160	12	6	IS4	3	2
ISC1173	1,173	46	8	ND	5	2
ISC1190	1,190	0	0	IS110	15	2
ISC1212	1,212	27	0	IS5	9	4
ISC1217	1,148	13	6	ND	11	2
ISC1225	1,225	17	4/5	IS4	11	1
ISC1229	1,229	0	0	IS110	7	2
ISC1234	1,234	19	4	IS5	16	1
ISC1250	1,250	9	0	IS256	3	0
ISC1290	1,290	40	0	IS5	3	2
ISC1316	1,316	0	0	IS605	13	1
ISC1332	1,332	22	9	IS256	1	0
ISC1359	1,367	25	4	IS4	10	5
ISC1395	1,395	68	0	IS630/T c1	4	3
ISC1439	1,439	20	9	IS4	32	0
ISC1476	1,476	20	0	IS605	2	1
ISC1491	1,491	0	0	IS110	5	1
ISC1904	1,904	0	0	IS605	11	3
ISC1913	1,913	0	0	IS605	2	3

The numbers given are averaged for each class of IS element. Partial copies are >250 bp.

Transposable Elements and Chromosomal Integration

Two hundred IS elements were identified (Table 1) that are spread around the genome and clustered into two broad areas (Fig. 2*A*). One hundred eighty-four belong to seven known bacterial/eukaryal families, whereas the remainder fall in new families, which may be archaea-specific (Table 1). A further 44 partial copies of IS elements (>250 bp) are present. In total, these elements constitute about 10% of the genome. ISC1058, ISC1217, ISC1359, and ISC1439 have been shown to be active (79), and the presence of identical copies of several IS elements suggests that others have duplicated recently. Five classes of IS elements contain two ORFs. The first ORFs of ISC1904 and ISC1913 show sequence similarity to a resolvase, common to transposons. For the shorter IS elements ISC774, ISC1212, and ISC1229, the ORFs overlap (Table 1). ISC1316, ISC1332, and ISC1913 also occur in conjugative plasmids pNOB8 and pING of *Sulfolobus* (55, 80). One hundred forty-three smaller (80- to 180-bp) elements, present in four classes, are likely to be mobilized by IS element-encoded transposases (81).

Integrase-mediated insertion of the virus SSV1 into the *S. shibatae* genome occurs in the downstream half of a $tRNA^{Arg}$ gene, producing a partitioned *int* gene (82). Similar partitioned integrase genes occur in the genome of *S. solfataricus* and other archaea (83), where the downstream part, *int(N)*, corresponds to the insertion sequence recognized by the integrase. This suggests that integrase-mediated insertion has occurred occasionally (84).

Conclusions

For many years, *S. solfataricus* has been the model organism for studying crenarchaeal biology, and the emerging genome se-

CELL BIOLOGY

quence has been exploited extensively. The complete genome shows a high degree of plasticity. It can also generate a rich diversity of metabolic reactions where it appears to use ferredoxin as the primary metabolic electron carrier. The data reveal a high proportion of archaea-specific genes and reinforce the major differences between archaea, and bacteria and eukarya. Clear differences are also discernible between the crenarchaea and euryarchaea in their DNA replication and translation mechanisms and in their cell cycle processes.

The European contribution to this paper was financed by European Union Grant BIO4CT-960270. The Canadian contribution was financed by the Canadian Genome Analysis and Technology Program (Grant GO12319) and the National Research Council of Canada (Institute for Marine Biosciences).

1. Zillig, W., Stetter, K. O., Wunderl, S., Schulz, W., Priess, H. & Scholz, I. (1980) *Arch. Microbiol.* **125,** 259–269.
2. Pfeifer, F., Palm, P. & Schleifer, K.-H., eds. (1994) *Molecular Biology of Archaea* (Gustav Fischer, Stuttgart), pp. 1–267.
3. Zillig, W., Arnold, H. P., Holz, I., Prangishvili, D., Schweier, A., Stedman, K., She, Q., Phan, H., Garrett, R. A. & Kristjansson, J. K. (1998) *Extremophiles* **2,** 131–140.
4. Sowers, K. R. & Schreier, H. J. (1999) *Trends Microbiol.* **7,** 212–219.
5. Sensen, C. W., Charlebois, R. L., Chow, C., Clausen, I. G., Curtis, B., Doolittle, W. F., Duguet, M., Erauso, G., Gaasterland, T., Garrett, R. A., *et al*. (1998) *Extremophiles* **2,** 305–312.
6. Sensen, C. W. (1999) in *Organization of the Prokaryotic Genome*, ed. Charlebois, R. L. (Am. Soc. Microbiol., Washington, DC), pp. 1–9.
7. Charlebois, R. L., Singh, R. K., Chan-Weiher, C. C., Allard, G., Chow, C., Confalonieri, F., Curtis, B., Duguet, M., Erauso, G., Faguy, D., *et al*. (2000) *Genome* **43,** 116–136.
8. She, Q., Confalonieri, F., Zivanovic, Y., Medina, N., Billault, A., Awayez, M. J., Thi-Ngoc, H. P., Pham, B. T., Van der Oost, J., Duguet, M. & Garrett, R. A. (2000) *DNA Sequence* **11,** 183–192.
9. Nelson, K. E., Clayton, R. A,, Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., *et al.* (1999) *Nature (London)* **399,** 323–329.
10. Charlebois, R. L., Gaasterland, T., Ragan, M. A., Doolittle, W. F. & Sensen, C. W. (1996) *FEBS Lett.* **389,** 88–91.
11. Gaasterland, T. & Ragan, M. A. (1998) *Microbial Comp. Genomics* **3,** 177–192.
12. Gaasterland, T. & Ragan, M. A. (1998) *Microbial Comp. Genomics* **3,** 199–217.
13. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
14. Gaasterland, T. & Sensen, C. W. (1996) *Biochimie* **78,** 302–310.
15. Moll, R. & Schäfer, G. (1988) *FEBS Lett.* **232,** 359–363.
16. Matin, A. (1990) *FEMS Microbiol. Rev.* **75,** 307–318.
17. Elferink, M. G. L., Albers, S. V., Konings, W. N. & Driessen, A. J. M. (2001) *Mol. Microbiol.* **39,** 1494–1503.
18. Ruepp, A., Graml, W., Santos-Martinez, M. L., Koretke, K. K., Volker, C., Mewes H. W., Frishman, D., Stocker, S., Lupas, A. N. & Baumeister, W. (2000) *Nature (London)* **407,** 508–513.
19. Schönheit, P. & Schäfer, T. (1995) *World J. Microbiol. Biotechnol.* **11,** 26–57.
20. Wakao, H., Wakagi, T. & Oshima, T. (1987) *J. Biochem. (Tokyo)* **102,** 255–262.
21. Albracht, S. P. & Hedderich, R. (2000) *FEBS Lett.* **485,** 1–6.
22. Finel, M. (1998) *Trends Biochem. Sci.* **23,** 412–414.
23. Grogan, D. W. (1989) *J. Bacteriol.* **171,** 6710–6719.
24. Lübben, M. (1995) *Biochim. Biophys. Acta* **1229,** 1–22.
25. Castresana, J., Lübben, M. & Saraste, M. (1995) *J. Mol. Biol.* **250,** 202–210.
26. Cramm, R., Pohlmann, A. & Friedrich, B. (1999) *FEBS Lett.* **460,** 6–10.
27. Charlebois, R. L., Sensen, C. W., Doolittle, W. F. & Brown, J. R. (1997) *J. Bacteriol.* **179,** 4429–4432.
28. Albers, S. V., Konings, W. N. & Driessen, A. J. (1999) *Mol. Microbiol.* **31,** 1595–1596.
29. Reeve, J. N., Sandman, K. & Daniels, C. J. (1997) *Cell* **89,** 999–1002.
30. Ng, H. H. & Bird, A. (2000) *Trends Biochem. Sci.* **25,** 121–126.
31. Geisberg, J. V. & Struhl, K. (2000) *Mol. Cell. Biol.* **20,** 1478–1488.
32. Bianchi, M. E. & Beltrame, M. (1998) *Am. J. Hum. Genet.* **63,** 1573–1577.
33. Burkle, A. (2000) *Ann. N.Y. Acad. Sci.* **908,** 126–132.
34. Ridgway, P. & Almouzni, G. (2000) *J. Cell Sci.* **113,** 2647–2658.
35. Graumann, P. L. (2000) *J. Bacteriol.* **182,** 6463–6471.
36. Mai, V. Q., Chen, X., Hong, R. & Huang, L. (1998) *J. Bacteriol*. **180,** 2560–2563.
37. McAfee, J. G., Edmondson, S. P., Datta, P. K., Shriver, J. W. & Gupta, R. (1995) *Biochemistry* **34,** 10063–10077.
38. Robinson, H., Gao, Y. G., McCrary, B. S., Edmondson, S. P., Shriver, J. W. & Wang, A. H. (1998) *Nature (London)* **392,** 202–205.
39. Forterre, P., Confalonieri, F. & Knapp, S. (1999) *Mol. Microbiol.* **32,** 669–670.
40. Wadsworth, R. I. & White, M. F. (2001) *Nucleic Acids Res.* **29,** 914–920.
41. Ng, W. V., Kennedy, S. P., Mahairas, G. G., Berquist, B., Pan, M., Shukla, H. D., Lasky, S. R., Baliga, N. S., Thorsson, V., Sbrogna, J., *et al*. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 12176–12181. (First Published October 3, 2000; 10.1073/pnas.190337797)
42. Myllykallio, H., Lopez, P., Lopez-Garcia, P., Heilig, R., Saurin, W., Zivanovic, Y., Philippe, H. & Forterre, P. (2000) *Science* **288,** 2212–2215.
43. Tye, B. K. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 2399–2401.
44. Brügger, K., Redder, P., She, Q., Confalonieri, F., Zivanovic, Y. & Garrett, R. A. (2001) *FEMS Microbiol. Lett.,* in press.
45. Ragan, M., Logsdon, J. M., Jr., Sensen, C. W., Charlebois, R. L. & Doolittle, W. F. (1996) *FEMS Microbiol. Lett.* **144,** 151–155.
46. Confalonieri, F., Elie, C., Nadal, M., de La Tour, C., Forterre, P. & Duguet, M. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 4753–4757.
47. Jaxel, C., Bouthier de la Tour, C., Duguet, M. & Nadal, M. (1996) *Nucleic Acids Res.* **24,** 4668–4675.
48. Bergerat, A., de Massy, B., Gadelle, D., Varoutas, P. C., Nicolas, A. & Forterre, P. (1997) *Nature (London)* **386,** 414–417.
49. Sioud, M., Possot, O., Elie, C., Sibold, L. & Forterre, P. (1988) *J. Bacteriol*. **17,** 946–953.
50. Nakasu, S. & Kikuchi, A. (1985) *EMBO J.* **42,** 705–710.
51. Bergerat, A., Gadelle, D. & Forterre, P. (1994) *J. Biol. Chem*. **269,** 27663–27669.
52. Assairi, L. M. (1994) *Biochim. Biophys. Acta* **121,** 107–114.
53. Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., *et al.* (1999) *DNA Res.* **6,** 83–101.
54. Hjort, K. & Bernander, R. (1999) *J. Bacteriol.* **181,** 5669–5675.
55. She, Q., Phan, H., Garrett, R. A., Albers, S.-V., Stedman, K. M. & Zillig, W. (1998) *Extremophiles* **2,** 417–425.
56. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., *et al.* (1996) *Science* **273,** 1058–1073.
57. Charlebois, R. L., She, Q., Sprott, D. P., Sensen, C. W. & Garrett, R. A. (1998) *Curr. Opin. Microbiol.* **1,** 584–588.
58. Sancar, G. B., Smith, F. W., Lorence, M. C., Rupert, C. S. & Sancar, A. J. (1984) *Biol. Chem*. **259,** 6033–6038.
59. Wagner, J., Gruz, P., Kim, S. R., Yamada, M., Matsui, K., Fuchs, R. P. & Nohmi, T. (1999) *Mol. Cell* **4,** 281–286.
60. Kulaeva, O. I., Koonin, E. V., McDonald, J. P., Randall, S. K., Rabinovich, N., Connaughton, J. F., Levine, A. S. & Woodgate, R. (1996) *Mutat. Res.* **357,** 245–253.
61. Seitz, E. M., Brockman, J. P., Sandler, S. J., Clark, A. J. & Kowalczykowski, S. C. (1996) *Genes Dev.* **12,** 1248–1253.
62. DiRuggiero, J. & Robb, F. T. (1998) in *New Developments in Marine Biotechnology*, eds. Le Gal, Y. & Halvorson, H. (Plenum, New York), pp. 193–196.
63. Kvaratskhelia, M. & White, M. F. (2000) *J. Biol. Chem.* **297,** 923–932.
64. Bell, S. D. & Jackson, S. P. (2001) *Curr. Opin. Microbiol.* **4,** 208–213.
65. Pühler, G., Leffers, H., Gropp, F., Palm, P., Klenk, H.-P., Lottspeich, F., Garrett, R. A. & Zillig, W. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 4569–4573.
66. Langer, D., Hain, J., Thuriaux, P. & Zillig, W. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 5768–5772.
67. Lanzendorfer, M., Langer, D., Hain, J., Klenk, H. P., Holz, I., Arnoldammer, I. & Zillig, W. (1994) *Syst. Appl. Microbiol*. **16,** 656–664.
68. Darcy, T. J., Hausner, W., Awery, D. E., Edwards, A. M., Thomm, M. & Reeve, J. N. (1999) *J. Bacteriol.* **181,** 4424–4429.
69. Hausner, W., Lange, U. & Musfeldt, M. (2000) *J. Biol. Chem.* **275,** 12393–12399.
70. Kyrpides, N. C. & Ouzounis, C. A. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 8545–8550.
71. Aravind, L. & Koonin, E. V. (1999) *Nucleic Acids Res.* **27,** 4658–4670.
72. Di Napoli, A., Van der Oost, J., Sensen C. W., Charlebois, R. L., Rossi, M. & Ciaramella, M. (1999) *J. Bacteriol.* **181,** 1474–1480.
73. Tolstrup, N., Sensen, C. W., Garrett, R. A. & Clausen, I. G. (2000) *Extremophiles* **4,** 175–179.
74. Lowe, T. M. & Eddy, S. R. (1997) *Nucleic Acids Res.* **25,** 955–964.
75. Lykke-Andersen, J., Aagaard, C., Semionenkov, M. & Garrett, R. A. (1997) *Trends Biochem. Sci.* **22,** 326–331.
76. Omer, A. D., Lowe, T. M., Russell, A. G., Ebhardt, H., Eddy, S. R. & Dennis, P. P. (2000) *Science* **288,** 517–522.
77. Trent, J. D., Nimmesgern, E., Wall, J. S., Hartl, F.-U. & Horwich, A. L. (1991) *Nature (London)* **354,** 490–493.
78. Trent, J. D., Kagawa, H. K., Yaoi, T., Olle, E. & Zaluzec, N. J. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 5383–5388.
79. Martusewitsch, E., Sensen, C. W. & Schleper, C. (2000) *J. Bacteriol.* **182,** 2574–2581.
80. Stedman, K. M., She, Q., Phan, H., Holz, I., Singh, H., Prangishvili, D., Garrett, R. & Zillig, W. (2000) *J. Bacteriol.* **182,** 7014–7020.
81. Redder, P., She, Q. & Garrett, R. A. (2001) *J. Mol. Biol.* **306,** 1–6.
82. Muskhelishvili, G., Palm, P. & Zillig, W. (1993) *Mol. Gen. Genet.* **237,** 334–342.
83. She, Q., Peng, X., Zillig, W. & Garrett, R. A. (2001) *Nature (London)* **409,** 478.
84. Peng, X., Holz, I., Zillig, W., Garrett, R. A. & She, Q. (2000) *J. Mol. Biol.* **303,** 449–454.