Transcriptional linkage of *Haloferax volcanii* proteasomal genes with non-proteasomal gene neighbours including RNase P, MOSC domain and SAM-methyltransferase homologues

Malgorzata A. Gil, Katherine E. Sherwood and Julie A. Maupin-Furlow

Correspondence
Julie A. Maupin-Furlow
jmaupin@ufl.edu

Department of Microbiology and Cell Science, University of Florida, Gainesville, FL 32611-0700, USA

Comparative genomics reveals a common theme of 20S proteasome and proteasome-activating nucleotidase genes dispersed throughout archaeal genomes yet arranged in conserved linkages with gene homologues of translation and/or transcription machineries. To provide biological evidence for these linkages as well as insight into proteasome operon organization, transcripts of the five proteasomal genes of the halophilic archaeon *Haloferax volcanii* were analysed by Northern (RNA) blotting, RT-PCR and primer extension. These included *psmA*, *psmB* and *psmC*, encoding the 20S proteasomal subunits $\alpha 1$, β and $\alpha 2$, as well as *panA* and *panB*, encoding the PanA and PanB proteasome-activating nucleotidase proteins, respectively. All five of these genes are dispersed throughout the *H. volcanii* genome. For each proteasomal gene, a distinct transcript was detected by Northern blotting that was similar in size to the respective coding region. For both *psmA* and *psmC*, an additional transcript was detected that was 1.34 and 0.85 kb greater, respectively, than the coding region. Further analysis by Northern blotting and RT-PCR revealed that *psmA* was co-transcribed with genes encoding a Pop5 homologue of the RNase P endoRNase as well as an S-adenosylmethionine (SAM)-dependent methyltransferase. Likewise, *psmC* was co-transcribed with a downstream gene encoding a molybdenum cofactor sulfurase C-terminal (MOSC) domain protein. Additional proteasomal and neighbouring gene-specific transcriptional linkages were detected by RT-PCR. These results provide the first evidence that proteasome and tRNA modification genes are co-transcribed, reveal that a number of additional enzymes including those predicted to facilitate metal–sulfur cluster assembly are co-regulated with proteasomes at the transcriptional level, and provide further insight into proteasome gene transcription in archaea.

Received 19 March 2007

Revised 24 May 2007

Accepted 29 May 2007

INTRODUCTION

Proteasomes are energy-dependent proteases that are highly conserved and universally distributed within the *Archaea* and *Eucarya* domains (Maupin-Furlow *et al.*, 2006). These enzymes form nanocompartments within the cell that selectively degrade specific proteins into oligopeptides that are hydrolysed to free amino acids by downstream peptidases. The 20S core, responsible for the proteolytic activity of proteasomes, is a cylindrical bundle of four stacked heptameric rings composed of related α - and β -type subunits in an $\alpha_7:\beta_7:\beta_7:\alpha_7$ configuration. Gated openings on each end of the core (formed by α -type

subunits) limit substrate access to the central proteolytic chamber (formed by β -type subunits). Members of the ATPases associated with various cellular activities (AAA) family, including the related regulatory particle ATPases (Rpts) of eucaryal 26S proteasomes and proteasome-activating nucleotidases (PANs) of archaea, associate with 20S proteasomes and stimulate the energy-dependent degradation of proteins (Smith *et al.*, 2006).

The halophilic archaeon *Haloferax volcanii* synthesizes at least five proteasomal proteins, including two PAN proteins (PanA and PanB) and three 20S proteasomal subunits ($\alpha 1$, β and $\alpha 2$) (Wilson *et al.*, 1999; Reuter *et al.*, 2004). The latter form at least two active 20S proteasome subtypes ($\alpha 1\beta$ and $\alpha 1\alpha 2\beta$) (Kaczowka & Maupin-Furlow, 2003). As is common among archaea, the genes encoding these various proteasomal proteins are dispersed throughout the genome. Although scattered, comparative genomics reveals that archaeal proteasome genes reside in what

Abbreviations: AAA family, ATPases associated with various cellular activities family; MOSC domain, molybdenum cofactor sulfurase C-terminal domain; PAN, proteasome-activating nucleotidase; SAM, S-adenosylmethionine.

Supplementary data are available with the online version of this paper.

appear to be evolutionarily conserved superoperons that include non-proteasomal genes (Maupin-Furlow *et al.*, 2000; Koonin *et al.*, 2001).

Little is known regarding the transcripts generated from archaeal proteasome genes and the organization of these operons. *In silico* analyses of archaeal genome sequences (Maupin-Furlow *et al.*, 2000; Koonin *et al.*, 2001) reveal a high conservation of gene order between proteasomal genes and their neighbours, suggesting that these genes are linked at the transcriptional level. Northern blotting and primer-extension analysis of the 20S proteasomal genes (*psmA* and *psmB* encoding α and β subunits) of the methanogenic archaeon *Methanosarcina thermophila* reveal single gene transcripts (Maupin-Furlow & Ferry, 1995). However, a TATA-like promoter element is only detected upstream of the *M. thermophila psmB* (vs *psmA*), thus suggesting that the *psmA*-specific transcript is modified by post-transcriptional RNA cleavages in this methanogen.

Regarding *H. volcanii*, our work has shown that transcripts specific for the genes encoding the five proteasomal proteins (PanA, PanB, and 20S proteasome $\alpha 1$, β and $\alpha 2$) are present in exponential phase, and the abundance of these transcripts increases in parallel as cells enter stationary phase (Reuter *et al.*, 2004). Although the levels of PanA, $\alpha 1$ and β proteins remain relatively unchanged during this transition, the levels of PanB and $\alpha 2$ proteins increase severalfold, similarly to their encoding transcripts (Reuter *et al.*, 2004). Thus, the levels of proteasome proteins appear to be regulated by transcriptional and post-transcriptional mechanisms.

To further understand the transcription of the *H. volcanii* proteasomal genes and the relationship of this to proteasome regulation and operon organization, proteasome-specific mRNAs were analysed by Northern blotting, RT-PCR and primer extension. This analysis is believed to provide the first evidence that 'non-proteasomal' genes [including RNase P, molybdenum cofactor sulfuryase C-terminal (MOSC) domain and S-adenosylmethionine (SAM)-dependent methyltransferase homologues] are co-transcribed with genes encoding known proteasome proteins. This study also identified the 5' ends of proteasomal and neighbouring gene-specific transcripts, the results of which suggest that transcripts from the proteasomal regions of the *H. volcanii* genome undergo 5'-end post-transcriptional RNA cleavages.

METHODS

Materials. Biochemicals were purchased from Sigma-Aldrich and Bio-Rad Laboratories. Other organic and inorganic analytical grade chemicals were obtained from Fisher Scientific. Restriction endonucleases and T4 DNA ligase were purchased from New England Biolabs. Desalted oligonucleotides were obtained from Integrated DNA Technologies.

Strains, media, plasmids and nucleic acid electrophoresis. Bacterial and archaeal strains, oligonucleotide primers, template DNA

and plasmids are summarized in Supplementary Tables S1–S4. *Escherichia coli* DH5 α (New England BioLabs) was grown in Luria–Bertani (LB) medium (37 °C, 200 r.p.m.). *H. volcanii* DS2 (Mullakhanbhai & Larsen, 1975) was grown in ATCC 974 complex medium (42 °C, 200 r.p.m.). Media were supplemented with 100 mg ampicillin l⁻¹ or 50 mg kanamycin l⁻¹, as needed. Plasmids were purified using the Qiagen Prep Spin kit. DNA and RNA were separated by agarose gel electrophoresis (35 min, 5.6 V cm⁻¹, 20 °C) using 0.8 and 2% (w/v) agarose gels in 1 × Tris/acetate/EDTA buffer, and detected by ethidium bromide according to standard procedures (Ausubel *et al.*, 1987).

RNA purification. Total RNA was isolated from *H. volcanii* DS2 (exponential phase; OD₆₀₀ 0.7) according to Nieuwlandt *et al.* (1995). RNA was treated with amplification grade DNase I according to the supplier's recommendations (Sigma-Aldrich), with the following modification: 3 U enzyme was added per microgram RNA and the mixture was incubated for 15 min at 37 °C. The integrity of RNA was determined by agarose gel electrophoresis. RNA concentration was determined by A₂₆₀ using a Bio-Rad SmartSpec 3000 instrument.

Northern blot analysis. Total RNA (12 µg per lane) was denatured and fractionated by electrophoresis (14 h, 20 V) using formaldehyde/0.8% agarose gels in 1 × MOPS buffer (20 mM MOPS, pH 7.0, 5 mM sodium acetate, 1 mM EDTA) using standard procedures (Ausubel *et al.*, 1987). RNA molecular mass standards (0.24–9.5 kb RNA Ladder, Invitrogen Life Technologies) were included in the gels. RNA was transferred from gels to BrightStar-Plus membranes (Ambion) by downward capillary action for 3 h using 10 mM NaOH and 5 × saline sodium citrate (SSC) (where 20 × SSC is 3 M NaCl, 0.3 M sodium citrate, pH 7.0). RNA was cross-linked to membranes using a UV Stratalinker 2400 (Stratagene) and was hybridized to gene-specific probes or stained with 0.03% (w/v) methylene blue in 0.3 M sodium acetate, pH 5.2. The probes were complementary nucleotide fragments of target gene-coding regions generated by linear amplification with digoxigenin-11-dUTP (Roche) and *Taq* DNA polymerase (New England Biolabs), according to the recommendations of the suppliers. The templates and primers used for amplification are summarized in Supplementary Table S2. For hybridization, membranes with cross-linked RNA samples were equilibrated in high SDS buffer (5 × SSC, 2%, w/v, blocking reagent, 0.1%, w/v, *N*-lauroylsarcosine, 0.2%, w/v, SDS, 50%, w/v, formamide) (2 h, 55 °C), followed by incubation with 25 ng labelled probe per millilitre of high SDS buffer (16 h, 55 °C). Membranes were washed with 2 × SSC supplemented with 0.1% (w/v) SDS (two times 5 min, 25 °C) and 0.5 × SSC supplemented with 0.1% SDS (w/v) (two times 15 min, 55 °C). Hybridization products were detected by colorimetric and chemiluminescent (CSPD*) digoxigenin immunoassay according to the supplier's recommendations (Roche).

Primer-extension analysis. The 5' ends of proteasomal operon-specific transcripts were mapped by primer extension. Total RNA (20 µg) was suspended in 5.5 µl nucleotide mix (7.5 µM dGTP, 7.5 µM dTTP and 7.5 µM dCTP), and 1 µl of 2 µM primer (listed in Supplementary Table S3) was added. Sample that was boiled (2 min), cooled (room temperature, 5 min) and equilibrated (42 °C, 1 min) was mixed with 2 µl 5 × avian myeloblastosis virus reverse transcriptase (AMV RT) buffer (Promega) and incubated at 40 °C (10 min). AMV RT (11.5 U) and 1 µl α [³²P]dATP (3000 Ci mmol⁻¹; 111 TBq mmol⁻¹; Perkin Elmer) were added to the RNA, and the sample was incubated at 30 min for 40 °C. Freshly made quench solution (3 µl) (7.5 mM each of dATP, dCTP, dGTP and dTTP) was added and incubated for 30 min at 40 °C to complete the replication. Stop mix (13 µl) was then added, which was composed of 95% (v/v) formamide, 20 mM EDTA, pH 8.0, 0.05% (w/v) bromophenol blue and 0.05% (w/v) xylene cyanol FF. For standards, DNA sequencing reactions using the same primers as those in the primer-extension

reactions were performed with the plasmid DNA templates listed in Supplementary Table S3 and a Sequenase 7-deaza-dGTP Sequencing kit according to the supplier's recommendations (US Biochemicals). DNA sequencing and primer-extension reactions were separated by electrophoresis (3 h at 1800 V, 50 °C) using a high-resolution denaturing 6% (w/v) polyacrylamide gel (Bio-Rad). Reaction products were visualized by autoradiography after exposure of Kodak RX Blue film to the dried polyacrylamide gels for 16–48 h at 20 °C.

RT-PCR. RT-PCR was performed using *H. volcanii* total RNA (0.5 µg) as template, appropriate primers (Supplementary Table S3), OneStep RT-PCR reaction mix (Qiagen) and an iCycler (Bio-Rad). OneStep RT-PCR included Q-solution, Omniscript and Sensiscript reverse transcriptases, and HotStart *Taq* DNA polymerase. After cDNA synthesis (50 °C, 30 min), reactions were preheated to 95 °C (15 min), followed by 30 amplification cycles consisting of denaturation (94 °C, 30 s), annealing (30 s, temperatures listed in Supplementary Table S3) and elongation (72 °C, 1 min). Final extension was performed at 72 °C (10 min). For each primer pair, negative and positive controls were included to exclude genomic DNA contamination and confirm primer pair function, respectively. Controls were identical to the RT-PCR reactions, with the following exceptions: sample was maintained on ice during the reverse-transcription step for the negative control, and *H. volcanii* genomic DNA (10–25 ng) (Ng *et al.*, 1995) was used as a template for the positive control.

RNA fold prediction. Secondary structure and ΔG values for RNA were estimated using mFold (Integrated DNA Technologies) using 50 maximum foldings, linear RNA, and 50% suboptimality. Invariable parameters were 25 °C and 1 M sodium concentration.

RESULTS

The five proteasomal genes of *H. volcanii* [*psmA* ($\alpha 1$), *psmB* (β), *psmC* ($\alpha 2$), *panA* (PanA) and *panB* (PanB)] are dispersed throughout its 2.84 Mb chromosome (Wilson *et al.*, 1999; Reuter *et al.*, 2004). In this study, transcripts specific for each proteasomal gene were analysed by Northern blotting, primer extension and RT-PCR, using total RNA prepared from exponential-phase cells (see Supplementary Tables S1–S4 and Methods for details). Comparative genomics was also performed to understand how these transcripts correlate with gene neighbourhoods conserved within haloarchaea (i.e. *H. volcanii*, *Haloquadratum walsbyi*, *Halobacterium* sp. NRC-1, *Haloarcula marismortui* and *Natronomonas pharaonis*).

psmA ($\alpha 1$) region

Northern blot analysis using a cDNA probe specific for *psmA* revealed two distinct transcripts of 1.0 and 2.1 kb, with the smaller transcript more abundant than the larger transcript (Fig. 1). Both transcripts were greater in size than the 0.76 kb coding capacity of *psmA*, thus suggesting that they may also be specific for gene neighbours. To investigate this, the genome sequence of *H. volcanii* (05/26/06 assembly; <http://archaea.ucsc.edu/>) was used to estimate the size of genes adjacent to *psmA*. The 234 bp region immediately downstream of *psmA* contained highly repetitive DNA sequence elements, lacked apparent coding

Fig. 1. Northern (RNA) blot analysis of transcripts generated from proteasome and proteasome-activating nucleotidase regions of *H. volcanii*. Lanes 1–7: total RNA was probed with cDNA fragments of the coding regions of *psmA* (I), *panA* (VII), *psmC* (VI), *panB* (VIII), *psmB* (V), *sam* (III) and *rnpB* (II), respectively. Probe numbers are indicated by Roman numerals and summarized in Figs 2, 4, 5 and 6, and Supplementary Table S2. RNA molecular mass markers are indicated (left of figure).

capacity for protein, and preceded the termination codon of *cinR* (a gene encoded on the DNA strand complementary to *psmA*) (Fig. 2). The region upstream of *psmA* included *rnpA*, *sam* and *rnpB* genes in the same orientation as *psmA*, and respectively encoding RNase P Rpp30, SAM-dependent methyltransferase and RNase P Pop5 protein homologues. This *rnpA-sam-rnpB-psmA* region spanned a total of 2559 bp, with the coding sequences of *rnpA* and *sam* overlapping by 17 bp and the intergenic regions of *sam-rnpB* and *rnpB-psmA* at 9 and 6 bp, respectively. The 17 bp overlap of *rnpA* and *sam* suggested that these two genes were co-transcribed; however, neither of the *psmA*-specific transcripts detected by Northern blotting was large enough to accommodate *rnpA* to *psmA*. Instead, the larger (2.1 kb) transcript was similar in size to the 1.9 kb *sam-rnpB-psmA* coding sequence plus the 0.23 kb downstream region. The 1.0 kb size of the smaller transcript, in contrast, was not large enough to accommodate the 1.2 kb required to encode the complete upstream *rnpB* gene in addition to *psmA*. This smaller transcript was similar in size to *psmA* (0.76 kb) plus the 0.23 kb downstream region.

To further analyse the transcripts generated in the *psmA* region, probes specific for *rnpA*, *sam* and *rnpB* were hybridized to total RNA by Northern blotting (Fig. 1). A single transcript of 2.1 kb was detected with the *sam*- and *rnpB*-specific probes that was comparable in size to the larger transcript observed with the *psmA*-specific probe (Fig. 1). No signal was detected with the *rnpA*-specific probe (data not shown). Although the results do not clarify whether *rnpA* is co-transcribed with its gene neighbours,

they do reveal that *sam-rnpB-psmA* are co-transcribed on a 2.1 kb transcript. Interestingly, a high degree of conservation was observed among haloarchaea and other archaea in which *psmA* genes were in the same orientation and linked with upstream *rnpA* and/or *rnpB* genes (Fig. 2) (Supplementary Fig. S1 for details). Similar to *H. volcanii*, SAM-dependent methyltransferase genes were also arranged between *rnpA* and *rnpB* genes in the *Halobacterium* sp. NRC-1 and *Haloquadratum walsbyi* genomes. Thus, the transcriptional linkage of *psmA*, *rnpA*, *rnpB* and *sam* genes is likely to be more widespread than in *H. volcanii* alone.

In addition to Northern blotting, RT-PCR was performed to analyse the transcriptional continuity of the genes within the *psmA* region (Fig. 2). Appropriate controls were included to confirm that each set of primer pairs was functional and to rule out the possibility that the RT-PCR product was due to genomic DNA contamination of the RNA samples (see Methods). RT-PCR products were analysed by agarose gel electrophoresis and DNA sequencing. Adjacent genes were characterized as either co-transcribed or not co-transcribed, based on the respective presence or absence of a single RT-PCR product homologous in size and sequence to that expected based on primer pair design. As for Northern blotting, transcriptional continuity between *sam-rnpB-psmA* was detected by RT-PCR (Fig. 2). Pairwise linkages were also detected for *sam-rnpA*, *rnpA*-ORF00853 and ORF00853-ORF00852, where ORF00852 and ORF00853 encode a putative M20/M25/M40-type peptidase/hydrolase and a conserved

membrane protein, respectively. Together, these six ORFs (ORF00852 to *psmA*) span a total of 4.4 kb of coding sequence. It is unclear whether a large transcript is synthesized within the *psmA* region that spans all six of these ORFs or whether multiple promoters are used to generate smaller transcripts which harbour these gene pairs. If large mRNA species are generated, either they are not detected by the Northern blot methods of this study and/or they are cleaved post-translationally by RNases to generate smaller transcripts which harbour the coding capacity for a subset of these gene pairs. In addition to the linkages described above, ORF00850 and ORF00851 (*lysA* and *dapF* homologues) were co-transcribed but not linked to the ORF00852 to *psmA* region based on RT-PCR analysis.

The 5' ends of selected transcripts within the *psmA* region were mapped by primer extension. Distinct products were identified 142 and 141 bp upstream of the *rnpA* and *psmA* start codons, respectively (Fig. 2). The 5' ends of these transcripts began with either a C or a G, and thus were consistent with previously characterized haloarchaeal transcripts. In contrast, the DNA sequences 19–32 bases upstream of these transcript start sites (Figs 2 and 3) were GC rich (64% G+C) and not related to the haloarchaeal TATA-like promoter consensus sequence described by Danner & Soppa (1996) (RGTWWWWRACYGSY, where R=A or G, Y=C or T, W=A or T, S=G or C). It is interesting to note that highly identical sequences were identified which were 38–20 bases upstream of the

Fig. 2. Comparative genomic and transcript analyses of the *psmA* (20S proteasomal subunit $\alpha 1$) region of *H. volcanii*. (a) Summary of comparative genomic and transcript analyses. *H. volcanii* ORFs are indicated by arrows underscored with numbers according to 05/26/06 assembly (<http://archaea.ucsc.edu/>). The proteasomal gene of interest is highlighted in black. Non-proteasomal ORFs conserved in organization between *H. volcanii* and other genomes are in grey (see supplementary figures for details). RT-PCR primers are indicated by arrows underscored by the primer number according to Supplementary Table S3, with the presence (–) and absence (×) of RT-PCR products indicated. Northern blot probes are indicated by boxes with Roman numerals. Transcripts detected by Northern blotting (→ with kb) and 5' ends of transcripts mapped by primer extension (P1 and/or P2) are also indicated. Sequences with similarity to the TATA-like (Box A) promoter sequence described by Danner & Sopper (1996) are indicated by and/or above ORFs. (b) RT-PCR analysis. RT-PCR products separated by DNA gel electrophoresis are depicted with molecular mass standards (M_r) on the left. RT-PCR analysis of genomic DNA (G, positive control), total RNA (R) and total RNA without the reverse-transcription step (N, negative control) are included. The arrowhead to the right indicates RT-PCR. Product size (bp) is based on genome sequence. u.d., Undetected: indicates RT-PCR did not generate detectable cDNA with total RNA as template. Primers are indicated below the gel and numbered according to Supplementary Table S3. (c) Primer-extension analysis. Left: primer-extension products generated from proteasomal region-specific transcripts (lanes 1 and 2, respectively, correlate with 2 and 4 μ l of quenched primer-extension reaction; see Methods for details) and DNA sequencing reactions (lanes A, T, G and C) using the primers listed in Supplementary Table S4. The arrowhead to the right of the gel indicates the primer-extension product, with the gene target and position of the 5' end of the transcript relative to the putative translational start site of the gene indicated in parentheses. Right: DNA sequences of regions analysed by primer extension. Indicated are: bases complementary to the 5' ends of primer-extension products ($\nabla + 1$); regions 19–32 bp upstream of the transcript start site in boxes; ORF coding sequence with start and stop codons in bold type, putative Shine–Dalgarno (SD) sites in boxes; Box A promoter sequences in boxes with bases identical to the TATA consensus in bold italic type; stem–loop structures ($\rightarrow \leftarrow$); and repetitive DNA sequences in bold italic type (not boxed).

Fig. 3. Multiple sequence alignment of regions upstream of the transcript start sites of the proteasomal genes. Bases complementary to the 5' ends of the proteasomal gene- and proteasomal gene neighbour-specific transcripts (▼+1) and the regions upstream of these sites are included. Identical bases are highlighted in white type on a black background.

psmA- and *psmB* (β)-specific transcript start sites (Fig. 3) (CGSKGMCRRYGCYGYCTG, where S=C or G, K=T or G, M=C or A, R=G or A, Y=C or T) (see discussion of *psmB* region, below). In addition, TATA-like (BoxA) sequences were identified 310 and 139 bp upstream of the start codon of *rnpA* and *psmA*, respectively (Fig. 2). While the Box A sequence upstream of *psmA* (b) is within the *rnpB* coding sequence, it overlaps (and thus cannot account for) the *psmA*-specific transcript start site. The second Box A sequence (a), in contrast, is positioned 168 bp upstream of the *rnpA*-specific transcript start site, thus opening the possibility that transcription initiates from this promoter-like element and that the resulting

transcripts are cleaved at the 5' end to generate *rnpA*- and/or *psmA*-specific transcripts.

psmC ($\alpha 2$) region

Northern blot analysis using a cDNA probe specific for *psmC* revealed two transcripts of 0.8 and 1.6 kb that were relatively similar in abundance (Fig. 1). Of these, the smaller transcript was comparable in size to the 0.75 kb coding capacity of *psmC*, and the larger was consistent with the 1637 bp distance spanning the start codon of *psmC* to the stop codon of *mosC* (Fig. 4). In further support of the *psmC*-*mosC* transcriptional linkage, the stop codon of

Fig. 4. Comparative genomic and transcript analyses of the *psmC* (20S proteasomal subunit $\alpha 2$) region of *H. volcanii*. See the legend of Fig. 2 for details. Lanes 1, 2, 3 and 4 of Fig. 4(c) correlate with 2, 4, 6 and 8 μ l, respectively, of quenched primer-extension reaction; see Methods for details.

psmC was found to immediately precede the start codon of *mosC*. In contrast, ORF02058-*psmC* and ORF02059-ORF02058-*psmC* span 1088 and 2084 bp, with intergenic regions of 75 and 159 bp upstream of ORF02058 and *psmC*, respectively. RT-PCR analysis of the *psmC* region (Fig. 4) revealed transcriptional continuity between *psmC* and *mosC*, consistent with these Northern blot results. In addition, RT-PCR analysis demonstrated that the *psmC*-*mosC* gene pair was also linked at the transcriptional level to additional upstream and downstream ORFs (02059, 02058, 02055 and 02054) that encode conserved or hypothetical proteins (Fig. 4). The transcriptional linkage of this ORF02059 to ORF02054 region, which spans 4235 bp, was based on the RT-PCR products generated between ORF02059 and ORF02058, ORF02058 and *psmC*, *psmC* and *mosC*, *mosC* and ORF02055, and ORF02055 and ORF02054. The linkage of the last two gene pairs, however, is based on relatively weak RT-PCR products, suggesting that the levels of these transcripts are limited under the conditions examined in this study.

The 5' ends of *psmC*- and ORF02058-specific transcripts were mapped by primer extension and found to be within intergenic regions 113 and 64 bp upstream, respectively, of each gene (Fig. 4). Both transcripts began with a G and had relatively GC-rich sequences 19–32 bases upstream of their start sites. A Box A sequence (a) preceded by repeats of CTCGTGA was identified 426 bp upstream of the ORF02058-specific transcript start site. Likewise, a Box A sequence (b) was identified 251 bp upstream of the *psmC*-specific transcript start site. The location of both of these Box A elements supports the suggestion that ORF02058- and *psmC*-specific transcripts undergo processing after expression from Box A (a) and (a or b), respectively. Although the primary (unprocessed) transcript generated from Box A (b) would include the 4 bases upstream of ORF02058, it would require ORF02058 to be translated using an internal ribosome-binding site. Likewise, the third Box A-like sequence 29 bp upstream of *psmC* (c) would generate a leaderless transcript devoid of the *psmC*-Shine-Dalgarno site. The examples of this type of translation

initiation occurring in archaea, however, are limited (Srinivasan *et al.*, 2006).

Comparative genomics of the *psmC* region was performed among archaea which encode a second α -type 20S proteasomal subunit (*H. volcanii*, *Haloquadratum walsbyi* and *Haloarcula marismortui*) (see Supplementary Fig. S2 for details). Of these, *Haloarcula marismortui* was unique, with a coding capacity for both $\alpha 2$ and $\beta 2$ in a region that to date displays little or no gene conservation with other organisms. In contrast, the *psmC* regions of *H. volcanii* and *Haloquadratum walsbyi* were somewhat related. Although the genes immediately 5' and 3' of the *H. volcanii psmC* gene are not conserved, the gene neighbours further upstream are conserved with *Haloquadratum walsbyi*, including *thyA* (thymidylate synthase) and *folA* (dihydrofolate reductase).

psmB (β) region

Northern blot analysis with a *psmB*-specific cDNA probe detected a single 0.75 kb transcript which correlated in size to the 0.73 kb *psmB* coding region (Figs 1 and 5). Primer-extension analysis mapped the 5' end of the *psmB*-specific transcript to a G residue within the intergenic region and 49 bp upstream of the translation start site of *psmB* (Fig. 5). Although a Box A consensus was not identified 32–19 bases upstream of this transcript start site, the –38 to –20 region was 57% identical to a sequence in a similar location upstream of the *psmA*-specific transcript start site (Figs 3 and 5). Based on the homology of these two regions, this GC-rich sequence may serve as an alternative promoter that coordinates the stoichiometric ratios of $\alpha 1$ and β proteins in 20S proteasome subtypes which are predominant during exponential-phase growth (Kaczowka & Maupin-Furlow, 2003). However, a number of Box A-like sequences (b, c, e) were also detected 305–446 bp upstream of the *psmB*-transcript start site (Fig. 5), thus opening the alternative possibility that a larger transcript is synthesized from one or more of these Box A promoters and cleaved at the 5' end to generate the *psmB*-specific transcript. A Box A consensus sequence located only 19 bp upstream of the *psmB* start codon was also identified (f) (Fig. 5); however, a leaderless *psmB*-specific transcript would be synthesized that would require internal translation initiation.

Interestingly, a transcriptional linkage of *psmB* to the upstream ORF00376, encoding a conserved transmembrane (TM)-spanning protein, was detected by RT-PCR (Fig. 5) that was not identified by Northern blotting. The orientation of this ORF in relation to *psmB* was highly conserved in the genomes of *H. volcanii*, *Haloquadratum walsbyi* and *Haloarcula marismortui* (Fig. 5) (see Supplementary Fig. S3 for details). In addition, two ORFs were conserved and typically in a divergent orientation with respect to the *psmB* genes of all haloarchaeal genomes to date. These ORFs included a conserved helix-turn-helix transcriptional regulator with a cystathionine β -synthase (CBS) domain proposed to bind the adenosyl portion of

molecules such as SAM, ATP and AMP (Ignoul & Eggermont, 2005), as well as a DUF555 family member with a predicted central TM-spanning helix (Fig. 5; ORF00378 and ORF00377 for *H. volcanii*, respectively) (see Supplementary Fig. S3 for details).

panA region

A single *panA*-specific transcript of 1.4 kb, similar in size to the 1.22 kb *panA*-coding region, was detected by Northern blot analysis (Figs 1 and 6). In addition, a distinct primer-extension product was identified at a C residue located 181 bp upstream of the translation start site of *panA* (Fig. 6). Although the region 19–32 bases upstream of the transcript start site was relatively GC-rich, two sequences with similarity to the Box A consensus were identified within the *arsR* coding sequence 81–117 bp upstream of the transcript start site (a, b), suggesting that the *panA*-specific transcript is cleaved at the 5' end. Interestingly, two Box A-like sequences were also identified downstream of the primer-extension site that were within 110–129 bp of the *panA* start codon (d, f) and which overlapped complementary strand Box A-like sequences (c, e) predicted to mediate *arsR* transcription. While these do not account for the primer-extension product detected in this study, the consensus of these four Box A-like sequences is high, ranging from 64 to 86% identity. Whether these promoter elements are negatively regulated during exponential-phase growth remains to be determined.

Interestingly, a transcriptional linkage of *panA* to the downstream *pdg* gene encoding a homologue of pyrimidine dimer glycosylase (UV endonuclease) was identified by RT-PCR (Fig. 6) that was not detected by Northern blotting. Comparison of haloarchaeal gene neighbourhoods revealed that this *panA*-*pdg* linkage was conserved in *Haloquadratum walsbyi* (Fig. 6) (see Supplementary Fig. S4 for details). In addition, the majority of haloarchaeal *panA* genes (including that of *H. volcanii*) were divergently orientated with *arsR*-, *mre11*- and *rad50*-like genes (respectively encoding homologues of ArsR transcriptional regulator, Mre11 DNA double-strand break repair, and Rad50 chromosome segregation proteins) (Fig. 6) (see Supplementary Fig. S4 for details). Thus, haloarchaeal *panA* genes may be transcriptionally regulated with *pdg* as well as other genes that encode DNA binding/modifying enzymes.

panB region

A single transcript of 1.35 kb was detected by Northern blotting with a *panB*-specific probe, which correlated in size to the 1.24 kb coding region of *panB* (Figs 1 and 7). This single gene transcript was also detected by RT-PCR, and its 5' end was mapped by primer extension (Fig. 7). Based on the latter analysis, a single distinct product was identified which mapped to a C residue within the intergenic region 221 bp upstream of the *panB* translation start site (Fig. 7). Unlike the other proteasomal genes, the region 19–32 bases

Fig. 5. Comparative genomic and transcript analyses of the *psmB* (20S proteasomal subunit β) region of *H. volcanii*. See the legend of Fig. 2 for details.

Fig. 6. Comparative genomic and transcript analyses of the *panA* (proteasome-activating nucleotidase A) region of *H. volcanii*. See legends of Figs 2 and 4 for details. The asterisk in (a) denotes that PCR amplification of genomic DNA was limited using primer pair A1 and A2. Thus, it cannot be ruled out that the transcript which spans this region is not synthesized in the cell.

upstream of this transcript start site was related to a Box A consensus with 57 % identity. Two additional Box A-like sequences were identified with similar identity; however, both were downstream of the transcript start site.

Although *panB* appeared monocistronic based on Northern blotting and RT-PCR, comparative genomics revealed high conservation of its gene neighbourhood with that of *Haloquadratum walsbyi* (Fig. 7) (see Supplementary Fig. S5 for details). N-terminal protein acetyltransferase (*rimI*) and aconitate hydratase (*citB*) genes were divergently oriented from the *panB* genes of both of these haloarchaea. In addition, homologues of deaminase (*dcd*) and decarboxylase (*pdaD*) genes were in the same orientation and immediately surrounding these *panB* genes. A *pdaD* homologue was also detected downstream of the *Halobacterium* sp. NRC-1 *panB*. In *H. volcanii*, this *pdaD* gene homologue was linked to ORF03057 (a signal-transducing histidine kinase of COG0642) at the transcript level by RT-PCR (Fig. 7).

DISCUSSION

In this study, transcripts specific for the five known proteasome genes (*psmA*, *psmB*, *psmC*, *panA* and *panB*) were analysed from exponential-phase *H. volcanii* cells. Based on Northern blotting, the majority of these transcripts were comparable in size to the coding region of each respective proteasome gene. Transcripts specific for the coding regions of the proteasomal genes plus their gene neighbours were also identified (with the exception of *panB*). Most notably, a 2.1 kb transcript specific for *rnpB-sam-psmA* and a 1.6 kb transcript comparable in size to the *psmC-mosC* coding region were readily detected by Northern blotting. Co-transcripts which included the coding sequences of neighbouring genes within the proteasomal regions were also detected by RT-PCR. These included transcripts with the coding sequences of: ORF00852-ORF00853, *rnpA-sam*, *sam-rnpB* and *rnpB-psmA* within the *psmA* region; ORF02059-ORF02058, ORF02058-*psmC*, *psmC-mosC* and *mosC*-ORF02055 within the *psmC* region; and *psmB*-ORF00376 and *panA-pdg* within the *psmB* and *panA* regions, respectively. The reason for the enhanced number of transcriptional linkages detected by RT-PCR compared to Northern blotting remains to be determined. However, it is speculated that these larger co-transcripts are less abundant than the single gene transcripts, particularly if the co-transcripts are precursor intermediates in mRNA processing via 5'-end cleavage in this region. Northern blotting is notorious for its low sensitivity, risk of mRNA degradation during electrophoresis, and difficulty in transfer of large transcripts from gel to hybridization membrane (Dvorak *et al.*, 2003). In contrast, RT-PCR is more sensitive and requires less manipulation of RNA, thus reducing the risk of mRNA degradation prior to analysis.

In addition to discovering transcriptional linkages in the proteasome-encoding regions of the genome, the 5' ends of

proteasomal gene-specific transcripts were determined. With the exception of *panB*, relatively GC-rich sequences were located 19–32 bases upstream of the transcript start site that were not related to the Box A (TATA) consensus sequence common to many archaeal gene promoters (Bell & Jackson, 2001). It is possible that novel promoter elements drive transcription of these genes (most notably a conserved GC-rich sequence 20–38 bases upstream of the *psmA* and *psmB* transcript start sites). Haloarchaea are particularly complex in the number of transcription factor IIB (TFB) and TATA-binding protein (TBP) pairs which appear to recruit RNA polymerase to promoters, suggesting that more than one TATA-like consensus sequence is utilized in these organisms (Baliga *et al.*, 2000; Facciotti *et al.*, 2007). Alternatively, the transcripts mapped in this study may have arisen by posttranscriptional RNA cleavages at the 5' end. This latter possibility is supported by the finding of TATA-like (Box A) consensus sequences 81–426 bp upstream of the transcript start sites.

Although the neighbouring genes that were linked to proteasomal genes at the transcriptional level have yet to be characterized in *H. volcanii*, several of the deduced proteins are closely related to orthologous enzymes with known biochemical properties. These include *rnpA* and *rnpB*, which encode the Rpp30 and Pop5 protein subunits of RNaseP, an enzyme responsible for processing the 5' ends of precursors to tRNA as well as cleaving other stable RNAs, intergenic transcripts (Li & Altman, 2003) and riboswitches (Altman *et al.*, 2005). Although the RNA component of the archaeal RNase P alone is capable of catalysing 5'-end tRNA maturation, substrate specificity (k_{cat}/K_m) is dramatically enhanced in the presence of Pop5 and Rpp30 (Tsai *et al.*, 2006). In addition to RNase P proteins, the deduced protein of the *H. volcanii* *pdg* is closely related (*E* value 2×10^{-32}) to the pyrimidine dimer glycosylase (UV endonuclease) of *Micrococcus luteus* (Piersen *et al.*, 1995).

The function of remaining genes transcribed with the proteasomal genes and/or proteasomal gene neighbours is less clear, as no close homologues have been characterized. The 17 bp overlap of *sam* and *rnpA* suggests that both genes may be involved in tRNA modification. Consistent with this, the *sam*-deduced protein has the sequence motifs (I–III) common to protein, DNA, RNA and small-molecule SAM-dependent methyltransferases (Ingrosso *et al.*, 1989) (Pfam08241; *E* value 2.1×10^{-14}); however, the specificity of this group of enzymes is difficult to predict by bioinformatics alone. The *mosC*-encoded protein harbours an N-terminal β -barrel MOSC domain, including the highly conserved Cys predicted to accept sulfur (S^0) from cysteine desulfurase (NifS-like) enzymes and deliver the abstracted sulfur to apoprotein targets during the assembly of metal-sulfur clusters (Anantharaman & Aravind, 2002). ORF02054 and ORF00852 respectively encode proteins related to metal-dependent hydrolases (cd01299 subgroup A; *E* value 4×10^{-44}) and M20/M25/M40 peptidases (pfam01546; *E* value 1×10^{-28}). Thus, these enzymes may be involved in peptide bond cleavage. The remaining ORFs (00853, 00376,

Fig. 7. Comparative genomic and transcript analyses of the *panB* (proteasome-activating nucleotidase B) region of *H. volcanii*. See the legend of Fig. 2 for details.

02059, 02058 and 02055) are conserved only with proteins of unknown function, including a subset predicted to be membrane-associated.

The reason for the assortment of putative RNA-modifying, apoprotein-maturation, DNA-repair and hydrolytic genes co-transcribed with the proteasomal genes remains to be determined. However, co-regulation of RNase P with its neighbours is not unusual, having also been observed in bacteria, in which *rnpA* (encoding C5 of RNase P) is linked at the transcriptional level to *rmpH* (encoding the small ribosomal protein L34) (Ellis & Brown, 2003). The transcriptional connections of RNase P to the proteasomes of *H. volcanii* are likely to serve multiple functions, including coordinating the balance of free amino acids generated by protein degradation (mediated by the proteasome and peptidases) with the free amino acids needed to charge tRNA after maturation by enzymes such as RNase P. Linking the transcription of proteasomal genes to those involved in apoprotein maturation (e.g. *mosC*) may enhance overall protein quality control. AAA+ proteins such as those associated with proteasomes (i.e. the PAN proteins) are important in the disaggregation and unfolding of proteins (Benaroudj & Goldberg, 2000), and thus may assist a 'MosC-dependent' metalloenzyme assembly pathway. Proteasomes also appear to be closely tied to the regulation of DNA repair (Sweder & Madura, 2002; Wang *et al.*, 2005); the transcriptional linkage of *panA* to *pdg* may facilitate this type of regulation in *H. volcanii*.

ACKNOWLEDGEMENTS

This research was funded in part by a grant from the National Institutes of Health (R01 GM057498) and Department of Energy (DE-FG02-05ER15650). The authors would like to extend a special thanks to Ms Wei Li for her technical assistance.

REFERENCES

- Altman, S., Wesolowski, D., Guerrier-Takada, C. & Li, Y. (2005). RNase P cleaves transient structures in some riboswitches. *Proc Natl Acad Sci U S A* **102**, 11284–11289.
- Anantharaman, V. & Aravind, L. (2002). MOSC domains: ancient, predicted sulfur-carrier domains, present in diverse metal-sulfur cluster biosynthesis proteins including molybdenum cofactor sulfases. *FEMS Microbiol Lett* **207**, 55–61.
- Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Smith, J. A., Seidman, J. G. & Struhl, K. (1987). Preparation and analysis of RNA. In *Current Protocols in Molecular Biology*, pp. 4.0.1–4.10.11. New York: Green Publishing Associates and Wiley-Interscience.
- Baliga, N. S., Goo, Y. A., Ng, W. V., Hood, L., Daniels, C. J. & DasSarma, S. (2000). Is gene expression in *Halobacterium* NRC-1 regulated by multiple TBP and TFB transcription factors? *Mol Microbiol* **36**, 1184–1185.
- Bell, S. D. & Jackson, S. P. (2001). Mechanism and regulation of transcription in archaea. *Curr Opin Microbiol* **4**, 208–213.
- Benaroudj, N. & Goldberg, A. L. (2000). PAN, the proteasome-activating nucleotidase from archaeobacteria, is a protein-unfolding molecular chaperone. *Nat Cell Biol* **2**, 833–839.
- Danner, S. & Soppa, J. (1996). Characterization of the distal promoter element of halobacteria *in vivo* using saturation mutagenesis and selection. *Mol Microbiol* **19**, 1265–1276.
- Dvorak, Z., Pascucci, J. M. & Modriansky, M. (2003). Approaches to messenger RNA detection – comparison of methods. *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub* **147**, 131–135.
- Ellis, J. C. & Brown, J. W. (2003). Genes within genes within bacteria. *Trends Biochem Sci* **28**, 521–523.
- Facciotti, M. T., Reiss, D. J., Pan, M., Kaur, A., Vuthoori, M., Bonneau, R., Shannon, P., Srivastava, A., Donohoe, S. M. & other authors (2007). General transcription factor specified global gene regulation in archaea. *Proc Natl Acad Sci U S A* **104**, 4630–4635.
- Ignoul, S. & Eggermont, J. (2005). CBS domains: structure, function, and pathology in human proteins. *Am J Physiol Cell Physiol* **289**, C1369–C1378.
- Ingrasso, D., Fowler, A. V., Bleibaum, J. & Clarke, S. (1989). Sequence of the D-aspartyl/L-isoaspartyl protein methyltransferase from human erythrocytes. Common sequence motifs for protein, DNA, RNA, and small molecule S-adenosylmethionine-dependent methyltransferases. *J Biol Chem* **264**, 20131–20139.
- Kaczowka, S. J. & Maupin-Furlow, J. A. (2003). Subunit topology of two 20S proteasomes from *Haloferax volcanii*. *J Bacteriol* **185**, 165–174.
- Koonin, E. V., Wolf, Y. I. & Aravind, L. (2001). Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach. *Genome Res* **11**, 240–252.
- Li, Y. & Altman, S. (2003). A specific endoribonuclease, RNase P, affects gene expression of polycistronic operon mRNAs. *Proc Natl Acad Sci U S A* **100**, 13213–13218.
- Maupin-Furlow, J. A. & Ferry, J. G. (1995). A proteasome from the methanogenic archaeon *Methanosarcina thermophila*. *J Biol Chem* **270**, 28617–28622.
- Maupin-Furlow, J. A., Wilson, H. L., Kaczowka, S. J. & Ou, M. S. (2000). Proteasomes in the archaea: from structure to function. *Front Biosci* **5**, d837–d865.
- Maupin-Furlow, J. A., Humbard, M. A., Kirkland, P. A., Li, W., Reuter, C. J., Wright, A. J. & Zhou, G. (2006). Proteasomes from structure to function: perspectives from archaea. *Curr Top Dev Biol* **75**, 125–169.
- Mullakhanbhai, M. F. & Larsen, H. (1975). *Halobacterium volcanii* spec. nov., a Dead Sea *Halobacterium* with a moderate salt requirement. *Arch Microbiol* **104**, 207–214.
- Ng, W.-L., Yang, C.-F., Halladay, J. T., Arora, P. & DasSarma, S. (1995). Isolation of genomic and plasmid DNAs from *Halobacterium halobium*. In *Archaea: a Laboratory Manual*, pp. 179–184. New York: Cold Spring Harbor Laboratory Press.
- Nieuwlandt, D. T., Palmer, J. R., Armbruster, D. W., Kuo, Y. P., Oda, W. & Daniels, C. J. (1995). A rapid procedure for the isolation of RNA from *Haloferax volcanii*. In *Archaea: a Laboratory Manual*, pp. 161–162. Plainview, NY: Cold Spring Harbor Laboratory Press.
- Piersen, C. E., Prince, M. A., Augustine, M. L., Dodson, M. L. & Lloyd, R. S. (1995). Purification and cloning of *Micrococcus luteus* ultraviolet endonuclease, an N-glycosylase/abasic lyase that proceeds via an imino enzyme-DNA intermediate. *J Biol Chem* **270**, 23475–23484.
- Reuter, C. J., Kaczowka, S. J. & Maupin-Furlow, J. A. (2004). Differential regulation of the PanA and PanB proteasome-activating nucleotidase and 20S proteasomal proteins of the haloarchaeon *Haloferax volcanii*. *J Bacteriol* **186**, 7763–7772.
- Smith, D. M., Benaroudj, N. & Goldberg, A. (2006). Proteasomes and their associated ATPases: a destructive combination. *J Struct Biol* **156**, 72–83.

Srinivasan, G., Krebs, M.P. & RajBhandary, U.L. (2006). Translation initiation with GUC codon in the archaeon *Halobacterium salinarum*: implications for translation of leaderless mRNA and strict correlation between translation initiation and presence of mRNA. *Mol Microbiol* **59**, 1013–1024.

Sweder, K. & Madura, K. (2002). Regulation of repair by the 26S proteasome. *J Biomed Biotechnol* **2**, 94–105.

Tsai, H. Y., Pulukkunat, D. K., Woznick, W. K. & Gopalan, V. (2006). Functional reconstitution and characterization of *Pyrococcus furiosus* RNase P. *Proc Natl Acad Sci U S A* **103**, 16147–16152.

Wang, Q. E., Wani, M. A., Chen, J., Zhu, Q., Wani, G., El Mahdy, M. A. & Wani, A. A. (2005). Cellular ubiquitination and proteasomal functions positively modulate mammalian nucleotide excision repair. *Mol Carcinog* **42**, 53–64.

Wilson, H. L., Aldrich, H. C. & Maupin-Furlow, J. A. (1999). Halophilic 20S proteasomes of the archaeon *Haloferax volcanii*: purification, characterization and gene sequence analysis. *J Bacteriol* **181**, 5814–5824.

Edited by: J. van der Oost