ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/41654417

Functional Characterization and High-Throughput Proteomic Analysis of Interrupted Genes in the Archaeon Sulfolobus solfataricus

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · MARCH 2010

Impact Factor: 4.25 · DOI: 10.1021/pr901166q · Source: PubMed

CITATIONS
12

READS
29

12 AUTHORS, INCLUDING:

Beatrice Cobucci-Ponzano
Italian National Research Council
51 PUBLICATIONS 566 CITATIONS
SEE PROFILE

Paola Londei
Sapienza University of Rome
81 PUBLICATIONS 1,619 CITATIONS
SEE PROFILE

Odile Lecompte
ICube Laboratory
42 PUBLICATIONS 1,238 CITATIONS
SEE PROFILE

Mosè Rossi
Italian National Research Council
305 PUBLICATIONS 5,614 CITATIONS
SEE PROFILE

Available from: Beatrice Cobucci-Ponzano
Retrieved on: 03 February 2016

research articles Journal of proteome research

Functional Characterization and High-Throughput Proteomic Analysis of Interrupted Genes in the Archaeon *Sulfolobus solfataricus*

Beatrice Cobucci-Ponzano,[†] Lucia Guzzini,[†] Dario Benelli,[‡] Paola Londei,[‡] Emmanuel Perrodou,[§] Odile Lecompte,[§] Diem Tran,[||] Jun Sun,[⊥] Jing Wei,[#] Eric J. Mathur,[∇] Mosè Rossi,[†,○] and Marco Moracci*,[†]

Institute of Protein Biochemistry−CNR, Via P. Castellino 111, 80131 Naples, Italy, Dipartimento di Biotecnologie Cellulari ed Ematologia, Università di Roma Sapienza, Policlinico Umberto I, Viale Regina Elena 324, 00161 Roma, Italy, Department of Structural Biology and Genomics, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), F-67400 Illkirch, France, AviaraDx, Inc., 11025 Roselle Street, Suite 200, San Diego, California 92121, Genomatica, Inc., 10520 Wateridge Circle, San Diego, California 92121, BiogenIdec Corporation, 5200 Research Place, San Diego, California 92121, Synthetic Genomics, Inc., 11149 North Torrey Pines Road, La Jolla, California 92037, and Dipartimento di Biologia Strutturale e Funzionale, Università di Napoli "Federico II", Complesso Universitario di Monte S. Angelo, Via Cinthia 4, 80126 Naples, Italy

Received December 18, 2009

Sequenced genomes often reveal interrupted coding sequences that complicate the annotation process and the subsequent functional characterization of the genes. In the past, interrupted genes were generally considered to be the result of sequencing errors or pseudogenes, that is, gene remnants with little or no biological importance. However, recent lines of evidence support the hypothesis that these coding sequences can be functional; thus, it is crucial to understand whether interrupted genes are expressed *in vivo*. We addressed this issue by experimentally demonstrating the existence of functional disrupted genes in archaeal genomes. We discovered previously unknown disrupted genes that have interrupted homologues in distantly related species of archaea. The combination of a RT-PCR strategy with shotgun proteomics demonstrates that interrupted genes in the archaeon *Sulfolobus solfataricus* are expressed *in vivo*. In addition, the sequence of the peptides determined by LCMSMS and experiments of *in vitro* translation allows us to identify a gene expressed by programmed −1 frameshifting. Our findings will enable an accurate reinterpretation of archaeal interrupted genes shedding light on their function and on archaeal genome evolution.

Keywords: Programmed frameshifting • gene expression • pseudogenes • translational recoding • Kae1/Bud32 • SUI-1

Introduction

Large sections of eukaryotic chromosomes contain nonfunctional DNA whose role is still unknown; such chromosomal stretches contain sequences known as pseudogenes, which are particularly abundant.[1] Pseudogenes are defined as disabled copies of genes or decayed remnants of genes. Some do not have introns or promoters while most have gene-like features (promoters, splice sites, etc.), but lack protein-coding ability resulting from various genetic disablements such as stop codons, frameshifts, or a lack of transcription.[1−3] For these reasons, in genome analysis, the identification of a disrupted gene (DG) is usually considered an evidence of a nonfunctional pseudogene. Nevertheless, the role of pseudogenes, which are more common in eukaryotic genomes,[1] is still poorly understood. In fact, some pseudogenes are indeed functional indicating their potentiality for becoming new genes useful for an organism's survival and adaptation to particular environmental changes.[3−5]

DG are far less common in prokaryotes that have compact genomes. Relevant exceptions are the genomes of the pathogenic Eubacteria *Mycobacterium leprae*, *Yersinia pestis*, and *Rickettsia prowazekii*[6−9] showing an elevated number of DG as an effect of their lifestyle change as intracellular parasites.[3,10] On the other hand, the identification of disruptions in prokaryotic genes can be misleading. In fact, while the quality of gene calling in eukaryotic genomes has dramatically improved by comparison with transcript sequences, the annotation of prokaryotic genomes still relies on *ab initio* prediction pro-

* To whom the correspondence should be addressed. Tel. +39-081-6132271. Fax +39-081-6132277. E-mail: m.moracci@ibp.cnr.it.
[†] Institute of Protein Biochemistry−CNR.
[‡] Università di Roma Sapienza.
[§] Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC).
[||] AviaraDx, Inc.
[⊥] Genomatica, Inc.
[#] BiogenIdec Corp.
[∇] Synthetic Genomics, Inc.
[○] Università di Napoli "Federico II".

10.1021/pr901166q © XXXX American Chemical Society

grams leading to errors in start codons predictions and in detection of authentic frameshifts and in-frame stop codons.[11]

More interestingly, disruptions do not necessarily are symptomatic of disabled genes; for instance, DG could be merged by RNA splicing, by natural suppressor tRNAs, which read stop codons as natural sense codons, or they could be the result of the splitting of a full-length gene into ORFs encoding separated polypeptides domains arranged in a multisubunit protein.[12] Moreover, genes interrupted by single stop codon or +/−1 frameshifts could be expressed by different mechanisms of translational regulation globally termed *recoding*, in which localized deviations from the standard translational rules take place (for reviews on recoding see refs 13−18). These observations challenge the popular belief that DG encountered in sequenced genomes are nonfunctional, explaining why several computational methods to identify sequencing errors, pseudogenes, or recoding events have been designed[11] and references therein.[19−21]

In this panorama, the identification of authentic DG and their functional characterization would be of great interest to understand their mechanism of expression *in vivo*. Despite the abundance of computational techniques, the experimental characterization on a whole genome of DG by high-throughput proteomics and biochemical characterization has never been performed.

The aim of this study is to identify DG in archaeal genomes and to test experimentally their functionality. We show here the identification of 98 interrupted genes in 16 archaeal genomes. We performed an in depth analysis of DG in *S. solfataricus* by combining several bioinformatics and experimental approaches. In this species, DG have interrupted homologues in other Archaea presenting the disruption in a conserved position. A high-throughput proteomic screening and the biochemical characterization allowed us to unequivocally demonstrate that most of the genes interrupted by single events are actively expressed *in vivo*. Remarkably, the ORFs encoding for the disrupted ortholog of Kae1/Bud32, an essential complex in yeast and humans implicated in transcription and telomere maintenance,[22] are expressed independently in *S. solfataricus* and the two polypeptides interact *in vitro*. In addition, experiments of *in vitro* translation by using *S. solfataricus* extracts and proteomic analysis indicated that the gene encoding for the universal translation regulator SUI-1 is expressed by programmed −1 frameshifting. This is the first proteomic high-throughput study on interrupted genes and it demonstrates that DG should be taken into account in the annotation of the sequenced genomes and in protein predictions.

Experimental Section

Identification of Interrupted Genes in Archaea. A preliminary manual identification of interrupted archaeal genes was performed by a keyword search in the annotations of 16 archaeal sequenced genomes downloaded from the NCBI data bank (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi), namely *Aeropyrum pernix*, *Archaeoglobus fulgidus*, *Halobacterium* sp. NRC-1, *Methanococcus jannaschii*, *Methanopyrus kandleri*, *Methanosarcina acetivorans*, *Methanosarcina mazei*, *Methanobacterium thermoautotrophicum*, *Pyrobaculum aerophilum*, *Pyrococcus abyssi*, *Pyrococcus furiosus*, *Pyrococcus horikoshii*, *S. solfataricus* strain P2, *Sulfolobus tokodaii*, *Thermoplasma acidophilum*, and *Thermoplasma volcanium*. The terms used were: “frag”, “truncation”, “part”, “amino”, “carboxy”, “termin”, “N-”, and “C-”. For the ORFs that were annotated as truncated at the N- or the C-termini, we analyzed the flanking regions to test if other contiguous ORFs encode for homologous proteins.

The amino acid sequences derived from all the ORFs identified were compared to the amino acidic sequences derived from the genomes of the 16 Archaea considered by using the BLASTP program[23] run on the site http://www-archbac.u-psud.fr/projects/sulfolobus/Blast_Search.html with default BLASTP parameters. The ORFs showing a BLASTP alignment score >80 to full-length ORFs in the database were selected.

A majority of the interrupted coding sequences identified are formed by two adjacent ORFs, showing the same orientation, and separated by single frameshifting events, stop codons, and insertions of longer sequences (shown in the Tables as +/−1, *n*T, and Ins/IS, respectively). The ORFs were classified as deleted (shown in the Tables as N-del and C-del) when they showed similarity only to the 5′- or the 3′-terminal part of the full-length gene found in the database.

In addition to this annotation-based search, the genome of *S. solfataricus*, strain P2, was scanned for interrupted genes using Interrupted CoDing Sequence (ICDS) detection program.[11] This program searches for neighboring ORFs in the same orientation that share common homologues and that are not paralogues. The results are independent from previous annotations of the considered genome since the program has been performed on the raw genomic sequence.

Analysis of the Available 3D-Structures. To understand whether the DG in *S. solfataricus* encode for independent protein domains, we searched the three-dimensional structures in the NCBI database derived from the RCSB protein structure data bank (http://www.rcsb.org/pdb/home/home.do) by employing the ‘domain parsing’ technique. Briefly, we used these entries to search by sequence similarity proteins for which the 3D-structure is available. We downloaded from the RCSB database the structure of the similar proteins and we examined manually the structure to find evidence of discrete domains. Then, we evaluated the frameshift position in the query sequence to see if a truncated protein might fall on a natural domain boundary, and finally, the translated ORFs were aligned to the corresponding full-length genes by using the program MultAlin (http://multalin.toulouse.inra.fr/multalin/multalin.html).[24]

Sequencing. To test if the frameshifts and the stop codons observed in *S. solfataricus* were corrected by RNA editing events or were the result of sequencing mistakes, we sequenced the region between the ORFs of each *S. solfataricus* DG from both genomic DNA and total RNA preparations. To this aim, we designed primers to amplify the region comprising the 3′- and the 5′-ends of the first and the second contiguous ORFs, respectively. We devoted particular attention to the design of the amplification primers. In the case of ORFs expressed from separate mRNAs, if the upstream primer matches too close to the second ORF, the RT-PCR could yield products originating from transcripts initiating between the two ORFs. To exclude this possibility, we designed the RT-PCR upstream primers so that they hybridize to a region of more than 200 nt from the first putative start codon of the downstream ORF (Figure 2). In fact, it is reported that in the mRNAs from *S. solfataricus* the 5′-untranslated region can vary from less than 10 nt to about 100 nt in the leaderless and leadered mRNAs, respectively.[25] The extraction of the genomic DNA and of the total RNA fragments and the methods used to perform the PCR and the RT-PCR reactions were described elsewhere.[26] The synthetic oligonucleotide primers (M-medical Florence, Italy) used for the amplification are listed in Supporting Table 1. The primers used for SSO11867/SSO3060 were described elsewhere.[26] We deposited the sequences of the genes of the entries 1 and 6−8 in Table 1 (accession numbers AY336520, AY336521, AY336522, and AY336523, respectively) that differ in single or multiple bases from those available in the database.

High-Throughput Proteomic Analysis. Cells of *S. solfataricus*, strain P2, from our collection were grown as previously reported[26] and lysated with Lysis Buffer containing 1% RapiGest, 50 mM Tris-HCl, pH 8.0, 100 mM NaCl, and 2 mM EDTA buffer. Cell lysates were sonicated for 15 min, boiled for 5 min, and then incubated at room temperature for 60 min to extract total proteins. After centrifugation at 20 000*g* for 15 min, supernatants were collected and protein concentrations were measured. The proteins were diluted 10× with water and then digested with trypsin. After complete digestion, RapiGest was degraded by acid treatment and resulted solution was centrifuged at 20 000*g* for 30 min. The supernatants contained hydrophilic peptides, whereas the pellets contained hydrophobic peptides and were resuspended with 70% isopropanol. The resulting peptides were analyzed by liquid chromatography online tandem mass spectrometry (LCMSMS) separately accordingly to previously published methods.[27] Collected MS/MS data were searched against a database containing the predicted *S. solfataricus* ORFs, which we named Interrupted Genes Data set file (IGD), and a reversed human proteins sequences as background. The peptide identifications were filtered as reported earlier[27] and high-confident peptide identifications were used for protein identifications.

Functional Analysis of *S. solfataricus* Interrupted Genes. The DNA fragment including the ORFs SSO0434/SSO0433 were amplified by PCR from genomic *S. solfataricus*, strain P2, as previously described[26] by using the oligonucleotides OsyaloBamHI and OsyaloEcoRI (Primm, Milan, Italy) (Supporting Table 1). The hydrolyzed fragment was cloned in vector pGEX2TK (GE Healthcare); direct sequencing identified the recombinant plasmid, and the gene was completely resequenced. In this vector, named pGST434/433, the ORF SSO0434 is fused in frame at its N-terminus with the Gluthatione S-Transferase-tag (GST). The downstream ORF SSO0433 was then removed by digesting pGST434/433 with *MunI* and *EcoRI* and religation of the linearized vector obtaining pGST434. The ORF SSO0433 was amplified by PCR from genomic *S. solfataricus*, strain P2, by using the oligonucleotides 0433Nde and 0433Xho (Primm, Milan, Italy) (Supporting Table 1). After digestion with *NdeI* and *XhoI*, the amplified fragment was ligated to the vector pCola Duet obtaining the recombinant plasmid pS-0433 in which the sequence of SSO0433 was checked by sequencing. In this plasmid, the SSO0433 ORF is fused in frame at its C-terminus with an S-tag.

GST-SSO0434 and SSO0433-S-tag were expressed independently in *Escherichia coli* cells BL21DE3RIL grown at 37 °C in LB medium supplemented with ampicillin (50 mg/L) and kanamycin (50 mg/L), respectively, and induced by addition of 1 mM IPTG when cells reach about 1.0 optical densities at 600 nm. After growing overnight, we prepared the extracts and performed the Western blot as previously described.[28] For the GST pull-down experiment, extracts containing GST-SSO0434 were loaded on a column containing gluthatione-sepharose matrix that was washed with 10 vol of PBS 1× buffer at room temperature (RT). Then, equal amounts of extracts containing SSO0433-S-tag were loaded on the matrix in which was bound GST-SSO0434 and on a brand new gluthatione-sepharose matrix. Again, the two columns were washed and samples analyzed by two Western blots with polyclonal anti-GST-tag and anti-S-tag antibodies (Figures 4C,D).

To search for mutant alleles of *sui-1*, amplification products from chromosomal DNA, obtained by using the primers SUI15/SUI1STOP and SUI1BamHI/SUI1XhoI (Supporting Table 1), were cloned in the plasmid vector pQE30UA and in the *BamHI*/*XhoI* sites of the plasmid pET29a, respectively. The inserts were completely sequenced.

The full-length gene encoding for SUI-1 was amplified from chromosomal DNA of *S. solfataricus* P2 strain and placed under transcriptional control of the T7 F10 promoter by means of PCR using the forward primer aIF1_FP and the reverse primer aIF1_RP (Supporting Table 1) obtaining vector pSUI_full.

The interrupted SUI-1 mutant was prepared by site-directed mutagenesis from the vector pSUI_full, by using the GeneTailor Site-Directed Mutagenesis System from Invitrogen and the oligonucleotides SUI-Mut and SUI-Rev (Supporting Table 1). Mutations convert full-length SUI-1 into the DG SSO5866/unclassified. Direct sequencing identified the plasmid containing the desired mutations and the mutant gene was completely resequenced.

Translation *in Vitro* of *sui-1*. The full-length *sui-1* gene was used as template for *in vitro* transcription with T7 RNA Polymerase (Fermentas). The runoff transcripts were purified on 6% polyacrylamide−8 M urea gels following standard procedures. The mRNA concentration was determined by measuring the absorbance at 260 nm. The *in vitro* translation reactions were performed as described before.[25] The samples contained in a final volume of 25 μL: 10 mM KCl, 20 mM Tris/HCl (pH 7), 20 mM $MgCl_2$, 7 mM β-mercaptoethanol, 3 mM ATP, 1 mM GTP, 5 μg of *S. solfataricus* tRNA, 1 μL (10 mCi) of [^{35}S]-methionine (Amersham Bioscience Biotech), 5 μL of *S. solfataricus* strain P2 extract, 0.4 mM of mRNA and 0.4−1.6 mM of aIF1 mRNA. The samples were incubated for 40 min at 70 °C, then resolved on a 15% SDS polyacrylamide gel, and the radioactive bands were visualized by an Instant Imager apparatus.

Data Availability. The deposited Accession Numbers of the *S. solfataricus* genes sequenced in this paper are AY336520, AY336521, AY336522, and AY336523.

Results

Identification of Disrupted Genes in Archaeal Genomes. Disrupted Genes (DG) are coding sequences containing insertion/deletion of single or multiple DNA base pairs leading to truncated ORFs (if compared to the full-length homologues) or to ORFs separated by +/−1 frameshifts, stop codons, and DNA fragments of different length. Their identification in entire genomes is far from easy and they are not compiled in archaeal genomes. To analyze DG in this domain of life, we made a preliminary manual keyword search in the annotations of 16 archaeal genomes including organisms from the two main archaeal phyla (Crenarchaeota and Euryarchaeota) with diverse lifestyles, namely, obligate/facultative aerobes and anaerobes, methanogens, halophiles, (hyper)thermophiles, and acidophiles. The keywords used include terms indicating gene fragmentation, interruption, and truncation of the N- or C-terminal parts. Then, we analyzed the flanking regions at the 5′ or the 3′ ends of the disrupted ORFs identified with this first screening for the presence of contiguous short ORFs corresponding to the missing parts. All the ORFs obtained were

Table 1. Interrupted Genes in *S. solfataricus*[a]

		S. solfataricus, strain P2, 2977 protein coding genes, 34 interrupted genes	
entry	ORFs	reported annotation	disruption[b]
1	SSO0273	Hypothetical protein	1T
	SSO5544	Hypothetical protein	
2	SSO0297	Transketolase	+1
	SSO0299	Transketolase	
3	SSO0434	O-sialoglycoprotein endopeptidase	−1
	SSO0433	O-sialoglycoprotein endopeptidase	
4	SSO1529	Dihydrolipoamide S-acetyltransferase	−1
	SSO1530	Dihydrolipoamide S-acetyltransferase	
5	SSO1787	Conserved hypothetical protein	−1
	SSO1788	Hypothetical protein	
6	SSO2020	Acetyl-CoA synthetase	−1
	SSO2021	Acetyl-CoA synthetase	
7	SSO5866	Protein translation factor SUI1 homologue	−1
	Unclassified[c]	−	
8	SSO10784	CoA-ligase/coenzyme F390 synthetase	−1
	SSO2627	CoA-ligase/coenzyme F390 synthetase	
9	SSO11867	α-fucosidase	−1
	SSO3060	α-fucosidase	
10	SSO5761	tRNA pseudouridine synthase subunit A	+1
	SSO0393	tRNA pseudouridine synthase subunit B	
11	SSO0142	Primase	+IS
	SSO0140	Primase	
12	SSO0450	Alanyl-tRNA synthetase truncated homologue	N-del
13	SSO0627	Methionyl-tRNA synthetase N-term homologue	N-del
14	SSO0782	SSV1 integrase fragment homologue	N-del
15	SSO1271	Alanyl-tRNA synthetase truncated homologue	N-del
16	SSO1281	Oligo/dipeptide transport, ATP binding protein	+IS
	SSO1279	Oligo/dipeptide transport, ATP binding protein	
17	SSO1342	Acetyl-CoA synthetase	+IS
	SSO1340	Acetyl-CoA synthetase	
18	SSO1562	Conserved hypothetical protein	Ins
	SSO1563	Conserved hypothetical protein	
19	SSO1653	Helicase of the snf2/rad54 family, hypothetical	+IS
	SSO1655	Helicase of the snf2/rad54 family, hypothetical	
20	SSO1903	Acetyl-CoA synthetase	+IS
	SSO9201	Acetyl-CoA synthetase	
21	SSO3191	Formate dehydrogenase homologue to ST0639	
	Unclassified[d]	−	Ins/1T
	SSO12272	homologue to ST0639	
22	SSO6520	homologue to SSO1007	
	Unclassified[e]	−	−1/1T/C-del
	SSO0790	Acetylornithine deacetylase homologue to SSO1007	
23	SSO8869	Ferredoxin	Ins
	SSO8958	Ferredoxin	
24	SSO1348	Second ORF in transposon ISC1904 (putative transposase)	+1
	SSO1347	Third ORF in transposon ISC1904 (putative transposase)	
25	SSO1497	First ORF in transposon ISC1904 (putative integrase/resolvase)	+1
	SSO8224	Hypothetical protein (putative integrase/resolvase)	
26	SSO1750	First ORF in transposon ISC1078	−1
	SSO1749	Second ORF in transposon ISC1078 (putative transposase)	
27	SSO1772	First ORF in transposon ISC1491 (transposase)	−1
	SSO1769	Second ORF in transposon ISC1491 (putative transposase)	
28	SSO1971	First ORF in transposon ISC1190 (putative transposase)	+1
	SSO1972	Second ORF in transposon ISC1190 (putative transposase)	
29	SSO1995	First ORF in transposon ISC1225 (putative transposase)	−1
	SSO9455	Second ORF in transposon ISC1225 (putative transposase)	
30	SSO2123	First ORF in transposon ISC1316 (putative transposase)	−1
	SSO2124	Second ORF in transposon ISC1316 (putative transposase)	
31	SSO7351	First ORF in transposon ISC1439 (putative transposase)	−1
	SSO1180	Second ORF in transposon ISC1439 (putative transposase)	
32	SSO8568	First ORF in transposon ISC1229 (putative transposase)	1T
	SSO1639	Second ORF in transposon ISC1229 (putative transposase)	
33	SSO9043	Transposase, putative amino-end fragment	C-del
34	SSO9135	Partial ORF from ISC1904 (integrase-resolvase)	+1
	SSO9134	Partial ORF from ISC1904 (integrase-resolvase)	

[a] Genes are listed in numerical order. For the sake of clarity, *S. solfataricus* protein encoding genes disrupted by single and multiple events are grouped in entries 1−10 and 11−23, respectively. Genes involved in mobile elements are ordered in entries 24−34. [b] N-del and C-del denote deletions of the N- and C-terminus, respectively; *n*T signifies the presence of *n* stop codons; −/+1 indicates −/+1 frameshifts, +IS and Ins stand for the introduction of an insertion sequence element or some other sequences over 10 nt, respectively. [c] ORF of 57 amino acids not identified in *S. solfataricus* genome, homologous to SUI1. [d] ORF of 55 amino acids not identified in *S. solfataricus* genome, homologous to ST0639. [e] ORF of 58 amino acids not identified in *S. solfataricus* genome, homologous to SSO1007. [f] N-terminal follows the C-terminal in the chromosome. Full list of interrupted genes in Archaea is provided in Supporting Table 2.

compared to the amino acidic sequences from all the available sequenced genomes leading to the identification of DG similar to full-length genes in the data bank. This analysis produced 98 genes from 8 different Archaea (Supporting Table 2), including 34 genes from the Crenarchaeon *S. solfataricus* (listed in Table 1). Among them, we detected the ORFs encoding for the α-fucosidase gene (entry 9) that is an authentic case of programmed −1 frameshifting.[26,28] Interestingly, with this method we found three ORFs (entries 7, 21, and 22 in Table 1) that are shorter than the threshold used by the annotation program, and for this reason, they were previously unidentified.

This manual approach is not exhaustive being based on the quality of the annotations of the sequenced genome. This explains the absence of interrupted genes in the genomes of *A. pernix*, *Halobacterium* sp. NRC-1, *M. jannaschii*, *M. acetivorans*, *M. mazei*, *M. thermoautotrophicum*, *P. horikoshii*, and *S. tokodaii*. In fact, a more detailed analysis described below demonstrated that interrupted genes are indeed present in at least 3 out of 8 of these Archaea. To complete our results, we used the Interrupted CoDing Sequence (ICDS) database (http://alnitak.u-strasbg.fr/ICDS/),[11] one of the few bioinformatics resources currently available to identify DG. The predicted ICDS are detected by a program relying on the analysis of physically adjacent genes in complete genomes that share a common homologue. The approach has been validated on the genomes of *Bacillus licheniformis* and *Mycobacterium smegmatis* and used to scan 116 genomes including five archaeal genomes showing 19, 59, 24, 71, and 33 ICDS entries for *A. pernix*, *A. fulgidus*, *P. abyssi*, *P. furiosus*, and *P. horikoshii*, respectively. These figures are much higher than those identified with our manual search (see Supporting Table 2). In fact the two approaches are complementary since none of the putative DG manually found in *A. fulgidus* and *P. abyssi* and only 3 out of the 10 found in *P. furiosus* were present in ICDS database. This is not surprising, in fact, most of the interruptions found in these three Archaea result from large deletions/insertions or truncations (Supporting Table 2) while ICDS database mainly compiles in-frame stop codons and small insertion/deletions.[11] Three such modifications were identified in *P. furiosus* by ICDS out of the five found with the manual method, confirming the validity of the bioinformatic tool.

Interrupted Genes in *S. solfataricus*. Considering the high number of DG manually detected in the *S. solfataricus* genome, we focused on this species for thorough analyses. First, we apply the ICDS program to investigate the genome of *S. solfataricus* leading to 178 entries, most of which are ORFs involved in transposable elements (Supporting Table 3). Interestingly, 60% (20 out of 34) of the entries manually identified were identified also by the ICDS detection program; those missing are mainly truncated ORFs. In fact, the score increases to 75% (15 out of 20) when DG interrupted by single events are considered. The ICDS data confirmed that DG are particularly abundant in the Crenarchaeon *S. solfataricus*.

To further investigate the characteristics of the interrupted genes, we restrict the analysis to protein encoding genes interrupted by single events such as +/−1 frameshifts and stop codons (entries 1−10 in Table 1) that have more chances to be functional *in vivo*. Therefore, we excluded the genes interrupted by large insertions/deletions and those involved in putative insertion sequence elements (IS) (entries 11−34). In addition, SSO11867/SSO3060 and SSO5761/SSO0393 (entries 9 and 10, respectively) were not further analyzed. The former is expressed by programmed −1 frameshifting[28] while SSO5761/SSO0393, together with its homologues from *A. pernix* and *S. tokodaii*, is expressed *in vivo* by mRNA splicing.[29]

By using the BLAST program, we searched the sequenced archaeal genomes for homologues to the *S. solfataricus* entries 1−8. This analysis revealed that SSO0297/SSO0299 and SSO0434/SSO433 (entries 2 and 3) have, respectively, 31 and 37 interrupted archaeal homologues. Among these, interrupted homologues of SSO0434/SSO0433 are also present in *A. pernix* and *S. tokodaii*, though the manual search of their genomes did not revealed interrupted genes, confirming that the annotations biased the screening.

The conservation of these DG in several archaeal genomes sequenced by different consortia strongly suggests that these genes contain authentic disruptions. The position of the disruption is remarkably conserved in all the genes (see a small survey in Figure 1), indicating that the interruptions did not occurred randomly as it would be expected for inactive pseudogenes, but, rather, that the genes are selectively maintained in their interrupted status for functional or regulation reasons.

To experimentally test whether the genes in the entries 1−8 were transcribed as independent or joined ORFs, we performed RT-PCR experiments. It is worth noting that all the RT-PCR experiments always produced cDNA fragments, demonstrating that the DG are co-transcribed *in vivo* (Figure 2).

To test if these were authentic interruptions, sequencing artifacts, or if they were removed by post-transcriptional editing events, we sequenced DNA fragments of about 350−700 nt in the region between the contiguous ORFs of the entries 1−8 amplified from the genomic DNA and from total RNA preparations by RT-PCR. The sequencing of the genomic DNA and of the total RNA produced identical results for all the analyzed genes, indicating that no RNA editing events occurred.

The resequencing confirmed that the DG in entries 2−5 are authentic interrupted genes. Instead, the sequence of the genes in the entries 1 and 6−8 differed in single or multiple bases from those in the deposited genome, allowing merging the separated ORFs into full-length genes. Remarkably, the resequencing allowed to complete SSO5866 with an unclassified downstream ORF, resulting in a gene encoding for a putative protein of 101 amino acids similar to the universal translation initiation factor SUI-1.[30] The merged unclassified ORF, encoding for a polypeptide of 57 amino acids homologue to other archaeal SUI-1 proteins, was not annotated in the *S. solfataricus* genome.

Shotgun Proteomic analysis of *S. solfataricus* Interrupted Genes. The functionality of the interrupted genes of *S. solfataricus* was tested by a shotgun proteomic analysis: cellular extracts of this archaeon were searched for peptides translated from the ORFs listed in entries 1−34 (Table 1). To this aim, we prepared an Interruped Genes Data set file (IGD) containing the translated ORFs listed in Table 1 (entries 1−34) and all the possible polypeptides resulting from the merged ORFs. For instance, for the ORFs separated by an in-frame stop codon, we prepared a list of 20 different protein sequences in which the in-frame stop codon was translated into one of the 20 different amino acids. Instead, for ORFs separated by +/−1 frameshifts, the merged sequences were designed by planning a single +1 or −1 shift (depending on the interruption found) in all the possible positions of the region of overlap between the two ORFs considered. Finally, for truncated genes or for those showing large insertions/deletions and/or multiple interruptions, we just considered the ORFs deposited in the *S.*

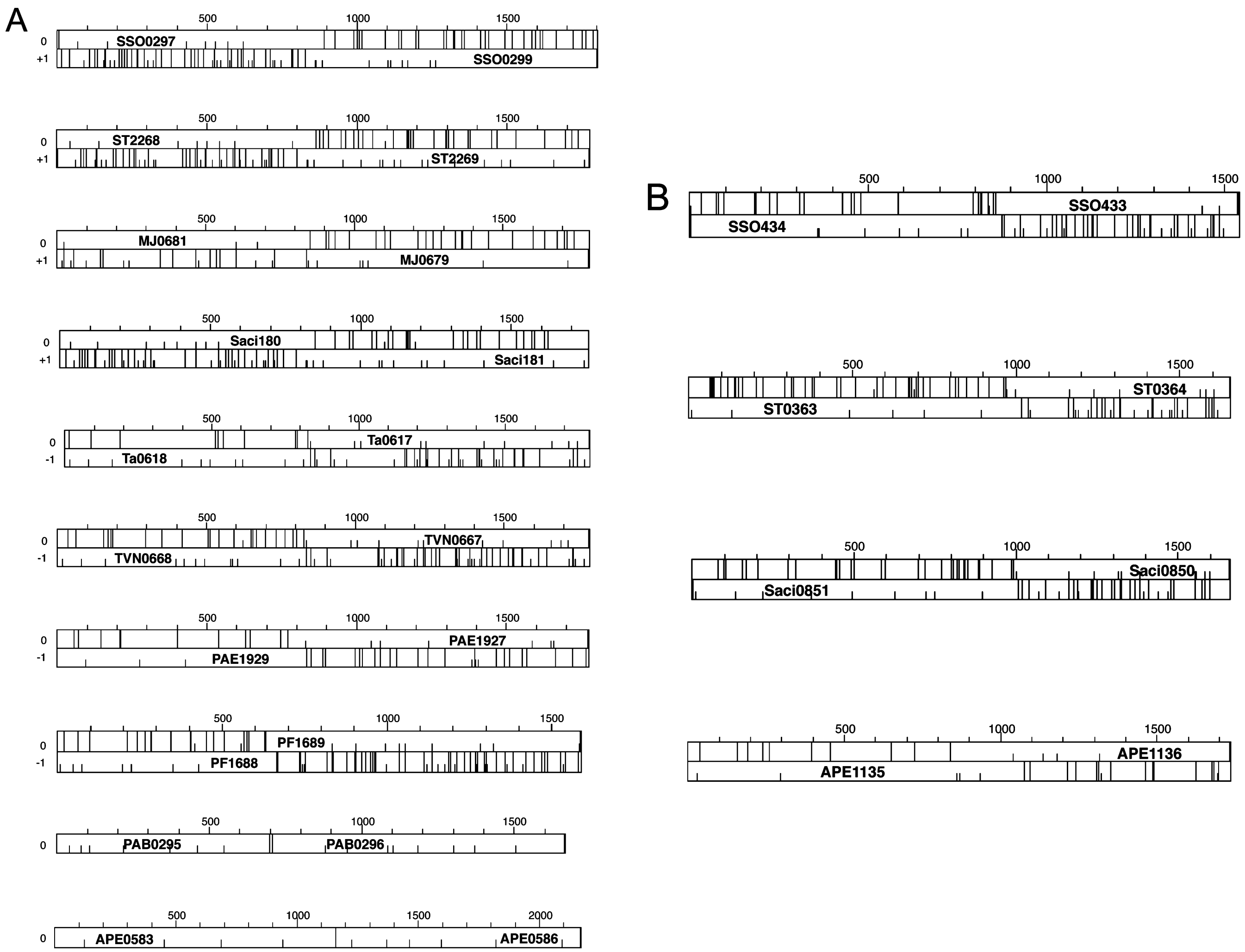

Figure 1. Schematic view of conserved interrupted genes. (A) Interrupted archaeal transketolases (entry 2); (B) interrupted archaeal O-sialoglycoprotein endopeptidases (entry 3). Long and short vertical lines represent stop and ATG codons, respectively. Minor initiation codons are not represented; thus, the TTG initiation codon of SSO0433 is not shown. The translational frames zero, +1, −1, and the names of the ORFs are indicated. ST, *S. tokodaii*; MJ, *M. jannaschii*; Saci, *Sulfolobus acidocaldarius*; Ta, *T. acidophilum*; TVN, *Thermoplasma volcanium*; PAE, *Pyrobaculum aerophilum*; PF, *P. furiosus*; PAB, *P. abyssi*; APE, *A. pernix*.

solfataricus genome since the events merging these coding sequences were not easy to foresee.

Protein extracts obtained from the *S. solfataricus* cells on which we performed the DNA/RNA amplification experiments described above were analyzed by liquid chromatography online tandem mass spectrometry (LCMSMS). Mass spectrometry data were searched against the *S. solfataricus* genome database, our IGD, and the reversed human proteins sequences as background. The description of the peptides identified in *S. solfataricus*, together with the results of a proteomic study of Chong and Wright,[31] is reported in Table 2. The LCMSMS run resulted in 128 655 peptide identifications from 10 528 distinct peptides that identified in total 1590 proteins, corresponding to 53.4% of all predicted ORF. Despite some discrepancies, our data are in good agreement with those previously reported.[31] For instance, the total number of identified proteins in our study is slightly lower than that reported on the same organisms in the other proteomic study. This could be explained by the combination of different techniques used by these authors that possibly increased the score.[31] In addition, it should be pointed out that we grew the *S. solfataricus*, strain P2, in a rich medium containing sucrose, yeast extract, and casaminoacids, while the data available from the literature are obtained from cells grown in a minimal medium including glucose and Wolfe's vitamins stock.[31] Therefore, the discrepancies in the expression of certain genes could be the result of a metabolic regulation rather than a technical artifact. Remarkably, according to the two proteomic studies, 22 out of the total 34 interrupted coding sequences reported in Table 2 revealed expressed products *in vivo*.

Most of the peptides detected by LCMSMS were identical to those found in the *S. solfataricus* genome database. This result would suggest that the ORFs of the interrupted genes are expressed independently. On the other hand, we might have missed some peptides encompassing the two ORFs because the corresponding full-length proteins were expressed at low level. For instance, the full-length α-L-fucosidase (entry 9) was not found in our proteomic screening (Table 2), but it was identified by Western blot in *S. solfataricus* extracts.[28]

Strikingly, LCMSMS experiments revealed two different peptides for the same region of the *sui-1* gene (entry 7 in Table 2). As expected, one peptide (underlined in Table 2) corresponds to the full-length gene that we identified after resequencing the genomic DNA (Figure 3A). However, a second

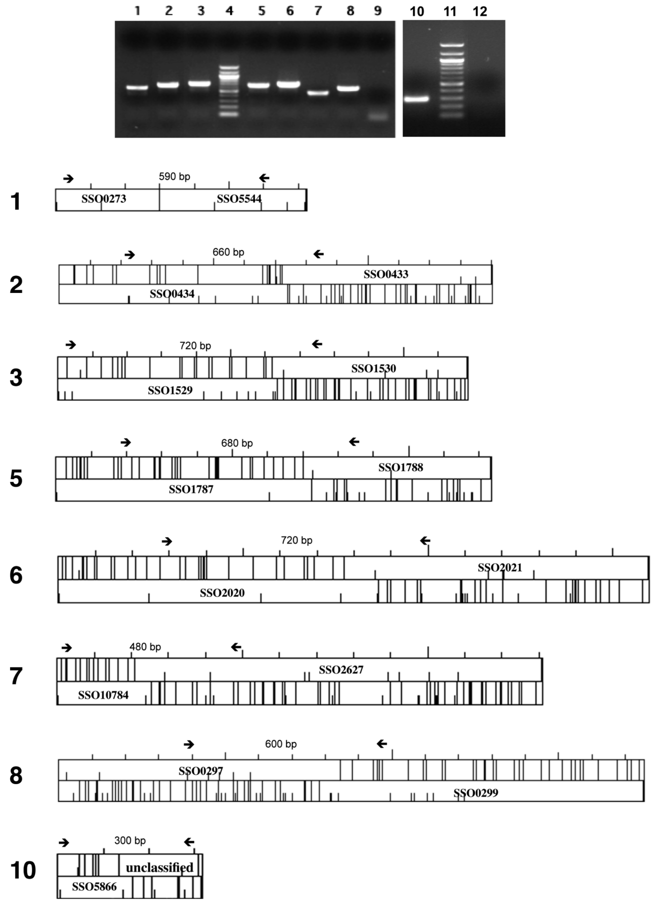

Figure 2. Reverse Transcriptase-PCR of interrupted genes of *S. solfataricus*. Lane 1, SSO0273/SSO5544 (entry 1 in Table 1); lane 2, SSO0434/SSO0433 (entry 3); lane 3, SSO1529/SSO1530 (entry 4); lane 4 and 11, 100 bp ladder; lane 5, SSO1787/SSO1788 (entry 5); lane 6, SSO2020/SSO2021 (entry 6); lane 7, SSO10784/SSO2627 (entry 8); lane 8, SSO0297/SSO0299 (entry 2); lane 9, control (same as lane 8, but without reverse transcriptase and with *Taq* polymerase); lane 10 SSO5866/unclassified (entry 7); lane 12, control (same as lane 10, but without reverse transcriptase and with *Taq* polymerase). The ORFs in the different frames, corresponding to the lanes of the agarose gel, are schematically represented. The amplification primers are shown with arrows and the expected size (base pairs) of the amplification products are indicated on top of the ORFs.

peptide corresponds to the interrupted gene present in the *S. solfataricus* database and it would result from a translational −1 frameshifting in the region of overlap between the ORF SSO5866 and the downstream unclassified ORF (Figure 3B). The second peptide sequence cannot result from the full-length gene even as a consequence of a single mutation or other recoding events. The genome of *S. solfataricus* showed only one copy of *sui-1*; therefore, we hypothesize that in the population from which we prepared the extracts there was a mix of cells in which the copy of the *sui-1* gene was either in a full-length or in a interrupted form. To test this eventuality, we sequenced several *sui-1* clones isolated by amplification from the genomic DNA prepared from *S. solfataricus*. Interestingly, 21% (7 out of 33) of the sequenced clones contained mutations (6 single mutants, one double mutant). No specific hot spots for mutations were identified, but they were spread along the sequence of full-length *sui-1* (Figure 3A). We could not identify the mutant shown in Figure 3B; however, it is worth noting that *sui-1* tends to accumulate mutations. Moreover, our results further suggest that *sui-1* is present in the genome as a single copy, otherwise we should find sequences resulting from the two alleles. Possibly, the disrupted *sui-1* leading to the peptide observed in our high-throughput proteomic analysis resulted from mutations of the full-length allele. In our study, we could not find the peptides resulting from the different mutants shown in Figure 3A; this is not surprising: the mutations fall in regions distant from that overlapping the ORFs SSO5866/unclassified, and, therefore, they were not included in the IGD file in which the SUI-1 protein sequences outside the overlapping region are invariant. More interestingly, our LCMSMS data clearly demonstrated that the interrupted form of *sui-1* gene produced a full-length polypeptide, presumably, by programmed-1 frameshifting.

Analysis of the Expression of Interrupted Genes in *S. solfataricus*. The mechanism of expression of three genes, namely, SSO0297/SSO0299, SSO0434/SSO0433, and SSO5866/unclassified (entries 2, 3, and 7 in Table 1), was analyzed in more detail. SSO0297/SSO0299 and SSO0434/SSO0433 were selected as authentic DG with conserved interrupted homologues in Archaea. In addition, we further investigated on SSO5866/unclassified, though the resequencing merged the two ORFs in a single full-length gene, because the proteomic study revealed a polypeptide produced *in vivo* by the frameshifted DG.

SSO0297/SSO0299 are homologues of the transketolase from *Saccharomyces cerevisiae*; the inspection of the available 3D-structure of this enzyme (pdb code 1AYO) showed that SSO0297 and SSO0299 correspond exactly to its N- and C-terminal domains (Figure 4A), suggesting that the two ORFs could be expressed independently in *S. solfataricus* forming a heterodimer in the native structure.

The SSO0434/SSO0433 ORFs, annotated in *S. solfataricus* genome as O-syaloglycoprotein endopeptidases, are homologues to MJ1130 from *M. jannaschii* encoding a Kae1/Bud32 fusion protein for which the three-dimensional (3D) structure has been recently obtained.[32] Figure 4B shows that SSO0434 and SSO0433 correspond to Kae1 and Bud32, respectively, with 52% and 38% identities. Kae1 (kinase associated endopeptidase 1) belongs to the small set of about 60 universal proteins present in all members of the three domains of life and corresponds to the human ortholog OSGEP (O-SyaloGlycoprotein EndoPeptidase).[33] Bud32 is present in all Eukaryotes and Archaea and corresponds to the human p53-related protein kinase (PRPK) involved in the phosphorylation of p53.[22] Interestingly, the biochemical characterization and the 3D-structure determination of Kae1 from *P. abyssi* demonstrated that this enzyme is an iron-protein, devoid of endopeptidase activity *in vitro*, showing a novel type of ATP-binding site, able to bind DNA, and exhibiting a class I apurinic endonuclease activity.[32] The orthologs of *kae1* and *bud32* are often found juxtaposed or even fused in archaeal genomes.[34] Remarkably, *kae1*/*bud32* genes separated by single frameshifting are present in the Crenarchaeota phylum in all cases but one (Supporting Table 4). Expression in *E. coli* of the SSO0434/SSO0433 ORFs gave no evidence of the expression of a single polypeptide from two ORFs by events of programmed −1 frameshifting (not shown); thus, we concluded that they were independently expressed. In particular, SSO0433 is expressed, presumably, from the first TTG codon as in this ORF there is no ATG codon close enough to express a large protein. SSO0434 fused at its N-terminus with a GST-tag and SSO0433 fused at the C-

Table 2. Proteomic Analysis of Interrupted Genes of *S. solfataricus*[a]

entry	ORFs	disruption	peptides identified by LCMSMS	peptides in 31
1	SSO0273	1T	√[b]	•
	SSO5544		√	•
2	SSO0297	+1	√	•
	SSO0299		√	•
3	SSO0434	−1	√	-
	SSO0433		√	-
4	SSO1529	−1	-	-
	SSO1530		-	-
5	SSO1787	−1	-	•
	SSO1788		-	•
6	SSO2020	−1	√	•
	SSO2021		√	-
7	SSO5866	−1	EVTIIEGLGGNDSELK and	-
	Unclassified		EVTIIEGIREVNDSEL	-
8	SSO10784	−1	√	-
			YAYPHGGDFL	
	SSO2627		√	•
9	SSO11867	−1	-	-
	SSO3060		-	•
10	SSO5761	+1	√	-
	SSO0393		√	•
11	SSO0142	+IS	-	•
	SSO0140		-	•
12	SSO0450	N-del	√	•
13	SSO0627	N-del	√	•
14	SSO0782	N-del	-	-
15	SSO1271	N-del	-	-
16	SSO1281	+IS	-	-
	SSO1279		-	-
17	SSO1342	+IS	-	-
	SSO1340		-	-
18	SSO1562	Ins	√	-
	SSO1563		-	-
19	SSO1653	+IS	-	•
	SSO1655		-	-
20	SSO1903	+IS	√	•
	SSO9201		-	-
21	SSO3191		√	-
	Unclassified	Ins/1T	-	-
	SSO12272		-	-
22	SSO6520		-	-
	Unclassified	−1/1T/C-del	-	-
	SSO0790		-	-
23	SSO8869	Ins	-	-
	SSO8958		-	•
24	SSO1348	+1	-	-
	SSO1347		-	-
25	SSO1497	+1	-	-
	SSO8224		-	-
26	SSO1750	−1	-	-
	SSO1749		-	•
27	SSO1772	−1	√	-
	SSO1769		-	-
28	SSO1971	+1	-	-
	SSO1972		-	-
29	SSO1995	−1	-	-
	SSO9455		-	-
30	SSO2123	−1	-	•
	SSO2124		-	-
31	SSO7351	−1	-	-
	SSO1180		-	•
32	SSO8568	1T	-	-
	SSO1639		-	-
33	SSO9043	C-del	-	•
34	SSO9135	+1	-	-
	SSO9134		-	-

[a] The listing order and the symbols used are the same of Table 1. [b] This symbol indicates that the sequence of one or more peptides matches the sequences available in the *S. solfataricus* genome database.

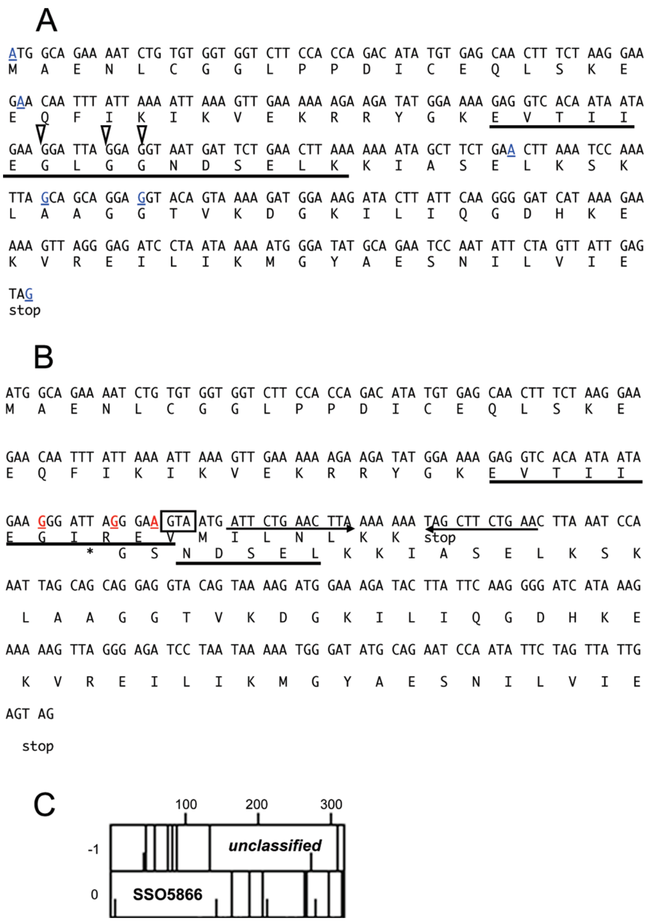
A
ATG GCA GAA AAT CTG TGT GGT GGT CTT CCA CCA GAC ATA TGT GAG CAA CTT TCT AAG GAA
M A E N L C G G L P P D I C E Q L S K E
GAA CAA TTT ATT AAA ATT AAA GTT GAA AAA AGA AGA TAT GGA AAA GAG GTC ACA ATA ATA
E Q F I K I K V E K R R Y G K E V T I I
GAA GGA TTA GGA GGT AAT GAT TCT GAA CTT AAA AAA ATA GCT TCT GAA CTT AAA TCC AAA
E G L G G N D S E L K K I A S E L K S K
TTA GCA GCA GGA GGT ACA GTA AAA GAT GGA AAG ATA CTT ATT CAA GGG GAT CAT AAA GAA
L A A G G T V K D G K I L I Q G D H K E
AAA GTT AGG GAG ATC CTA ATA AAA ATG GGA TAT GCA GAA TCC AAT ATT CTA GTT ATT GAG
K V R E I L I K M G Y A E S N I L V I E
TAG
stop
B
ATG GCA GAA AAT CTG TGT GGT GGT CTT CCA CCA GAC ATA TGT GAG CAA CTT TCT AAG GAA
M A E N L C G G L P P D I C E Q L S K E
GAA CAA TTT ATT AAA ATT AAA GTT GAA AAA AGA AGA TAT GGA AAA GAG GTC ACA ATA ATA
E Q F I K I K V E K R R Y G K E V T I I
GAA GGG ATT AGG GAA GTA ATG ATT CTG AAC TTA AAA AAA TAG CTT CTG AAC TTA AAT CCA
E G I R E V M I L N L K K stop
* G S N D S E L K K I A S E L K S K
AAT TAG CAG CAG GAG GTA CAG TAA AAG ATG GAA AGA TAC TTA TTC AAG GGG ATC ATA AAG
L A A G G T V K D G K I L I Q G D H K E
AAA AAG TTA GGG AGA TCC TAA TAA AAA TGG GAT ATG CAG AAT CCA ATA TTC TAG TTA TTG
K V R E I L I K M G Y A E S N I L V I E
AGT AG
stop
C
100
200
300
-1
0
unclassified
SSO5866

Figure 3. Sequence of the *sui-1* genes. (A) Sequence of the full-length gene obtained after the resequencing of the *S. solfataricus* genomic DNA. The nucleotide mutated in the *sui-1* clones sequenced, namely, clone #1: A1 deletion; #2 A62G (Glu21Gly) and A168G (Glu56Glu); #3 and #4 G184T (Ala62Ser); #5 G193A (Gly65Ser); #6 G303T (Amber101Tyr); #7 G303A (Amber101Ochre) are underlined in blue. (B) Sequence of the interrupted gene SSO5866/uncharacterized ORF (entry 7 in Table 1) deposited in the *S. solfataricus* data bank. The amino acid sequences of the peptides identified by LCMSMS are underlined; the differences between the two *sui-1* genes are indicated by arrowheads (in A) and underlined in red (in B). The codon in which frameshifting might occur is boxed. The putative stem-loop is indicated with arrows. (C) Schematic view of the SSO5866/unclassified ORFs. For details see the legend of Figure 1.

terminus with a S-tag were expressed in *E. coli* producing the expected molecular weights identified by Western blot (61 and 28 kDa for GST-SSO0434 and SSO0433-S-tag, respectively) (Figure 4C,D). SSO0434 was expressed in *E. coli* with low efficiency, as we could not observe SSO0434 in *E. coli* extracts by Western blot (Figure 4C, lane 6). However, the GST pull-down experiments indicated that the two proteins interact *in vitro* (Figure 4D, lane 1).

To test the mechanism of expression of SSO5866/unclassified, we performed experiments of translation *in vitro* by using *S. solfataricus* extracts; to this aim, the full-length *sui-1* gene from the *S. solfataricus*, strain P2, was cloned under the control of a T7 polymerase promoter. Full-length *sui-1* was used as template for site-directed mutagenesis experiments to insert two guanines and the mutation G to A obtaining the interrupted *sui-1*; the sequence of the two genes is shown in Figure 3, psnels A and B, respectively. The *in vitro* translation products

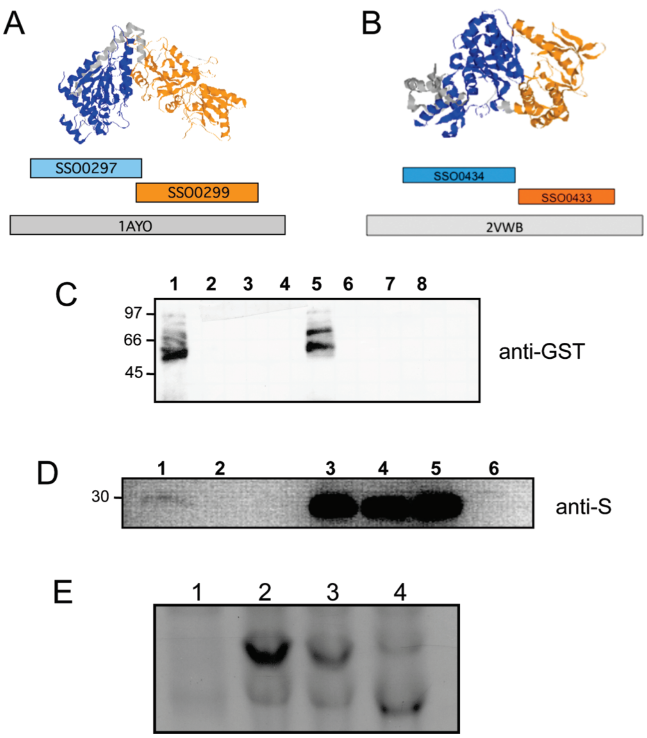

Figure 4. Functional analysis of interrupted genes in *S. solfataricus*. (A and B) Prediction analysis of putative structural domains of SSO0297/SSO0299 and SSO0434/SSO0433. Full-length ORFs are shown as gray boxes; N- and C-terminal ORFs of *S. solfataricus* are colored in blue and orange, respectively. The 3D-structures follow the same color code. The regions in gray in the 3D-structure could not be found in the archaeal sequences by alignment of the *S. solfataricus* translated ORFs to the full-length translated genes. The gene IDs and the PDB accession numbers are shown within the boxes. (C) Anti-GST Western blot of the GST pull-down experiment on SSO0434/SSO0433. Lane 1, gluthatione-sepharose matrix loaded with GST-SSO0434 and SSO0433-S (GST pull-down); lane 2, gluthatione-sepharose matrix loaded with SSO0433-S only; lanes 3 and 4 flow-through of the same matrices of 1 and 2; lane 5 gluthatione-sepharose matrix loaded with GST-SSO0434 only; lane 6, cellular extract containing GST-SSO0434; lane 7, cellular extract containing SSO0433-S; lane 8, wash of the GST pull-down. (D) Anti-S Western blot of the GST pull-down experiment on SSO0434/SSO0433. Lane 1, gluthatione-sepharose matrix loaded with GST-SSO0434 and SSO0433-S (GST pull-down); lane 2, gluthatione-sepharose matrix loaded with SSO0433-S only; lanes 3 and 4 flow-through of the same matrices in 1 and 2; lane 5, cellular extract containing SSO0433-S-tag; lane 6, cellular extract containing GST-SSO0434. (E) Analysis of the *in vitro* translation of SSO5866/uncharacterized; 10 μL of each sample was loaded on 17.5% acrylamide-SDS gel and the newly synthesized proteins were revealed by autoradiography. Lane 1, no mRNA added; lane 2, control mRNA encoding for a *S. solfataricus* protein of 10 kDa (ORF104);[25] lane 3, full-length *sui-1*; lane 4, truncated *sui-1*.

were analyzed by autoradiography of a SDS-PAGE revealing that full-length *sui-1* gene produced a clear band of the expected molecular weight of about 11 kDa. The interrupted *sui-1* produced, as expected, a major band of about 6 kDa, resulting from the early termination of translation (theoretical molecular weight of about 6.1 kDa). Remarkably, however, a tiny but clear band of the same molecular weight as the full-length *sui-1* was also consistently produced (Figure 4E). The identity of the 6 and 11 kDa bands in lanes 3 and 4 of Figure

4E as SUI-1 was confirmed by Western blotting with anti-SUI-1 polyclonal antibodies (not shown). Quantification of the full-length and truncated bands of three replicate experiments with the ImageJ program enabled us to evaluate an approximate 15% of frameshifting, a percentage similar to that observed for the SSO11866/3060, another *S. solfataricus* split-gene.[28] Altogether, the above data unequivocally demonstrate that the ribosomes of *S. solfataricus* can decode the interrupted *sui-1* by programmed −1 frameshifting producing a full-length polypeptide from the two ORFs SSO5866/unclassified.

Discussion

The assignment of correct gene and protein sequence is often hampered by technical bottlenecks and depends on reliable sequence data. This is particularly true for prokaryotic genomes in which annotations largely rely on computer programs.[11] The subsequent challenge is to understand whether authentic interruptions lead to nonfunctional pseudogenes (not transcribed nor translated), or to functional interrupted genes whose expression is regulated by post-transcriptional events. Reliable methods to predict these features are still limited; thus, usually, interrupted genes are not considered in biochemical studies and their functionality is determined serendipitously. Therefore, insights into the possible expression mechanisms of interrupted genes and the role played *in vivo* are urgently needed. To this aim, we searched 16 archaeal genomes for interrupted genes, we tested if they were expressed *in vivo* by a high-throughput proteomic screening, and we analyzed their functionality.

A manual search supported by the ICDS program revealed an high number of disrupted genes (DG), especially abundant in the Crenarchaeon *S. solfataricus* where we grouped DG in three categories: (i) those showing single +/−1 frameshifts or stop codons, (ii) those truncated or interrupted by large insertions/deletions, and (iii) those involved in mobile elements. We functionally characterized those disrupted by single events. The mechanism of expression of two of these (tRNA pseudouridine synthase and a α-fucosidase) was experimentally characterized as mRNA splicing and programmed −1 frameshifting, respectively.[28,29] Four genes (entries 1 and 6−8 in Table 1), whose resequencing produced full-length coding sequences, were the result of sequencing errors, occurring in archaeal genomes, on the average, every 5000−10000 nt.[35] For the remaining genes, several lines of evidence demonstrated that they are expressed *in vivo* and are maintained in this status for functional reasons. First, the search of archaeal genomes by using *S. solfataricus* interrupted genes revealed DG archaeal homologues within the interruption at the same position. Second, RT-PCR experiments demonstrated that all the genes interrupted by single events are transcribed in *S. solfataricus*, excluding the possibility that they are nonfunctional pseudogenes. Third, our shotgun proteomic analysis and a high-throughput screening[11] allowed the unequivocal identification of peptides in three out of the four remaining DG interrupted by single events (entries 2, 3, and 5 in Table 2). The last DG SSO1529/SSO1530 (entry 4), a putative gene annotated as a dihydrolipoamide S-acetyltransferase for which we could not identify confirmed peptides, is transcribed in this archaeon; however, further studies are required to confirm its expression. It is worth noting that we identified peptides matching 48% of all the DG found in *S. solfataricus*, demonstrating that these genes were expressed *in vivo*.

The mechanism by which the interrupted genes are expressed in *S. solfataricus* cannot be simply foreseen. For most of the DG we could not identify peptides in the region of the gene that include the interruption, suggesting that these ORFs are translated independently. This was experimentally confirmed in the case of the *S. solfataricus* ortholog of Kae1/Bad32 (SSO0434/SSO0433) by expressing the two ORFs in *E. coli*. Nevertheless, we could not find peptides from the frameshifted gene encoding for the α-L-fucosidase, which produced a full-length polypeptide by programmed −1 frameshifting *in vivo*.[28] Therefore, it cannot be ruled out that the proteomic analysis misses some peptides also in other interrupted genes.

The high-throughput proteomic analysis gave interesting results on the ORF SSO5866 (entry 7) encoding for a putative protein similar to the universal translation initiation factor SUI-1[30] but including only the first 50 amino acids of the latter, indicating the presence of a split gene for this protein. The resequencing of the genomic DNA of the *S. solfataricus*, however, retrieved an uninterrupted gene encoding a full-length SUI-1; this was confirmed by the proteomic analysis revealing the presence in the *S. solfataricus* extracts of a peptide perfectly matching the product of the full-length gene (Figure 3A). While the above evidence suggested that the split gene in the data bank could be due to a sequencing error, our proteomic analysis intriguingly identified also a second peptide whose sequence was only explainable by supposing the presence of an interrupted *sui-1* gene expressed by programmed −1 frameshifting (Figure 3B). Remarkably, this was confirmed by experiments of *in vitro* translation, demonstrating that the interrupted *sui-1* gene can drive the expression of a full-length polypeptide in *S. solfataricus* extracts.

In programmed −1 frameshifting, the translating ribosomes are induced to shift to an alternative overlapping reading frame one nucleotide 5′-wards of the mRNA. This process is triggered by several elements in the mRNA: a slippery sequence (usually a heptanucleotide) has the function of favoring the tRNA misalignment and it is the site where the shift takes place.[15,36] Other elements flanking the slippery sequence, namely, a codon for a low-abundance tRNA, a stop codon, an mRNA secondary structure, and, in bacteria, a Shine-Dalgarno sequence, enhance frameshifting by pausing the translating ribosome on the slippery sequence.[17,36,37] Without direct evidence, the identification of the regulatory sequences of the putative translational recoding of *sui-1* is rather uncertain and specific experiments are required to this aim. The inspection of the sequence of the peptide experimentally determined and the region of overlap between the ORFs encoding for interrupted *sui-1* suggested that the frameshifting occurs in the site G-GAA-GUA that is followed by a putative stem-loop (Figure 3B). Such frameshifting sites do not belong to the list of the experimentally identified slippery sequences promoting programmed −1 frameshiftings (http://recode.genetics.utah.edu/); however, this list only includes experimentally determined slippery sequences and the sole case known in Archaea has been elucidated only recently by our group.[28] In addition, the regulatory signals of programmed −1 frameshifting are not widely conserved: in prokaryotes, the stimulatory signals include tetra-, hexa-, or heptanucleotide motifs (see 36, 38 and references therein),[39] and the mRNA secondary structure is not always present.[15]

SUI-1, which is essential in yeast and present in all Archaea in which the genomes were sequenced, forms the translation initiation complex and monitors the maintenance of the correct translational reading frame in eukaryotes.[30] The function of this translation factor in Archaea and in *E. coli* is still unknown; interestingly, it has been proposed that it could govern programmed −1 frameshifting as a *trans*-acting factor.[41] From our data, it would be tempting to speculate that SUI-1 regulates its own expression in *S. solfataricus* by programmed −1 frameshifting. This, together with the gene encoding for the α-L-fucosidase, would be the second documented case of a gene expressed by this translational mechanism in *S. solfataricus*. However, our results cannot exclude that the peptide identified by proteomic analysis could be expressed by transcriptional slippage or RNA editing. However, the A or T stretches observed in typical transcriptional frameshifting sites of bacterial genomes[40] are missing in the putative frameshifting region of *sui-1*. Further studies are required to define the precise mechanism of programmed −1 frameshifting for this gene.

Surprisingly, the proteomic studies shown here demonstrated that 64% of the genes suffering truncations, large deletions/insertions, or multiple interruptions (entries 11−23, 33) are expressed *in vivo*; possibly the gene products are independent polypeptide domains forming functional multimeric proteins. This resembles the case of the split genes frequently found in the genome of *Nanoarchaeum equitans*. Interestingly, in this organism, the ORFs encoding for the two subunits of an alanyl-tRNA synthetase are separated by half of the chromosome and a fully functional enzyme was obtained by expressing the ORFs independently in *E. coli* and combining the two polypeptides.[42] The presence of expressed split genes in *N. equitans* has been interpreted as both a derived[43] and an ancestral character.[44] Our data might be useful to understand the origin of these genes under an evolutionary perspective.

Interestingly, we showed that about 75% of all the interrupted genes found in the 16 archaeal genomes analyzed are present in the Crenarchaeota *S. solfataricus* and *P. aerophilum* (Supporting Table 2). The reason for the high frequency of interrupted genes in these organisms is not clear. During the sequencing of the genome of *P. aerophilum*, the authors explained the elevated number of frameshifts with the lack of DNA mismatch repair activity in this organism.[45] Similarly, a relevant degree of gene sequence variability, possibly resulting from the movement of insertion elements during cell culture, was also observed during the sequencing of the genome of *S. solfataricus*[46,47] and we reported here that the *sui-1* gene tends to accumulate mutations with an unusually high frequency. If this is a peculiarity of *sui-1* gene, and if it occurs also in other genes and has an effect on interrupted genes, it is not clear and would merit further investigations. It was pointed out that in *P. aerophilum* the split genes may generate diversity responding to changing environments, but a permanent mutator lifestyle could not be feasible.[44] Our study suggests that yet uncharacterized mechanisms might be used to express the interrupted genes: high expectations are justified for this field in the future.

Acknowledgment. We want to thank Maurizio Iaccarino and Maria Ciaramella for useful discussions and one of the reviewers for the transcriprional frameshifting and RNA editing hypothesis of the *sui-1* gene. B.C.-P., P.L., M.R. and M.M. were supported by the project MoMa n. 1/014/06/0 of the Agenzia Spaziale Italiana. We thank the TIGEM-IGB DNA sequencing core for the DNA sequencing. The IBP-CNR belongs to the Centro Regionale di Competenza in Applicazioni Tecnologico-Industriali di Biomolecole e Biosistemi.

Supporting Information Available: Tables of synthetic oligonucleotides, interrupted genes in Archaea, interrupted coding sequences in *S. solfataricus* strain P2, and Archaeal homologues of Kae1/Bad32. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Zhang, Z.; Carriero, N.; Gerstein, M. Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet.* **2004**, *20*, 62–67.

(2) Brosius, J.; Gould, S. J. On “genomenclature”: a comprehensive (and respectful) taxonomy for pseudogenes and other “junk DNA. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10706–10710.

(3) Harrison, P. M.; Gerstein, M. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J. Mol. Biol.* **2002**, *318*, 1155–1174.

(4) Balakirev, E. S.; Ayala, F. J. Pseudogenes: are they “junk” or functional DNA? *Annu. Rev. Genet.* **2003**, *37*, 123–151.

(5) Hirotsune, S.; Yoshida, N.; Chen, A.; Garrett, L.; Sugiyama, F.; Takahashi, S.; Yagami, K.; Wynshaw-Boris, A.; Yoshiki, A. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* **2003**, *423*, 91–96.

(6) Andersson, S. G.; Zomorodipour, A.; Andersson, J. O.; Sicheritz-Ponten, T.; Alsmark, U. C.; Podowski, R. M.; Naslund, A. K.; Eriksson, A. S.; Winkler, H. H.; Kurland, C. G. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **1998**, *396*, 133–140.

(7) Andersson, J. O.; Andersson, S. G. Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. *Mol. Biol. Evol.* **2001**, *18*, 829–839.

(8) Parkhill, J.; Wren, B. W.; Thomson, N. R.; Titball, R. W.; Holden, M. T.; Prentice, M. B.; Sebaihia, M.; James, K. D.; Churcher, C.; Mungall, K. L.; Baker, S.; Basham, D.; Bentley, S. D.; Brooks, K.; Cerdeno-Tarraga, A. M.; Chillingworth, T.; Cronin, A.; Davies, R. M.; Davis, P.; Dougan, G.; Feltwell, T.; Hamlin, N.; Holroyd, S.; Jagels, K.; Karlyshev, A. V.; Leather, S.; Moule, S.; Oyston, P. C.; Quail, M.; Rutherford, K.; Simmonds, M.; Skelton, J.; Stevens, K.; Whitehead, S.; Barrell, B. G. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **2001**, *413*, 523–527.

(9) Cole, S. T.; Eiglmeier, K.; Parkhill, J.; James, K. D.; Thomson, N. R.; Wheeler, P. R.; Honore, N.; Garnier, T.; Churcher, C.; Harris, D.; Mungall, K.; Basham, D.; Brown, D.; Chillingworth, T.; Connor, R.; Davies, R. M.; Devlin, K.; Duthoy, S.; Feltwell, T.; Fraser, A.; Hamlin, N.; Holroyd, S.; Hornsby, T.; Jagels, K.; Lacroix, C.; Maclean, J.; Moule, S.; Murphy, L.; Oliver, K.; Quail, M. A.; Rajandream, M. A.; Rutherford, K. M.; Rutter, S.; Seeger, K.; Simon, S.; Simmonds, M.; Skelton, J.; Squares, R.; Squares, S.; Stevens, K.; Taylor, K.; Whitehead, S.; Woodward, J. R.; Barrell, B. G. Massive gene decay in the leprosy bacillus. *Nature* **2001**, *409*, 1007–1011.

(10) Lawrence, J. G.; Hendrix, R. W.; Casjens, S. Where are the pseudogenes in bacterial genomes. *Trends Microbiol.* **2001**, *9*, 535–540.

(11) Perrodou, E.; Deshayes, C.; Muller, J.; Schaeffer, C.; Van Dorsselaer, A.; Ripp, R.; Poch, O.; Reyrat, J. M.; Lecompte, O. ICDS database: interrupted CoDing sequences in prokaryotic genomes. *Nucleic Acids Res.* **2006**, *34*, 338–343.

(12) Beier, H.; Grimm, M. Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. *Nucleic Acids Res.* **2001**, *29*, 4767–4782.

(13) Atkins, J. F.; Weiss, R. B.; Gesteland, R. F. Ribosome gymnastics—degree of difficulty 9.5, style 10.0. *Cell* **1990**, *62*, 413–423.

(14) Gesteland, R. F.; Weiss, R. B.; Atkins, J. F. Recoding: reprogrammed genetic decoding. *Science* **1992**, *257*, 1640–1641.

(15) Farabaugh, P. J. Programmed translational frameshifting. *Annu. Rev. Genet.* **1996**, *30*, 507–528.

(16) Baranov, P. V.; Gurvich, O. L.; Hammer, A. W.; Gesteland, R. F.; Atkins, J. F. RECODE 2003. *Nucleic Acids Res.* **2003**, *31*, 87–89.

(17) Namy, O.; Rousset, J. P.; Napthine, S.; Brierley, I. Reprogrammed genetic decoding in cellular gene expression. *Mol. Cell* **2004**, *13*, 157–168.

(18) Cobucci-Ponzano, B.; Rossi, M.; Moracci, M. Recoding in archaea. *Mol. Microbiol.* **2005**, *55*, 339–348.

(19) Jacobs, J. L.; Belew, A. T.; Rakauskaite, R.; Dinman, J. D. Identification of functional, endogenous programmed−1 ribosomal frameshift signals in the genome of *Saccharomyces cerevisiae*. *Nucleic Acid Res.* **2007**, *35*, 165–174.

(20) Moon, S.; Byun, Y.; Kim, H. J.; Jeong, S.; Han, K. Predicting genes expressed via−1 and +1 frameshifts. *Nucleic Acids Res.* **2004**, *32*, 4884–4892.

(21) Yao, A.; Charlab, R.; Li, P. Systematic identification of pseudogenes through whole genome expression evidence profiling. *Nucleic Acids Res.* **2006**, *34*, 4477–4485.

(22) Hecker, A.; Lopreiato, R.; Graille, M.; Collinet, B.; Forterre, P.; Libri, D.; van Tilbeurgh, H. Structure of the archaeal Kae1/Bud32 fusion protein MJ1130: a model for the eukaryotic EKC/KEOPS subcomplex. *EMBO J.* **2008**, *27*, 2340–2351.

(23) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.

(24) Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **1988**, *16*, 10881–10890.

(25) Condò, I.; Ciammaruconi, A.; Benelli, D.; Ruggero, D.; Londei, P. Cis-acting signals controlling translational initiation in the thermophilic archaeon *Sulfolobus solfataricus*. *Mol. Microbiol.* **1999**, *34*, 377–384.

(26) Cobucci-Ponzano, B.; Trincone, A.; Giordano, A.; Rossi, M.; Moracci, M. Identification of an archaeal alpha-L-fucosidase encoded by an interrupted gene. Production of a functional enzyme by mutations mimicking programmed−1 frameshifting. *J. Biol. Chem.* **2003**, *278*, 14622–14631.

(27) Wei, J.; Sun, J.; Yu, W.; Jones, A.; Oeller, P.; Keller, M.; Woodnutt, G.; Short, J. M. Global proteome discovery using an online three-dimensional LC-MS/MS. *J. Proteome Res.* **2005**, *4*, 801–808.

(28) Cobucci-Ponzano, B.; Conte, F.; Benelli, D.; Londei, P.; Flagiello, A.; Monti, M.; Pucci, P.; Rossi, M.; Moracci, M. The gene of an archaeal α-L-fucosidase is expressed by translational frameshifting. *Nucleic Acids Res.* **2006**, *34*, 4258–4268.

(29) Watanabe, Y.; Yokobori, S.; Inaba, T.; Yamagishi, A.; Oshima, T.; Kawarabayasi, Y.; Kikuchi, H.; Kita, K. Introns in protein-coding genes in Archaea. *FEBS Lett.* **2002**, *510*, 27–30.

(30) Kyrpides, N. C.; Woese, C. R. Universally conserved translation initiation factors. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 224–228.

(31) Chong, P. K.; Wright, P. C. Identification and characterization of the *Sulfolobus solfataricus* P2 proteome. *J. Proteome Res.* **2005**, *4*, 1789–1798.

(32) Hecker, A.; Leulliot, N.; Gadelle, D.; Graille, M.; Justome, A.; Dorlet, P.; Brochier, C.; Quevillon-Cheruel, S.; Le Cam, E.; van Tilbeurgh, H.; Forterre, P. An archaeal orthologue of the universal protein Kae1 is an iron metalloprotein which exhibits atypical DNA-binding properties and apurinic-endonuclease activity in vitro. *Nucleic Acids Res.* **2007**, *35*, 6042–6051.

(33) Koonin, E. V. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* **2003**, *1*, 127–136.

(34) Marcotte, E. M.; Pellegrini, M.; Ng, H. L.; Rice, D. W.; Yeates, T. O.; Eisenberg, D. Detecting protein function and protein-protein interactions from genome sequences. *Science* **1999**, *285*, 751–753.

(35) Bhatia, U.; Robison, K.; Gilbert, W. Dealing with database explosion: a cautionary note. *Science* **1997**, *276*, 1724–1725.

(36) Baranov, P. V.; Gesteland, R. F.; Atkins, J. F. Recoding: translational bifurcations in gene expression. *Gene* **2002**, *286*, 187–201.

(37) Namy, O.; Moran, S. J.; Stuart, D. I.; Gilbert, R. J.; Brierley, I. A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. *Nature* **2006**, *441*, 244–247.

(38) Sekine, Y.; Eisaki, N.; Ohtsubo, E. Translational control in production of transposase and in transposition of insertion sequence IS3. *J. Mol. Biol.* **1994**, *235*, 1406–1420.

(39) Licznar, P.; Mejlhede, N.; Prere, M. F.; Wills, N.; Gesteland, R. F.; Atkins, J. F.; Fayet, O. Programmed translational−1 frameshifting on hexanucleotide motifs and the wobble properties of tRNAs. *EMBO J.* **2003**, *22*, 4770–4778.

(40) Baranov, P. V.; Hammer, A. W.; Zhou, J.; Gesteland, R. F.; Atkins, J. F. Transcriptional slippage in bacteria: distribution in sequenced genomes and utilization in IS element gene expression. *Genome Biol.* **2005**, *6*, R25.

(41) Cui, Y.; Dinman, J. D.; Kinzy, T. G.; Peltz, S. W. The Mof2/Sui1 protein is a general monitor of translational accuracy. *Mol. Cell. Biol.* **1998**, *3*, 1506–1516.

(42) Waters, E.; Hohn, M. J.; Ahel, I.; Graham, D. E.; Adams, M. D.; Barnstead, M.; Beeson, K. Y.; Bibbs, L.; Bolanos, R.; Keller, M.; Kretz, K.; Lin, X.; Mathur, E.; Ni, J.; Podar, M.; Richardson, T.; Sutton, G. G.; Simon, M.; Soll, D.; Stetter, K. O.; Short, J. M.; Noordewier, M. The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12984–12988.

(43) Makarova, K. S.; Koonin, E. Evolutionary and functional genomics of the Archaea. *Curr. Opin. Microbiol.* **2005**, *8*, 586–594.

(44) Di Giulio, M. Formal proof that the split genes of tRNAs of Nanoarchaeum equitans are an ancestral character. *J. Mol. Evol.* **2009**, *69* (5), 505–511.

(45) Fitz-Gibbon, S. T.; Ladner, H.; Kim, U. J.; Stetter, K. O.; Simon, M. I.; Miller, J. H. Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 984–989.

(46) She, Q.; Singh, R. K.; Confalonieri, F.; Zivanovic, Y.; Allard, G.; Awayez, M. J.; Chan-Weiher, C. C.; Clausen, I. G.; Curtis, B. A.; De Moors, A.; Erauso, G.; Fletcher, C.; Gordon, P. M.; Heikamp-de Jong, I.; Jeffries, A. C.; Kozera, C. J.; Medina, N.; Peng, X.; Thi-Ngoc, H. P.; Redder, P.; Schenk, M. E.; Theriault, C.; Tolstrup, N.; Charlebois, R. L.; Doolittle, W. F.; Duguet, M.; Gaasterland, T.; Garrett, R. A.; Ragan, M. A.; Sensen, C. W.; Van der Oost, J. The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 7835–7840.

(47) Brugger, K.; Torarinsson, E.; Redder, P.; Chen, L.; Garrett, R. A. Shuffling of *Sulfolobus* genomes by autonomous and non-autonomous mobile elements. *Biochem. Soc. Trans.* **2004**, *32*, 179–183.

PR901166Q