PhD Thesis
Yang Guo

Functional characterization and Gene regulation of the archaeal virus SIRV2

Supervisor: Xu Peng
Submitted: 05/11/14
Preface

The thesis entitled ``Functional characterization and gene regulation of the archaeal virus SIRV2`` was submitted to the Faculty of Science, University of Copenhagen to obtain the degree. I have been financed by a scholarship from China Scholarship Council and a stipend from the European Union.

Almost all the experimental work presented in this thesis was performed at Danish Archaea Centre (DAC), Department of the Biology, University of Copenhagen, Copenhagen, Denmark, under the supervision of Associate Professor Dr. Xu Peng. Protein Circular dichroism (CD) spectroscopy was performed at the SBIN lab, Department of the Biology, University of Copenhagen and the high-throughput sequencing was carried out at Department of System Biology, Technical University of Denmark, Copenhagen, Denmark.

The thesis starts with a briefly summary of archaea and its viruses. Some typical viruses with unexpected morphotypes and genome structures were described and SIRV2 infection life cycle was also presented in detail. Then the second parts is the summary of the results mainly described ssDNA binding, annealing and nuclease activities of a conserved gene cluster of SIRV2, and a regulation map of two transcription regulators of *Sulfolobus solfataricus* P2 upon SIRV2 infection. At last, it ends with the conclusions and further perspectives for future work. Two manuscripts are enclosed behind.

Author: Yang Guo

Date: 05/11/14

Place: Copenhagen, Denmark
Acknowledgments

First and most, I would like to greatly thank my academic supervisor Xu Peng for endless support and always positive attitude, for her patient guidance and fruitful discussions. I also like to thank Xu for bringing me along the inspiring and exciting research trips from the very beginning of my biology study.

Further, I would like to thank Qunxin She and Roger A. Garrett for the valuable suggestions and scientific discussions.

Moreover, I want to thank the lab. technicians at Denmark Archaea Center, Hein Phan and Mariana Awayez for the great technical assistance in the laboratory.

Many great thanks to all the members in Denmark Archaea Center, Ling Deng, Fei He, Laura Alvarez, Daniel Jensen, Soley Gudbergsdottir, Wenyuan han, Wenfang Peng, Guannan Liu ,Carlos Sobrino, Kristine Uldahl and Marzieh Mousaei. You are inspring and talented people on various projects, and bring me many fun times in the laboratory.

Last but not least, I would like to thank my beloved family and friends outside the scientific environment for unreserved love and support throughout the duration of my Ph.D.
# Table of content

Preface II
Acknowledgments III
Summary (English) V
Resumé (Danish) VII
List of the Publications IX
Abbreviations X
Thesis Objective XIII
Introduction 1
1. Archaea 2
1.1 Classification of Archaea 3
1.2 *Sulfolobus* 5
2. Archaeal Viruses 8
2.1 Crenarchaeal viron morphotypes and their genome 9
2.2 *Sulfolobus islandicus* rod-shaped virus 2 13
2.3 SIRV2 life cycle 15
2.3.1 Attachment and Entry 15
2.3.2 SIRV2 Gene Transcription and Regulation 17
2.3.3 SIRV2 Replication 18
2.3.4 SIRV2 Release Mechanism 20
Summary of Results 23
Future Perspectives 28
Reference 30
Manuscript I 40
Manuscript II 85
Summary (English)

Viruses infecting hyperthermophilic archaea have gained wide attention during recent years owing to its remarkable diversity on morphology and genome structures. Although a substantial work was made to decipher the functions of the unique proteins encoded by archaeal viruses and to characterize the relationship of the viruses and host cells, the knowledge on the biology of the archaeal viruses is still limited. The crenarchaeal virus Sulfolobus islandicus rod-shaped virus 2 (SIRV2), was emerging as a promising model for genetic and biochemical studies as well as for the characterization of different stages in viral infection cycle. However, similar to other archaeal viruses, the majority of the SIRV2 genome sequence showed little similarity to the public databases, which hindered the virus functional researches and raised challenges in protein comparison and prediction.

This thesis comprises two parts of results. Firstly, the functional characterization of a highly conserved operon of SIRV2 was described, revealing their unique protein structures, biochemical activities as well as possible biological process they may participate in. In the second part, the genome wide regulations of two Sulfolobus solfataricus P2 transcription regulators upon SIRV2 infection were firstly constructed.

A SIRV2 gene operon (gp17, gp18 and gp19) was found to be the only and highly conserved gene clusters in rudi viruses and filamentous viruses, suggesting an important function in both viral families. The experimental results showed that ORF131b (gp17) was a novel ssDNA binding protein, without a canonical ssDNA binding domain. A few positively charged residues forming a U-shaped binding channel on the gp17 dimer are crucial for its ssDNA binding activity. The intrinsically disordered C-terminus of gp17 was demonstrated to be involved in the interaction with gp18, which was predicted previously as a helicase but showed a ssDNA annealing activity in this study. gp19 was shown to possess a 5’ to 3’ ssDNA nuclease activity, in addition to the previously demonstrated endonuclease activity, and a weak interaction between gp18 and gp19 was also detected. The functional characterization of the entire operon and the strand-displacement replication mode proposed previously for SIRV2 strongly point to a role of the operon in genome maturation and/or DNA recombination in viral gene DNA replication and repair.

Two transcription regulators sso2474 and sso10340 from Sulfolobus solfataricus P2 were differently expressed upon SIRV2 infection. A method similar to, but simpler than,
Chromatin immunoprecipitation combined with subsequent high-throughput sequencing (Chip-seq) was applied in this study to get insight into the gene composition of the two protein regulons in vivo. After mapping the sequence data with the genomes of Sulfolobus solfataricus P2 and SIRV2, protein sso2474 was detected to have a high binding affinity to virus genome by an unknown mechanism, whereas sso10340 or its interacted protein preferred to bind and regulate the host genes on several binding sites. A total of 27 enriched DNA fragments extracted from sso10340 complex were selected as candidate binding targets from the host genome for the further analysis using EMSA (Electrophoretic mobility shift assay) and foot printing assay. A palindromic sequence motif was defined based on the enriched sequences, and most of these target genes were involved in energy metabolism, transport and amino acid metabolism. The genome-wide binding profile presented here reflected two different kinds of regulon conditions and contribute to the knowledge expansion of the transcription regulation upon virus infection in Sulfolobus.
**Resumé (Danish)**

Vira der inficerende hypertermofile archebakterier har fået stor opmærksomhed i de seneste år på grund af deres bemærkelsesværdige mangfoldighed indenfor morfologi og genom strukturer. Selv om et stort arbejde bliver gjort for at identificere funktionen af de unikke proteiner arke virus kodet for og at beskrive forholdet mellem vira og værtsceller, er viden om arke viras biologi stadig begrænset. Den crenarchaeal virus *Sulfolobus islandicus* stavformet virus 2 (SIRV2), er et lovende model for genetiske og biokemiske undersøgelser samt til karakterisering af forskellige stadier af virusinfektionscyklus. Men i lighed med andre arke vira, har størstedelen af SIRV2 genom sekvens ringe lighed med sekvenser i de offentlige databaser, dette hindrede funktionel virus forskning og giver store udfordringer ved sammenligning af og funktionelle forudsigtelse af proteiner.

Denne afhandling består af to dele. Første beskrives hvordan en særdeles konserveret operon fra SIRV2 bliver funktionelle karakteriseret, her afsløres operons proteiners unikke strukturer, biokemiske aktiviteter samt mulig biologisk processer, de kan deltage i. I den anden del bliver to *Sulfolobus sofatarius* P2 transskription regulators mål identificeret i hele host genomet for første gang.

En SIRV2 gen operon (*gp*17, *gp*18 og *gp*19) blev anset for at være de eneste og højkonserverede genklynger i rudivirus og trådformede vira, hvilket tyder på en vigtig funktion i begge vira familier. De eksperimenterlige resultater viste, at ORF131b (*gp*17) var en hidtil ukendt ssDNA bindende protein uden et kanonisk ssDNA bindende domæne. Et par positivt ladede aminosyre, danner en U-formet substrat kanal på *gp*17 dimer. Dette er afgørende for *gp*17s ssDNA bindende aktivitet. Den naturlige uordnet C-terminalen del af *gp*17 blev påvist at være involveret i interaktionen med *gp*18. Som tidligere forudsigelser har klassificeret som en helicase, men i denne undersøgelse viste *gp*18 ssDNA bindende aktivitet. Det blev påvist at *gp*19 har 5' til 3' ssDNA nuklease aktivitet, udover den tidligere påvist endonukleaseaktivitet. Ydermere blev en svag interaktion mellem *gp*18 og *gp*19 blev også påvist. Den funktionelle karakterisering af hele operonet og streng-fortrængning replikation metode som tidligere er foreslået for SIRV2 peger kraftigt på operonens rolle i genomet modning og / eller DNA-rekombination af viral-DNA under replikation og reparation.
To transskription regulatorer sso2474 og sso10340 fra *Sulfolobus solfataricus* P2 blev forskelligt udtrykt ved SIRV2 infektion. En metode, der ligner, men er enklere end, Chromatin immunopræcipitation kombineret med efterfølgende høj-throughput sekventering (Chip-seq) blev anvendt i denne undersøgelse for at få indsigt i den genetiske opbygning af de to proteiners regulon *in vivo*. Efter kortlægning af sekvens data mod genomerne fra *Sulfolobus solfataricus* P2 og SIRV2 blev det påvist protein sso2474 have en høj affinitet til virus genom via en ukendt mekanisme, hvorimod sso10340 eller dets interaktion partner foretrak at binde og regulere værtsgener på flere steder på genomet. I alt 27 berigede DNA-fragmenter blev udvundet fra sso10340 kompleks blev udvalgt som mulige bindings mål i værtsgenomet og yderligere analyse ved hjælp af EMSA (Electrophoretic mobility shift assay) og fodaaftryk analyse. Et palindromt mønster blev defineret på basis af de berigede sekvenser. De fleste af de genre relateret til dette mønster var involveret i stofskiftet, aminosyretransport og metabolismen. Profilen for de to proteiners binding til DNA, der dækker hele genomet, afspejler to forskellige typer af regulons og er med til at udvide viden om regulation af transskription i relation til virus infektion i *Sulfolobus*. 
List of the Publications:

- Article I


  Submitted to Nucleic acid research.

- Article II

  Guo, Y., and Peng, X. Genome-wide binding profile of two transcription regulators from Sulfolobus solfataricus.

  In prep.

- Article III

**Abbreviations**

ABV, Acidianus bottle-shaped virus
ACV, Aeropyrum coil-shaped virus
AFV1, Acidianus filamentous virus 1
APBV1, Aeropyrum pernix bacilliform virus 1
ARV1, Acidianus rod-shaped virus 1
ASV1, Acidianus spindle-shaped virus 1
ATP, Adenosine Triphosphate
ATV, Acidianus two-tailed virus
Ala, Alanine
Amp, Ampicillin
bp, Base pair
BSA, Bovin serum albumin
Cam, Chloramphenicol
CRISPR, Clustered Regularly Interspaced Short Palindromic Repeats
dsDNA, Double-Strand DNA
DTT, Dithiothreitol
EB, Elution Buffer
E.coli, Escherichia coli
EDTA, Ethylenediaminetetraacetic acid
EMSA, Electrophoric mobility shift assay
GST, Glutathione-S-transferase
Hjr, Holliday junction resolvases
ICTV, International Committee on Taxonomy of Viruses
IPTG, Isopropyl-beta-D-thiogalactopyranoside
ITR, Inverted terminal repeats
Kan, Kanamycin
KDa, Kilo-Dalton
LB medium, Lysogeny Broth medium
Ni-NTA, Ni-nitritotriacetic acid
MALDI-TOF, Matrix-assisted laser desorption/ionization-time of flight
MCP, Major capsid protein
M.O.I, Multiplicity of infection
mM, Mili Molar
OD, Optical density
ORF, Open Reading Frame
PAV, Pyrococcus abyssi virus
PAGE, Poly acrylamide gel electrophoresis
PBS, Phosphate Buffered Saline
PCNA, Proliferating cell nuclear antigen
PCR, Polymerase Chain Reaction
PDB, Protein Data Bank
PMSF, Phenylmethylsulfonyl fluoride
PSV1, Pyrobaculum spherical virus 1
Pfu, Pyrococcus furiosus
RCR, Rolling-circle replication
rRNA, Ribosomal RNA
SIFV, Sulfolobus islandicus filamentous virus
SIRV1/2, Sulfolobus islandicus rod-shaped virus ½
SNDV, Sulfolobus neozealandicus droplet-shaped virus
SMRV1, Sulfolobales Mexican rudivirus 1
SRV, Stygiolobus rod-shaped virus
SSU, Small-subunit
SSV, Sulfolobus spindle-shaped virus
SSVK1, Sulfolobus spindle-shaped virus K1, Kamchatka
SSVrh, Sulfolobus spindle-shaped virus RH, Yellowstone
STIV1/2, Sulfolobus turreted icosahedral virus 1/2
STSV1, Sulfolobus tengchongensis spindle-shaped virus 1
TEMED, N, N, N’, N’-tetramethylethylenediamine
TTSV1, Thermoproteus tenax spherical virus 1
TTV1, Thermoproteus tenax virus 1
VAP, Virus-associated pyramid
WT, Wild-type
Thesis Objective

The objective of this PhD is mainly focused on the functional characterization of a conserved archaeal viral gene cluster-ORF131b (gp17), ORF436 (gp18) and ORF207 (gp19) of SIRV2 to investigate their possible roles in the whole virus life cycle. Besides, the genome wide regulation of two Sulfolobus solfataricus transcription regulators upon SIRV2 infection were also studied to get a better understanding of the regulation network between virus and host cells.
Introduction
1 Archaea

Evolution is a process through which the composition of genes in a population changes over generations, and it seems to progress in a quantized way, from one lever or domain of organization rising ultimately to a more complex one. In the early to middle 20th century, microbiologists tried to classify microorganisms based on the structures of their cell walls, their shapes, and the substances they consume. Until five decades ago, Zuckerkandl and Pauling claimed that it is at the level of molecules (particularly molecular sequences) that one really becomes privy to the workings of the evolutionary process. The comparative analysis of the molecular sequences started to become a powerful approach for determining evolutionary relationship (Zuckerkandl and Pauling, 1965).

The Ribosomal RNA was chosen to be a candidate molecule to detect relatedness among distant species due to broad distribution, slowly changed sequence and a component of self-replicating systems (Zablen et al., 1975). In 1977, Woese and Fox digested the 16S (18S) ribosomal RNA of the organisms with T1 RNase and subjected the products to two-dimensional electrophoretic separation, producing oligonucleotide fingerprint to identify the relationships of the living system. They found out that many of the prokaryotes once classified as bacteria belong to their own domain, which was later classified as a third domain – Archaea, meaning ancient and primitive in ancient Greek language (Woese and Fox, 1977).

The discovery of the new microbial kingdom eventually led to the classification of all known life into three major Domains: Eucarya (all eukaryotes), Archaea, and Bacteria, which was a significant breakthrough in the history of biology (Forterre et al., 2002). Actually in the early 1980s, people already realized that Thermoplasma and Halobacterium had close evolutionary affinity with Methanogens, all of which were the representatives of known archaea (Woese et al., 1990). For a long time, archaea were seen as extremophiles that only exist in extreme habitats such as hot springs and salt lakes with high salt concentration (Oren, 2002), low pH (Johnson et al., 2008) or high temperature (Stetter, 2006). At the end of the last century, more organisms were discovered along with new habitats were studied, archaea have been found in a wide variety of non-extreme environments, including marine waters (DeLong, 1992), freshwater sediments (Schleper et al., 1997) as well as all kinds of soil environments (Bintrim et al., 1997; Oline et al., 2006).
They are globally distributed in nature and have become common microbes in environment. Since archaea can survive in such harsh conditions, they can provide a source of enzymes that resist to heat and/or to acidity, which is a valuable treasure for industry (Breithaupt, 2001). The most familiar application of an archaeal enzyme is the thermostable *Pfu* DNA polymerase from *Pyrococcus furiosus*, allowing the accurate polymerase chain reaction (PCR) to be widely used in biology science researches. There are many hidden treasures in archaea still waiting to be deciphered by old and new Archaea lovers.

### 1.1 Classification of Archaea

Based on the pioneering work of Carl Woese, the small subunit ribosomal RNA (ss rRNA) is widely used in molecular phylogenetic studies to investigate the relationship between organisms, rather like some classification systems that trying to group archaea based on the shared structural features and common ancestors (Gevers et al., 2006). At the early stage, archaea was further classified into two distinct groups, that the methanogens as well as their relatives were named as Euryarchaeotes and the thermoacidiphiles, sulfurdependent ones were categorized as Crenarchaeota (Woese et al., 1990). Most of the cultivable and well-studied archaeal species exhibit in these two main phyla. In 2002, the peculiar species Nanoarchaeum equitans was found. It harbor the smallest archaeal genome with a spherical cell shape and had been given its own phylum – Nanoarchaeota (Hohn et al., 2002). Another small new group of thermophilic archaeal species, exhibiting an apparent affinity to the Crenarchaeota, but also sharing features with Euryarchaeota, were identified as Korarchaeota (Elkins et al., 2008;Anderson et al., 2008). A fifth group has also been created as Thaumarchaeota in recent years (Guy and Ettema, 2011).

Euryarchaeota, as one of the major phyla in Archaea, encompasses the most diversified phenotypes. The cultivated Euryarchaeota is subdivided into eight groups (Thermococci, Methanopyri, Methanococci, Methanobacteria, Thermoplasmata, Archaeoglobi, Halobacteria and Methanomicrobia ), while Methanogenesis was the main invention that occurred in the euryarchaeal phylum along with halophiles, some thermoacidiphiles as well as some hyperthermophiles (Gribaldo and Brochier-Armanet, 2006). In contrast, most of the cultivable Crenarchaeota strains belong to the thermophilic or hyperthermophilic species, showing a very limited phenotypic diversity (Forterre et al., 2002). However, since the marine archaeal group was discovered and identified as characteristic Crenarchaeota by
environmental rRNA, it is thought that they may be the extremely abundant archaea in the marine environment and could be a significant component of deep-sea metabolism (Fuhrman et al., 1992).

The orders -Thermoproteales, Caldisphaerales, Desulfurococcales and Sulfolobales represent the four lineages of the Crenarchaeotal branch of Archaea. Thermoproteales are rod-shaped extreme thermophiles or hyperthermophiles. They are the only organisms known to lack the canonical SSB proteins, instead possessing the protein ThermoDBP specifically bound to ssDNA (Paytubi et al., 2012). The Sulfolobus species are relatively easy to cultivate due to the aerobic lifestyle and relatively short doubling times, and as the only genetic manipulatable representatives in Crenarchaeota, have developed into model organisms to study their DNA repair, replication, transcription, chromosome integration, RNA processing, cell division, virus-host interaction systems as well as many other cellular aspects (Bernander, 2007).
Figure 1. Small subunit ribosomal RNA-based phylogenetic tree. The thick lineages represent Hyperthermophiles. (modified from Stetter, 2006)

1.2 Sulfolobus

Since the first description by T. Brock in 1972 about Sulfolobus acidocaldarius, isolated from a hot spring in Yellowstone National Park, this new group of sulfur-oxidizing
organisms has been of interest both evolutionarily and geochemically (Brock et al., 1972). *Sulfolobus* species has been isolated from a wide variety of acid thermal areas (in the USA, Italy, Iceland, Russia and elsewhere), with optimal growth occurring at pH 2-3 and temperatures of 75-80 °C, making them acidophiles and thermophiles respectively. So far, most strains isolated are able to grow heterotrophically as well as autotrophically. Since the *Sulfolobus* species have a wide geographic distribution, they are normally named after the location where they were first isolated, e.g. *Sulfolobus islandicus* strains were isolated in Iceland (‘island’ is German for ‘Iceland’) (Zillig et al., 1994), *Sulfolobus tengchongensis* from Teng Chong, China (Xiang et al., 2003) and *Sulfolobus solfataricus* from volcanic hot springs at Pisciarelli Solfatara (Zillig et al., 1980). Among these species, *S. solfataricus* is one of the best-characterized and most commonly used strains in laboratories.

**Figure 2.** Electron micrographs of *Sulfolobus solfataricus* strain, DSM 1617, thin section (from Zillig et al., 1980).

The *Sulfolobus* strains DSM 1616 and DSM 1617 (Fig 2.) were firstly named as *Sulfolobus solfataricus* by Zillig for having a similar GC content but significantly different RNA polymerase molecular weights with respect to *S. acidocaldarius* (Zillig et al., 1980). Moreover, they were newly renamed as *S. solfataricus* P1, P2 and have developed as the main model species that researchers work on, especially when the genome sequence of *S. solfataricus* strain P2 was published by She, the transcriptome map was drawn by Wurtzel, providing rich detailed information for the further work on DNA replication mechanism, cell cycle, transcription and large numbers of unknown genes (She et al., 2001; Wurtzel et al., 2010). Wealthy data, standardized methods, maturing genetic system, the easy
lab-cultivating advantages as well as the host species for studying virus-host interactions contributed to the construction of *Sulfolobus solfataricus* as a model organism (Albers et al., 2009; Worthington et al., 2003; Deng et al., 2009).

The widely used strain *S. solfataricus* P2 (DSM1617) showed a low fraction of susceptible cells to both *Sulfolobus islandicus* rod-shaped virus 2 (SIRV2) and *Sulfolobus* turreted icosahedral virus (STIV) infection (Okutan et al., 2013; Ortmann et al., 2008). In this work, the highly susceptible mutant strain *S. solfataricus* P2 5E.6 was selected as the host strain for SIRV2 infection study. The strain carries a deletion of CRISPR (Clustered regularly interspaced short palindromic repeats) clusters A-D, but shows similar phenotype on chromosome degradation and virus life cycle as the natural host strain *S. islandicus* LAL14/1 upon SIRV2 infection (Okutan et al., 2013; Bize et al., 2009).
2 Archaeal Viruses

Along with archaeal communities, viruses thrive in extreme conditions and play an important role in ecosystem dynamics. Like members of the other domains of life, archaea are infected with viruses. At the initial stage, the thermophilic viruses isolated from archaea domain resembled bacteriophages with head-tail in morphotype (Martin et al., 1984), and the subsequently discovered Euryarchaeal viruses were also similar with bacterial viruses. In contrast, as more habitats were studied, viruses were found infecting members of the kingdom Crenarchaeota (Sulfolobus, Acidianus, Pyrobaculum and Thermoproteus) and exhibiting highly diverse morphotypes and genomic properties. The extraordinary shape of some viruses have never been observed before (Prangishvili, 2003). Most crenarchaeal viruses have been isolated from hot terrestrial habitats, and they show adaption to their extreme environments like their host.

Due to the abundance and unique biology, some challenges arise with archaeal viruses. The low level of sequence similarity to public databases, novel biochemical mechanisms as well as difficulties in virus and host cultivation need to be addressed (Prangishvili and Garrett, 2005). After a relatively intensive study on archaeal viruses in the last decade, about 100 viral species have been sequenced and their genomic properties as well as relationships with the host cell have also been described. Among these viruses, only two single-strand DNA virus species have been discovered (Pietila et al., 2010; Mochizuki et al., 2012), the others all possess double-strand DNA genomes. According to the International Committee on Taxonomy of Viruses (ICTV), bacterial viruses comprise nine morphotypes, which belong to ten families. While archaeal viruses, exhibit 16 different morphotypes, and are classified into 15 families (Ackermann and Prangishvili, 2012) (Fig 3.). Although limited in number, compared with the viruses infecting bacteria, the diverse and unique morphotypes of archaea viruses revealed new insights into the viral world.
Figure 3. Virion morphotypes of prokaryotic viruses. Names of viral genera or families based on International Committee on Taxonomy of Viruses (ICTV) are indicated below the schematic virus particles. If an archaeal virus has not been assigned to any genus or family, individual virus names are given. The virions are not drawn to scale (from Pietila et al., 2014).

2.1 Crenarchaeal virion morphotypes and their genomes

Thermophilic viruses infecting the crenarchaea have been classified into ten families based on their morphology, eight families were approved by ICTV and the remaining two are waiting for approval (Pietila et al., 2014). The ten crenarchaeal virus families are: One tail spindle –shaped Fuselloviridae (SSV1-7, SSVK1, SSVrh, ASV1); two tail spindle–shaped Bicaudaviridae (ATV); Bottle-shaped Ampullaviridae (ABV); Droplet-shaped Guttaviridae (SNDV); Linear filamentous Lipothrixviridae (AFV1-9,SIFV,TTV1); Linear rod-shaped rudiviridae (SIRV1-2,SRV,ARV1); Spherical Globulaviridae (PSV1,TTSV1); Bacilliform Clavaviridae (APBV1), Tailless icosahedral `Turriviridae` (STIV,STIV2) and Coil-shaped Spiravridae (ACV), the last two families are waiting for the approval. Some other viruses
like *Sulfolobus tengchongensis* spindle-shaped virus (STSV1) and *Pyrococcus abyssi* virus (PAV1) are awaiting assignment to a viral family. Some well-studied and intriguing viruses will be described in more detail as below:

**Sulfolobus spindle-shaped virus 1 (SSV1).** The *Sulfolobus* spindle-shaped viruses (SSVs) of the family *Fuselloviridae* were the first discovered family of archaeal viruses. Most of the SSVs (except for SSV6 and ASV1) are spindle-shaped, 100 x 60 nm in size and carry tail structures at one pole (Fig 4.).

The virus SSV1 was isolated from UV-induced growing cultures of *Sulfolobus shibatae* (strain B12). It contains a 15.5-kb positively supercoiled circular double-stranded DNA, with a GC-content of 39.7 %, resembling that of the host DNA (Palm et al., 1991). During infection SSV1 is stably carried by its lysogenic host *S. shibatae* and is found intracellularly either in a covalently closed circular (plasmid) form or site-specifically integrated within an arginine tRNA gene in the host chromosome (Yeats et al., 1982). The transcription pattern of the SSV1 genome is relatively simple, some genes are significantly upregulated by UV irradiation and the genes can be clearly divided into early, late and UV-inducible categories (Reiter et al., 1987; Frols et al., 2007).

**Acidianus two-tailed virus** (ATV). This archaeal virus was discovered in an acidic hot spring (85–93 °C; pH 1.5) at Pozzuoli, Italy. As the sole member of the viron family *Bicaudaviridae*, ATV contains a lemon-shaped central structure, but when it exits the host cell, it then develops elongated tails protruding from both pointed ends, specifically at
temperatures above 75°C, close to the temperature of the natural habitat of the host (Fig. 5). The circular, dsDNA genome contains 62730 bp, encodes 72 predicted proteins, 11 of which are structural proteins with molecular masses in the range of 12 to 90 kDa. The unique host-independent as well as extracellular functional activity might be associated with an 88.7-kDa ATV viron protein P800, which is rich in coiled-coil motifs and can generate structures that resemble intermediate filaments (Haring et al., 2005c; Prangishvili et al., 2006c). ATV was the first known virus from hot, acidic habitats that causes lysis of its host cell, whereas most archaeal viruses maintain a stable relationship with their host.

**Figure 5.** Electron micrographs of *Acidianus convivator* and different forms of the *Acidianus* two-tailed virus. **a,** Virions in an enriched sample taken from acidic hot springs in Pozzuoli, Italy (pH 1.5, 85–93 °C). **b,** Extrusion of lemon-shaped virions from an ATV-infected *A. convivator* cell. **c,** Virions in a growing culture of ATV-infected *A. convivator,* 2 days after infection. **d,** Cultured virions after purification and incubation at 75 °C for 0, 2, 5, 6 and 7 days (panels from right to left, respectively) (from Haring et al., 2005c).

**Acidianus bottle-shaped virus (ABV).** The enveloped virion of ABV, has a complex form resembling a bottle (230-nm long, 4–75-nm wide, Fig 6.C), the morphology is so unique that it has been assigned to a new family *Ampullaviridae.* The narrow end of the bottle is likely to be involved in cellular adsorption and in channeling of viral DNA into the host cell (Fig 6.A) and the broad end exhibits 20 thin filaments, which are inserted into a disk and interconnected at the base, the function of these filaments remains unclear but very intriguing (Fig 6.B) (Haring et al., 2005a).

ABV was isolated from the same hot spring in Pozzuoli, Italy, where ATV was isolated. It infects strains of the *hyperthermophilic* archaeal genus *Acidianus,* and contains a linear double-stranded DNA. The viron genome has a length of 23,814 bp, with a G+C content of
35%, and a 590-bp inverted terminal repeat. It encodes 57 predicted ORFs, of which a putative RNA molecule was predicted to have a notable secondary structural similarity to the bacteriophage RNA molecule, which has been implicated in DNA packaging. Moreover, in contrast to other crearchaeal viruses, ABV encodes a Family B DNA polymerase (Peng et al., 2007).

Fig. 6. Electron micrographs of particles of ABV after negative staining with 3% uranyl acetate. (A) ABV particles adsorbed with their pointed end toward a membrane vesicle of the host “A. convivator.” (B) ABV particles attached to each other with their thin filaments at the broad end. Bars, 100 nm. (C) A scheme of the structure of an ABV virion (from Haring et al., 2005a).

Acidianus filamentous virus 1 (AFV1). AFV is a Lipothrixivirus that infects the Acidiannus genus of the Crenarchaeota in a stable carrier state and was observed in an enrichment culture from a hot spring at 80 °C in Crater Hills region of Yellowstone National Park (Rachel et al., 2002). AFV1 is composed of a protein core covered with a lipid envelope, containing at least five different proteins with molecular masses in the range of 23-130 kDa. The 20.8-kb-long linear genome contains 40 ORFs and particles of AFV1 are measured with size of 900 × 24 nm.

AFV1 exhibits claw-like terminal structures, connected to the virion body by appendages at the both ends (Fig 7.A). Apparently, the unusual termini of the virions have a special function in the process of adsorption, which was detected to have an attachment with the host pili and the contact seems rather strong (Fig 7.B) (Bettstetter et al., 2003). Crystal structures of two major coat proteins AFV1-132 and AFV1-140 have been resolved, both carry a novel four-helix-bundle fold and AFV1-140 also carries an extra C terminal domain possibly interacting with the virion envelope (Goulet et al., 2009b). Recently, a new replication model was proposed by analyzing the replicative intermediates on
two-dimensional (2D) agarose gel, revealing that the genome replication started from a D-loop formation, proceeded via strand displacement, and terminated by recombination. This process in some degree resembled the T4 DNA replication, but further studies are still needed to support the proposed model (Pina et al., 2014).

**Fig 7.** (A) Electron micrographs of particles of AFV1 with tail structures in their native conformation. (B) Electron micrographs of particles of AFV1 adsorbed to pili of host *A. hospitalis* CH10/1, stained with 3% uranyl acetate. Black arrows indicate pili; white arrows show knots which are putative viral terminal structures separated from the virus body. Bars, 100 nm. (from Bettstetter et al., 2003)

2.2 *Sulfolobus islandicus* rod-shaped virus 2 (SIRV2).

*Rudiviridae* and *Lipothrixviridae* belong to the linear viruses, and they are ubiquitous in high temperature (>75°C) and low pH (pH <3) terrestrial geothermal environments. Comparative-genomic analysis suggests a common evolutionary ancestry of the rudiviruses and lipothrixviruses, based on the conservation of orthologous core genes and the similarity of the major viron coat proteins (Prangishvili et al., 2006a). Together with *Sulfolobus islandicus* rod-shaped virus 1 (SIRV1), *Stygiolobus* rod-shaped virus (SRV) (Vestergaard et al., 2008b), *Acidianus* rod-shaped virus 1 (ARV1) (Vestergaard et al., 2005) and *Sulfolobales* Mexican rudivirus 1 (SMRV1) (Servin-Garciduenas et al., 2013), SIRV2 was grouped as rudiviruses by their rod-shaped morphology, gene architecture and sequence.

SIRV2 is one of the most extensively studied archaeal viruses and has developed into the archaeal model virus thanks to the structural, genomic and transcription studies. SIRV2 was first isolated from the colony-cloned *S.islandicus* stains HVE 10/2 isolated from solfataric fields in Iceland-Hveragerdi. This non enveloped virus is a stiff rod of 23 nm in width, 900 nm in length. As shown in the electron micrographs, a central cavity ends plugging by
approximately 50 nm stoppers were clearly visible in Figure 8, and both ends decorated with three short tail fibers. Sensitivity of SIRV2 genome to BAL31 but not λ exonuclease indicated the existence of covalently closed hairpin ends (Prangishvili et al., 1999).

The linear 35,502 bp double stranded DNA SIRV2 genome carries inverted terminal repeats (ITRs) of 1628 bp at each end, and with a low G+C content of 25%. The virion body is a superhelix formed by genomic DNA and multiple copies of the highly glycosylated 20-kDa capsid protein. SIRV2 genome encodes 54 ORFs, sharing 44 homologs with SIRV1, and approximately half of the encoded proteins have been characterized by sequence, structural and biochemical analysis, which is the highest proportion on recognizing gene functions among crenarchaeal viruses. Four SIRV2 viron proteins were identified as virion structure proteins: the major capsid protein (MCP), P134 (gp26), shares a common fold with MCPs of lipothrixivirus AFV (Acidianus filamentous virus) (Goulet et al., 2009a). The largest viral protein P1070 (gp38), has a MW of 105 kDa, possesses a coiled-coil domain. It is a component of the three fibers (Steinmetz et al., 2008). Besides, the two structural proteins ORF488 (gp33) and ORF564 (gp39) are also found in the SIRV2 virions, although in a low amount (Vestergaard et al., 2008b). Crystal structure resolution of SIRV1 P119 as well as the detection of nicking and joining activities by experiments suggest that this protein could be involved in initiation of SIRV1 genome replication (Oke et al., 2011). P121 has a high sequence similarity with archaeal Holliday junction resolvases (Hjrs) and the Hjr activity was experimentally examined (Birkenbihl et al., 2001). There are more hypothetical proteins predicted by sequence analysis and experimental evidence that involved in transcription, replication and nucleic acid metabolism, more detailed information about the whole life cycle of SIRV2 will be discussed below.
2.3 SIRV2 life cycle

2.3.1 Attachment and Entry

The archaeal viruses display an unusual and diverse morphotypes, genome sequences, as well as the structure of proteins (Krupovic et al., 2012). Recent researches revealed that the interaction between archeal viruses and their hosts seem also to be unique (Bize et al., 2009). However, compared with wealth of data available on bacterial and eukaryotic systems, the studies on archaeal viruses mainly consistent of biochemical and genetic characterization of their virions, while the attachment and entry process are still elusive.

Virus infection is initiated by entry into the host cell, and the first step of the entry process is to recognize the receptors present on the host cell surface by specific interaction. Then they must have ways to transporting their genetic information to the cell compartment where their genome is replicated (Poranen et al., 2002). The vast majority of known viruses have a tail structure decorated to one or two ends of nucleocapsid, which facilitate the attachment of virions to the host membrane. In the *Lipothrixviridae* family, each of the virion is tapered and carries different specific terminal structures. These structures can represent claws (AFV1), T-bars (AFV9), mop-like structures (SIFV), three (AFV3) or six (SFV) short
filaments or tips resembling bottle brushes (AFV2), which are implicated in cellular adsorption (Bettstetter et al., 2003; Bize et al., 2008; Haring et al., 2005b).

Both termini of the SIRV2 virion are connected with three tail fibers composed of the minor structural protein P1070. These termini were detected to bind the tips of the pilus-like filaments, which are abundant on the surface of host cells, by transmission electron microscopy and whole-cell electron cryotomography (cryo-ET). Figure 9 demonstrated the interaction between SIRV2 termini fibers and purified host cellular filaments. The virus adsorption was very fast and irreversible, but the infected cells were no longer able to adsorb more virus efficiently (Quemin et al., 2013). Many bacterial viruses like Ff inoviruses, utilize the filamentous cellular appendages as primary receptors. Then retraction of the host pilus bring the viron close to the host cell surface, where it could bind to the secondary receptor (Rakonjac et al., 2011). Although no retracting pili have been identified in archaea, there should be secondary receptors on the host cell surface to adsorb the virus particles. Indeed, Sulfolobus mutant strain lacking cluster sso3138-sso3141 and cluster sso2386-sso2387 was resistant to SIRV2. No growth retardation was observed when this mutant strain was diluted and infected with SIRV2 at the same M.O.I, compared with wide type strain. The first clusters sso3138 to sso3141 were predicted to possess transmembrane helices and to be located extracellularly, probably acting as a receptor for SIRV2. The proteins encoded by the other gene cluster may be involved in the secretion of the receptor components. Besides, the genetic complementation experiments confirmed the involvement of the mutation in virus resistance and further support that these proteins are responsible for SIRV2 entry (Deng et al., 2014).

**Figure 9.** Transmission electron micrographs of SIRV2 interaction with purified cellular filaments. The filaments were removed from S. islandicus LAL14/1 cells (from Quemin et al., 2013).
However, how the virus overcomes 12.5-\(\mu\)m-long filament to reach the cell body, and the mechanism of removing their coat protein as well as the association of the two identified gene clusters with the structure of pili is still poorly understood and need further studies.

2.3.2 SIRV2 Gene Transcription and Regulation

SIRV2 enters into the cell, removes the coat protein and is likely to recruit the host RNA polymerase complex for transcribing the SIRV2 genes, as no viral gene was shown to encode a RNA polymerase. Generally, the virus transcription is time regulated and could be classified as early, middle and late transcribed genes that encode the proteins involved in regulation, translation, replication and structure proteins for assembly in a chronological way.

In bacterial and eukaryal virus-host systems, modification of cellular transcription as a result of virus infection is well studied, such as T7 bacteriophage. Besides transcribed by host \textit{E.coli} RNA polymerase, T7 bacteriophage encodes its own RNA polymerase, which is a single subunit enzyme of 99 kDa. The DNA genome of T7 is transcribed entirely from left to right, firstly in the early region by \textit{E.coli} RNA polymerase, and then from a portion of the early region to the entire late region by newly-made T7 RNA polymerase (Dunn and Studier, 1983; Steitz, 2004). Whereas in archaeal domain, the mechanisms and controls of viral gene expression as well as host gene regulation upon virus infection are still not elucidated. To date, several archaeal viruses have developed as suitable models to study molecular details of the archaeal viron life cycle and host responses, \textit{e.g.} the temperate \textit{Sulfolobus} spindle-shaped viruses (SSV) (Frols \textit{et al.}, 2007) and the lytic \textit{Sulfolobus} turreted icosahedral virus (STIV) (Ortmann \textit{et al.}, 2008; Maaty \textit{et al.}, 2012).

Reminiscent to the life cycles of bacteriophages and eukaryotic viruses, SSV1 exhibits a tight temporal regulation of its own transcription after UV treatment, initiating from a small UV-specific gene and then continues as three distinct sets of genes representing immediate-early, early and late transcripts. But very few of host genes was regulated upon virus infection (Frols \textit{et al.}, 2007). However, the microarray study about transcription of the lytic virus STIV was completely an opposite story. STIV transcription did not show a typical temporal regulation. In the virus life cycle, transcription signals of nine early viron
genes were detected at 8h, then all the remaining genes were transcribed subsequently. Surprisingly, a total of 177 host genes were determined to be differentially expressed during the infection, of which two thirds were up-regulated and one third were down-regulated (Ortmann et al., 2008).

SIRV2 infects Sulfolobus spp. but does not integrate into the host chromosome. It was thought to be a lysogenic virus existing in a stable carrier state in the host, corresponding to the uniform transcription pattern revealed in an earlier study (Kessler et al., 2004). During the characterization of the special release mechanism of SIRV2, it was then demonstrated to be a lytic virus (Bize et al., 2009). Independent microarray study performed in infected S. solfataricus 5E6 cells and transcriptomic analysis of infected S. islandicus LAL14/1 cells exhibited that SIRV2 transcription starts from the terminal genes located at both ends of the linear genome (Quax et al., 2013;Okutan et al., 2013). Although SIRV2 transcription is not tightly regulated chronologically like SSV1, the gene expression showed a temporal pattern. Some early genes like two identical ORF83a, ORF83b as well as ORF119C, the viral replication initiation protein (Oke et al., 2011), were detected to be transcribed in 15-30 min and highly expressed at 1h. Whereas the structure proteins like the major capsid protein and virus-associated pyramids protein were most abundant at the late stage of the infection cycle. Despite lacking strong temporal regulation of transcription on its own virus, the host response to SIRV2 infection was significant. More than one third of the host genes were differentially regulated, with a similar number of down and upregulated genes. Most of the host genes that are strongly activated upon infection are assumed to function in defense against viruses, as well as cellular collapse, energy metabolism and membrane transport, which may suggest that the virus control the replication phases less dependent on its own differential gene expression, but co-opted host genes.

2.3.3 SIRV2 Replication

As we know that SIRV2 genome is a linear duplex with covalently closed hairpin termini and long ITRs at both ends (Blum et al., 2001). This termini structure is normally involved in replication initiation that parallels to that of Poxviridae and other large cytoplasmic eukaryotic viruses. Both DNA sequence and structure within the termini are important for template recognition, which is nicked by Rep initiation protein and exposed a 3'-OH group as a primer for DNA synthesis (Du and Traktman, 1996).
By sequence and structure analysis, ORF119 of SIRV2 was found to be a member of the replication initiator (Rep) family, having a conserved key active-site motif with rolling-circle replication (RCR) rep proteins (Vega-Rocha et al., 2007). It forms a dimer and sequence-specifically nicks one strand of the SIRV2 terminal hairpin only when the substrate is in the single-strand form. The joining activity of ligating the fragments by a strand transfer (flip-flop) mechanism was also confirmed (Oke et al., 2011). According to all these features, along with the detection of head-to-head concatamers of the replicative intermediates (Peng et al., 2001), a related but unrestricted mechanisms to rolling-circle replication (RCR) was proposed. The Rep protein recognizes and nicks one ori site of the genome, then one subunit of rep protein covalent connected with the new generated ori site and the other subunit of the protein ligated the old two fragments, forming a contiguous DNA circle. Displacement replication is then used to replicate the rest of the genome and generated a double strand DNA circle, but with a nicked hairpin termini adducted rep protein. The next steps of the replication are similar to RCR of the poxviruses. At the junctions between genome monomers, opposing inverted terminal repeats can be extruded to form hairpin fourway junctions. Therefore, a Holiday junction resolving enzyme (Hjr) was supposed to introduced to resolve the concatamers, producing monomer copies with linear hairpin ends (Culyba et al., 2006; Oke et al., 2011).

Holliday junction resolving enzymes are ubiquitously found in all the domains of life, such as RuvC in Bacteria, Human GEN1 in Eukarya (Declais and Lilley, 2008), and two different holiday junction resolving enzymes (Hjr and Hje) from Sulfolobus solfataricus of crenarchaeon (Kvaratskhelia and White, 2000). As expected, SIRV2 encodes a 14 kDa Holliday junction resolving enzyme (SIRV2 Hjr), which is conserved among rudiviruses. Unlike the bacteriophage resolving enzymes, which cleave a variety of branched DNA structures formed during replication, the SIRV2 Hjr showed a very narrow substrate range, only cleaves the four-way junctions DNA structures, and the cleavage pattern is also unique by nicking only exchange strand pairs (Gardner et al., 2011a). This protein was presumed to be important for the processing of replicative DNA intermediates late during the infection cycle before packaging into newly synthesized heads commences.

Unlike some bacterial viruses encoding DNA replication related proteins, most of the archaeal viruses lack its own DNA polymerase genes, indicating that their replication probably rely on the host replication machinery. It was proved by recently published work that two of the heterotrimeric S. solfataricus sliding clamp (SsoPCNA1 to 3) (proliferating
cell nuclear antigen) interacted with some SIRV2 viral proteins. PCNA is a key protein functioning as a cofactor of DNA polymerases recruiting different crucial DNA metabolism proteins in DNA replication and repair (Moldovan et al., 2007). Most of the interacting viral proteins have not been assigned function, except SIRV2 Hjr, which agreed to previous research released that SsoPCNA could stimulate the Hjr enzyme activity in *S. solfataricus* (Dorazi et al., 2006). It is intriguing that the early transcribed genes ORF83a/b were also shown to interact with PCNA, suggesting its important roles during the replication cycle of SIRV2 (Gardner et al., 2014).

Although the functions of some replicative viral proteins were confirmed, and a preliminary model was proposed, a lot of further studies are still needed to discover and explain the virus replication mechanism.

### 2.3.4 SIRV2 Release Mechanism

The final step for completion of the viral replication cycle is the release of virus particles. In bacterial domain, most lytic viruses cross the cell envelope and spread to the environment with the assistant of phage-encoded small integral membrane proteins--holins (Krupovic and Bamford, 2008). How the archaeal viruses overcome the challenging task of rupturing the cell membrane and escape from the host cells have attracted a lot of attention in recent years, especially after the discovery of a unique release mechanism (Bize et al., 2009).

Among archaeal viruses, SIRV2 and STIV are the best studied viruses with respect to host cell interactions. Both of them are lytic viruses, and shared the same extraordinary virion egress mechanism. SIRV2 induces the degradation of the host chromosome and assemble virus particles in cytoplasm (Fig. 10A). In the late stages of the virus infection cycle, numerous prominent virus-associated pyramids (VAPs) were formed on the host cell surface (Fig. 10B and C), and these special structures open outward at the end of infection cycle, allowing the escape of the mature viruses through the created apertures (Bize et al., 2009; Brumfield et al., 2009). Apparently, this release mechanism is not universal for hyperthermophilic viruses. Although sharing the release mechanism, the two viruses are dramatically different in their morphological properties. Therefore, it is possible that the morphogenetic and egress systems evolved independently.
In order to investigate the special structural protein components of SIRV2-infected cells, three different virus-infected cell fractions were collected, compared and analyzed: the total cell lysate, the membrane and the cytosol fractions. It was found that the 10 kDa P98 of SIRV2 is the only protein appearing specifically in the membrane fraction of infected cells and is exposed on the surface that rupturing the S-layer, no other viral protein is involved in the assembly of pyramids. After overexpression of SIRV2-ORF98 in E. coli and S. acidocaldarius, the VAPs were also formed with the same size and shape as those formed in S. islandicus infected with SIRV2 (Fig. 10F) (Quax et al., 2010;Quax et al., 2011). The sequence alignment data revealed that no other archaeal virus carried the homologue of the SIRV2-ORF98, except for STIV and the Rudiviridae (SIRV1/2, SRV). It is also intriguing that the VAPs was a separate structural unit which can be isolated and purified, and the solo protein SIRV2-ORF98 is capable of self-assembling into ordered sevenfold isosceles
triangle–shaped pyramid (Fig. 10D and E), which seems to be a baseless and hollow structure (Quax et al., 2011).

Although the similar VAPs were observed in heterologous expressed *E. coli* and *S. acidocaldarius*, they only existed in the surface of the inner membrane and all of the VAPs were in closed state. There must be at least one special factor induced the VAPs opening, which is absent in *S. acidocaldarius* and *E. coli*, but is present in its native host cells.
Summary of Results
*Sulfolobus islandicus* rod-shaped virus 2 (SIRV2), as a member of the family *Rudiviridae*, is a promising candidate to become a general model for detailed studies of archaeal virus biology due to its relatively easy laboratory cultivation and sufficient yields. To date, several important stages of its biological life cycle have been characterized such as viral entry, transcription pattern, genome replication as well as its unique egress mechanism, which provide us a much better understanding of the unknown archaeal viral world.

Even so, similar to the vast majority of other archaeal viruses showing little sequence similarity to public databases, the functions of many SIRV2 proteins remain to be identified. Only one fifth of the 54 ORFs encoded by SIRV2 genome were experimentally confirmed a function, and the knowledge of its basic molecular processes like DNA repair, recombination, genome maturation as well as the interaction with its host are still limited.

Although possessing limited sequence similarities with the public gene bank, a CRISPR-associated Cas4-like protein ORF207 (*gp19*), previously identified as a ssDNA endonuclease, has drawn a lot of interests (Gardner *et al.*, 2011b). It was detected to be transcribed from a single promoter with two other proteins (*gp17* and *gp18*), located at its upstream, and generated a polycistronic transcript (Kessler *et al.*, 2004). This organization of proteins suggests related functions. Moreover, the bioinformatic analysis revealed that this operon constitutes the most conserved gene cluster in archaeal linear viruses including rudiviruses and filamentous viruses. Then it has raised questions regarding the functions of this entire operon and the related virus infection stages they may be involved in.

The resolved crystal structure of *gp17* homolog encoded by SIRV1 indicated a DNA binding activity, although no obvious structural similarity was matched in Protein Data Bank. Different structural DNA substrates were tested for its binding activities, and the results demonstrated that either ssDNA or dsDNA with a single or double flaps can be shifted with the protein *gp17*, no blunt-end dsDNA could form the protein-DNA complex, which indicate that *gp17* is a ssDNA binding protein. However, none of the documented classical ssDNA binding domains were found in the structure of *gp17*, therefore this protein constitutes a novel non-canonical ssDNA binding protein. Sequence alignment of *gp17* homologs revealed 3 highly conserved and 5 relatively conserved residues. Mutagenesis of a
few conserved basic residues distributed in two adjacent loops within each monomer suggested a U-shaped binding path for ssDNA. As gp18 couldn’t be cloned into *Sulfolobus* cells due to its toxicity and as recombinant expression in *E.coli* resulted in the formation of inclusion bodies, a denaturation and refolding strategy was employed to purify the His-tagged gp18 from *E.coli*. Both the circular dichroism spectroscopy and gel-filtration chromatography assay showed that the refolded protein gp18 is functional stable to be used for the further study. BlastP search of gp18 sequence suggested a weak similarity to bacterial ATPase domains of Lon protease, and a tertiary structure prediction suggested a function as hexameric helicase. However, neither protease or helicase activity of gp18 could be detected under all possible conditions. Instead, gp18 was detected to be able to increase the dsDNA yield from two complementary oligonucleotides. The failure of detecting the helicase activity could be due to the lack of proper experimental conditions or possible mask of helicase activity by the stronger annealing activity. It also could be that gp18 carries no helicase activity, but only annealing activity, possessing the similar features as the annealing helicases (HARP, AH2). (Yusufzai and Kadonaga, 2008; Yusufzai and Kadonaga, 2010).

To better understand the function of the entire gene operon, the protein product of the third gene, gp19, was further characterized in this study, which was detected to have a 5’-3’ ssDNA exonuclease activity, in addition to the previously demonstrated ssDNA endonuclease activity.

There are 38 aa residues missing at the C-terminus of gp17 in the crystal structure, which was predicted as disordered region by two different program IUpred and PONDR. The disordered C-terminus of bacterial SSB proteins are normally involved in protein-protein interactions. Since the entire operon all work in the same type of substrate, ssDNA, the interactions among the three proteins were performed by GST affinity chromatography. The experimental results demonstrated that gp17 interacts with gp18 and the C-terminal disordered domain of gp17 is essential for the interaction. No interaction was detected between gp17 and gp19, but a weak interaction was shown between gp18 and gp19. In order to confirm whether gp17 could recruit some ssDNA-processing proteins as bacterial SSBs, this gene was inserted into plasmid and transformed into host *Sulfolobus solfataricus* P2 cells for the pull-down assay *in vivo*. Two host proteins sso2277 and reverse gyrase
(sso0422) were detected to interact with gp17. Protein structure prediction of sso2277 revealed a high similarity with RecF and RecN proteins, which involved in DNA replication / recombination.

The operonic or clustered organization of the three genes in rudi- and filamentous viruses and the observed interactions between their protein products strongly suggest their close cooperation in a same process(es) involving ssDNA. This process could be the SIRV2 genome maturation, replication or recombination, and new evidences are needed to support the hypothesis.

Besides available information concerns unusual viral morphological and genomic properties, the SIRV2 transcription pattern as well as the regulation of host genes during virus infection was studied, either by microarray analysis or by deep transcriptome sequencing (Okutan et al., 2013; Quax et al., 2013). Although lacking of strong temporal regulation of transcription on its own virus, the host response to SIRV2 infection was significant. More than one third of the host genes were differentially regulated, with a similar number of downregulated and upregulated genes. Among these regulated genes, there are two transcription regulators sso2474 and sso10340 from Sulfolobus sulfataricus P2 were responded differently. Then we are curious to find out if any host genes or virus genes were regulated by these two regulators, and whether they are a local or global acting transcription factors. In this work, we investigate the binding targets of the two proteins in an in vivo context by performing a method similar to chromatin immunoprecipitation combined with DNA sequencing.

In order to detect the regulation on both host genes and viral genes, after the proteins were overexpressed for 15h, the cells were infected with SIRV2 at about M.I.O of 10 for 2.5 h. His-tagged protein purification was carried out from virus infected cells, the protein sso2474 was detected to bind hundreds of fold more DNA than the control group, and sso10340 exhibited a range of oligomeric states resembling the feature of Lrp/AsnC family proteins. The bound DNA was separated from DNA-protein complex from the two purified proteins, respectively, and sent for the high-throughput sequencing. The alignment between sequenced data and virus genome or host genome revealed that most of the DNA bound by sso2474 is viral DNA and sso10340 is mainly associated with the host regulation.
However, the specific binding targets and the binding mechanisms of sso2474 are still not identified, and the experiments showed that this protein purified from *E.coli* preferred to bind dsDNA than ssDNA. A total 27 binding target regions by protein sso10340 or its interacted proteins in *S.solfataricus* P2 were identified, and half of them located in the upstream or partial upstream of the corresponding genes, while the other half fell within the gene coding regions. A 11bp palindromic binding motif was defined by analysis of the enriched oligonucleotide sequences, which was present in 96% of the binding targets. The functions of these related genes were categorized and most of which were involved in energy metabolism, transport and amino acid metabolism.
Future Perspectives
The PhD work is the first study providing the functional characterization of an entire gene operon conserved in archaeal rudiviruses and filamentous viruses as well as the general regulation profile of two host regulators. There are still some work to be done in the future to enlarge the knowledge of the archaeal viral biology and virus-host interaction.

The toxicity of gp18 to *Sulfolobus* cells as well as its insolubilities in *E.coli* hindered the progress of characterization of the whole operon. Although insertion of both gp17 and gp18 genes into *Sulfolobus solfataricus* cells could decrease the strong toxicity of gp18, almost no expressed gp18 can be purified (data not shown). One method we would like further to try is co-expressing the two or three proteins in *E.coli* in a suitable system to test if gp17 or gp19 could help the folding of gp18, since both of them were demonstrated to interact with gp18. If so, the next plan is crystalizing the complex with a synthesized oligonucleotide to further investigate the interaction in detail between the complex and ssDNA.

The ssDNA binding, annealing and nuclease activities *in vitro* were all characterized in this study, and the possible function in viral infection cycle are discussed, whereas more *in vivo* evidence is still needed to complete the scenario we constructed. Due to the limited viral genetic technologies and relatively large size of this virus, silencing these genes in virus is not possible until now. We already tried to overexpressing the c-terminal truncated gp17 in the host cells for competition with the wide-type one in virus to detect its influence either in virus replication or genome maturation. However, the result is not conclusive due to different reasons. It would be interesting and exciting to use some good ideas and methods to detect the viral function of this operon *in vivo*.

The very surprising thing about sso2474 is its special high affinity to viral DNA in a non-sequence binding way. The DNA binding mechanism of sso2474 was not clear, and the phenotype changes and no growth retardation to virus are expected to be observed in the sso2474 mutated organism, if it can be knocked out. At last, the global regulation of sso10340 need to be further validated. Is there any other DNA binding protein or regulators interacting with sso10340 and whether this regulator activates or represses the transcription of the corresponding genes are still need to be confirmed.
Reference


Manuscript I

Single-stranded DNA binding, annealing and nuclease activities encoded by a conserved archaeal viral gene cluster

Yang Guo, Birthe B. Kragelund, Malcolm F. White and Xu Peng

Submitted to Nucleic Acid Research
Single-stranded DNA binding, annealing and nuclease activities encoded by a conserved archaeal viral gene cluster

Yang Guo¹, Birthe B. Kragelund¹, Malcolm F. White² and Xu Peng¹*

¹ Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, 2200 CPH N. Denmark

² Biomedical Sciences Research Complex, University of St. Andrews, North Haugh, St. Andrews, Fife, UK

* To whom correspondence should be addressed. Tel: +45 35322018; Fax: +45 35322128; Email: peng@bio.ku.dk
ABSTRACT

Single-stranded DNA (ssDNA) occurs in various cellular and viral processes of DNA metabolism, including DNA replication, homologous recombination and repair pathways. Here, we describe a novel type of ssDNA binding protein, a novel ssDNA annealing protein and a ssDNA nuclease encoded by an operon comprised of ORF131b (gp17), ORF436 (gp18) and ORF207 (gp19), respectively, of Sulfolobus islandicus rod-shaped virus 2 (SIRV2). Rather than comprising one of the canonical ssDNA binding domains, SIRV2 gp17 forms a dimer with each monomer containing two α-helices and three β-strands. Mutagenesis of a few conserved basic residues distributed in two adjacent loops within each monomer suggested a U-shaped binding path for ssDNA. Although predicted previously as a helicase, the recombinant gp18 showed a ssDNA annealing activity often associated with helicases and recombinases. Moreover, gp19 was shown to possess a 5´ to 3´ ssDNA exonuclease activity, in addition to the previously demonstrated ssDNA endonuclease activity. Further, in vitro pull-down assay demonstrated interactions between gp17 and gp18 and between gp18 and gp19 with the former being mediated by the intrinsically disordered C-terminus of gp17. The strand-displacement replication mode proposed previously for rudiviruses and the close interaction between the ssDNA binding, annealing and nuclease activities strongly point to a role of the gene operon in genome maturation and/or DNA recombination which may function in viral DNA replication/repair.

INTRODUCTION

Viruses that infect extreme hyperthermophilic archaea, the third domain of life, are unusual in their morphology, genome structure and proteins. In the last decade, a major effort has been undertaken to study the archaeal viruses, which have attracted intense interest as model systems to understand the biochemistry and molecular biology required for life at high temperatures. Based on their morphological and genomic characteristics, 15 viral families have been classified and about 100 viral isolates described, all with either linear or circular double-stranded (ds) DNA, except two species possessing a single-stranded (ss) DNA genome (1-3).

The Sulfolobus islandicus rod-shaped virus 2 (SIRV2) (4), together with SIRV1 (5), Stygiolobus rod-shaped virus, SRV (6), Acidianus rod-shaped virus 1, ARV1 (7) and Sulfolobales Mexican rudivirus 1 (SMRV1) (8), belong to the family Rudiviridae. The
Rudiviruses have linear dsDNA genomes (24.6 to 35.8 kbp) with inverted terminal repeats and the two strands at the genomic termini are covalently linked. Recently SIRV2 has been the focus of genomic, structural, genetic and transcriptional studies, which have provided important insights into its entry, gene regulation and unique release mechanisms (5;9-13). Even so, similar to the vast majority of other archaeal viruses showing little sequence similarity to public databases (14), the functions of many SIRV2 proteins remain to be identified.

Among the 54 ORFs encoded in the genome of SIRV2, only one fifth had been experimentally assigned a function (15). The virus is coated with one major capsid protein gp26 and three minor structural proteins gp33, gp38 and gp39 (16). Together with the genes encoding the viral structural proteins, gp49, encoding the component of the pyramidal egress structure (11), is repressed by the transcription regulator gp15 (SvtR) during the early virus infection cycle (17;18). gp16, belonging to the replication initiator (Rep) family and nicking one strand of the viral genomic termini, was proposed to be involved in the initiation of the DNA replication (19). The Holiday junction resolving enzyme (Hjr) gp35 was suggested to resolve the concatamers of the replicative intermediates, producing monomeric copies with linear hairpin ends (20). Taken together, the functions of many SIRV2 genes remain unknown and the knowledge of its biology and basic molecular processes such as DNA replication, recombination and maturation is still limited.

In this work we studied a SIRV2 operon containing three genes, gp17, gp18 and gp19 that are highly conserved in rudiviruses and filamentous viruses. gp19 was previously shown to be an endonuclease specifically cutting ssDNA (21). We demonstrate here that gp17 is a ssDNA binding protein and interacts with gp18 while the latter stimulates annealing of complementary oligonucleotides. In addition to the previously identified ssDNA endonuclease activity, we detected the 5’ to 3’ ssDNA exonuclease activity of gp19, which also interacts with gp18. Based on the data, the possible functions of the gene operon are discussed.

MATERIALS AND METHODS

Cloning, expression and purification of C-terminally His-tagged recombinant proteins

The coding sequences of gp17, gp18 and gp19 were amplified by PCR from SIRV2 genome using primers listed in Table S1, digested with NdeI and XhoI and subsequently inserted
into a similarly digested pET-30(a) (Novagen) expression vector. To introduce single or multiple amino acid (aa) mutations into the recombinant gp17, fusion PCR using 4 primers (Table S1) was performed for each mutant.

_E.coli_ BL21 CodonPlus cells were transformed with individual plasmid construct and a single clone transformant was inoculated in LB medium containing 30 µg/ml kanamycin and 25 µg/ml chloramphenicol. At an optical density (OD$_{600}$) of 0.4, IPTG (0.5 mM) was added to the culture and the cells were further cultured at 25°C for 12 hours. Harvested cell pellet was resuspended in lysis buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 1 mM EDTA, 1% Triton X-100 and 1 mM PMSF) and lysed by sonication. The lysate was cleared by centrifugation at 10000 x g for 20 minutes and the supernatant was then incubated with Ni-NTA-agarose beads (Qiagen, Germany) for 1 h at room temperature. The beads were washed three times with washing buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 40 mM Imidazol) and the protein eluted with elution buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 250 mM Imidazol). The purity of the proteins was evaluated by SDS-PAGE and the gel (12.5%) was stained with PAGE blue (Sigma Aldrich, UK). In the case of gp18, a denaturing and refolding method was applied (see below).

**Cloning, expression and purification of N-terminally GST-tagged proteins**

The wild-type gp17, gp17-I, its truncated mutants gp17-II (1-121aa) and gp17-III (1-111aa) and gp19 were amplified by PCR using primers listed in Table S1 and the PCR products were digested with BamHI and XhoI and ligated to a BamHI and XhoI digested pGEX-6p-2 (GE Healthcare Life Science, Sweden) expression vector. The constructs were introduced individually into BL21 CodonPlus cells. Transformed cells were grown in LB medium containing 100 µg/ml ampicillin and 25 µg/ml chloramphenicol and induced with 0.5 mM IPTG at an OD$_{600}$ of 0.4. After 12 hours incubation at 25°C, the cells were pelleted, resuspended in lysis buffer (PBS buffer pH 8.0, 1 mM EDTA, 1% Triton X-100 and 1 mM PMSF) and lysed by sonication. The supernatant was incubated with Glutathione Sepharose 4 Fast Flow beads (GE Healthcare Life Science, Sweden) for 1 h at room temperature. The beads were washed three times with PBS buffer and the proteins remain bound on the beads for subsequent pull-down assays with His-tagged proteins.

**Refolding of His-tagged gp18 from inclusion bodies**
The cell lysate was centrifuged at 10000 × g for 20 min, and the inclusion bodies in the pellet were solubilized in lysis buffer containing 8 M urea at room temperature for 1h. The purification of the denatured protein using Ni-NTA-agarose beads followed the same procedure as described above for other His-tagged proteins, except that 8 M urea was included in both washing and elution buffers. The eluted 2 ml protein from 1L E.coli cells was dialysed first in 200 ml 0.5 M L-Arginine buffer for 2 h, and then in 2 L PBS buffer for 3 h and the latter dialysis was repeated for 3 times.

Preparation of substrates for DNA mobility shift and nuclease activity assays

To prepare DNA substrates for DNA mobility shift and exonuclease assays, oligonucleotide 1 (Table S2) was annealed to a series of fully or partially complementary ssDNA oligonucleotides (Table S2) to generate either a 23-nucleotide (23-nt) 5´-ssDNA tailed duplex (substrate B), a 23-nt 3´-ssDNA tailed duplex (substrate C), or a blunt-ended duplex (substrate A) (Table 1). Oligonucleotide 4 (Table S2) was annealed to its partially complementary ssDNA oligonucleotides 5 (Table S2) to generate a Y-shaped dsDNA (substrate D). The annealing mixture was heated at 95°C for 2 min and then slowly cooled to room temperature (25°C) over a period of 1 h. M13mp18 DNA (New England Biolabs, America) was chosen as circular single-stranded DNA substrate for the endonuclease assay.

Gel mobility shift analysis

50 nM of the ssDNA (oligo 4) or dsDNA (substrate A, B and D in Table 1) were incubated for 20 min at 50°C with increasing concentrations of gp17 or its mutant variants (0-2000 nM) in 20 µl DNA-binding buffer (10 mM Tris-Cl, pH 8.0, 100 mM KCl, 2 mM DTT, 10% [vol/vol] glycerol). The samples were loaded onto 12% acrylamide gel and electrophoresed in 0.5 × TBE buffer for 1 h 50 min. Following electrophoresis, the gels were stained with SYBR® Gold (Life Technologies) and scanned by Typhoon FLA 7000 (GE Healthcare Life Science). The bands were quantified using ImageQuant TL (GE Healthcare Life Science).

Gel-filtration chromatography

Gel-filtration chromatography was carried out using an ÄKTA–FPLC system. Briefly, purified proteins in PBS buffer were applied individually to a Superdex 200 HR 10/300 GL column (GE Healthcare Bio-Sciences, America) equilibrated with the same buffer. The column was
operated at a flow rate of 0.5 ml/min, and 0.5 ml fractions were collected. The proteins were detected by measuring the absorbance at 280 nm, 254 nm and 215 nm. The column was calibrated with proteins of known molecular weight: Thyroglobulin, Bovine (669 kDa), Apoferritin, Horse Spleen (443 kDa), β-Amylase, Sweet Potato (200 kDa), Alchol Dehydrogenase, Yeast (150 kDa).

Circular dichroism (CD) spectroscopy

A far-UV CD spectrum was recorded on a Jasco 810 spectropolarimeter at a wavelength range from 260 to 190 nm, a scan rate of 20 nm/min, 15 accumulations and 2 s response time, at room temperature. Samples were recorded in a quartz cuvette with a 1mm path length. A corresponding spectrum of the buffer was recorded and subtracted and the resulting spectrum smoothed (Jasco software). The spectrum was recorded of 3.85 µM protein in PBS, pH 8.0 and the ellipticity given as mean residue ellipticity \[
\theta_{\text{MRW}} \text{ in } \text{deg}^*\text{cm}^2\text{dmol}^{-1}.
\]
A temperature denaturation profile was recorded at 220 nm by heating the sample from 25°C to 95°C with a rate of 1°C/min, and apparent melting temperatures T\text{m}^\text{app}s derived from fitting of the data to the following equation:

\[
\theta(T) = \frac{(\beta_F + \alpha_F T) + (\beta_U + \alpha_U T) \cdot \exp \left( \frac{T\Delta S - \Delta H}{RT} \right)}{1 + \exp \left( \frac{T\Delta S - \Delta H}{RT} \right)}
\]

Where \(\Delta H(T_m)\) is the enthalpy change at Tm, and \(\Delta S(T_m)\) the entropy change at Tm.

ssDNA annealing activity

For ssDNA annealing assay, the \(^{32}\text{P}\) end-labelled 57-nt oligo 4 (1 nM) (Table S2) was incubated in annealing buffer (30 mM Tris-HCl, pH7.5, 5 mM MgCl\(_2\), 75 mM NaCl, 50 mM KCl and 1 mM DTT) with increasing amounts of gp18 at 25°C for 5 min. The reaction was initiated by adding the unlabelled complementary oligonucleotide 5 (1.2 nM) (Table S2) and incubated at 50°C for 15 min. The reaction was then stopped by the addition of 20 mM cold oligo 4, 0.5% [wt/vol] SDS and 1 mg/ml proteinase K. The deproteination was carried out at 25°C for 10 min and the samples were loaded on a 10% native polyacrylamide gel and run at 100V for 1 h 20 min in 0.5 × TBE buffer. Following electrophoresis, gels were dried and exposed to X-ray film for documentation. DNA was quantified using ImageQuant TL (GE Healthcare Life Science).
**Nuclease activity assays**

The nuclease activity assays (20 µl) were performed by mixing 0.08 µM DNA duplex (substrate A, B or C) or 0.05 µM M13mp18 DNA in reaction buffer containing 20 mM Tris-HCl, pH 7.5, 50 mM NaCl, 5 mM MgCl₂, 1 mM DTT and 10% glycerol. Reactions were initiated by the addition of 0.5 µM SIRV2 gp19, and the mixtures were incubated at 50°C for the indicated length of time. Time course analyses were carried out by scaling up the reaction volume to 150 µl and withdrawing 20 µl aliquots at the indicated times. Reactions were terminated by the addition of 6 µl stop solution (0.1% [wt/vol] bromophenol blue, 0.1% [wt/vol] xylene cyanol, 8% [vol/vol] glycerol, 1% [wt/vol] SDS, 50 mM EDTA and 2 mg/ml protease K). As a negative control, the substrates were incubated in the reaction mix in the absence of protein gp19. Samples for the exonuclease assay were analyzed by 12% polyacrylamide gel electrophoresis in 0.5 x TBE buffer and stained with SYBR® Gold (Life Technologies). Samples for the endonuclease assay were resolved in 0.7% agarose gel.

**Western blot analysis**

Proteins were separated in a 12.5% SDS polyacrylamide gel and transferred with transfer buffer (39 mM Glycine, 50 mM Tris pH 8.7, 0.04% SDS, 20% Methanol) onto a nitrocellulose membrane (Whatman, Germany) at 70 mA for 1h 15 min. The membrane was blocked for 1 h with PBST buffer (PBS buffer containing 0.05% Tween® 20) containing 8% milk powder and then incubated with anti-His antibody (Qiagen, Germany, 1:2000 dilution in PBS buffer containing 3% BSA). The membrane was further incubated for 1 h with a peroxidase-coupled secondary antibody (anti-mouse 1:10 000 IgG, Sigma-Aldrich, America) followed by 3 times wash with PBS buffer. An alkaline phosphate (Sigma-Aldrich, America) substrate was used for detection according to the instructions provided.

**In vitro detection of protein-protein interactions**

The GST-tagged gp17-I, gp17-II (1-121aa), gp17-III (1-111aa) or gp19 remaining on the GSH beads (50 µg) was incubated at 25°C for 1 h with His-tagged prey proteins (90 µg) in PBS buffer (0.1 µg/ml BSA, 0.1% Triton X-100). The mixture was centrifuged for 3 min at 3000 x g and the supernatant was stored for later analysis. The beads were washed several times with PBS buffer to remove unbound components and then heated in SDS loading buffer at 98°C for 10 min before SDS-PAGE in a 12.5% gel. The gel was stained with PAGE
blue (Sigma Aldrich, UK) and the presence of His-tagged interaction partners was detected by Western blot using anti-His antibody, as described above.

**Identification of *Sulfolobus* proteins interacting with gp17**

The *gp17* fragment amplified by PCR (Table S1) was inserted between NdeI and NotI restriction sites of the *Sulfolobus/E.coli* shuttle vector pEXA2 (22), allowing the expression of His-tagged gp17 under the control of arabinose promoter. The constructed plasmid, as well as the empty plasmid pEXA2, were then electroporated individually into the uracil deficient competent cells (23). Single colonies of the transformants were inoculated into test tubes containing 5 ml SCV (basal medium supplemented with 0.2% sucrose, 0.2% casamino acids and 1% vitamin solution) (23), and incubated in an Innova 3100 oil-bath shaker. Large-scale culturing was performed in ACV medium (0.2% D-arabinose was substituted for sucrose) with Erlenmeyer flasks of long necks.

Purification of the His-tagged gp17 from the transformed *Sulfolobus* cells was performed as described above, and the eluted proteins were evaluated by 12.5% SDS-PAGE and stained with PAGE blue (Sigma Aldrich, UK). Protein bands present exclusively in the gp17-containing transformant were sliced from the gel and subjected to MALDI-TOF analysis (Alphalyse A/S, Odense, Denmark).

**RESULTS**

**Bioinformatic analysis revealed high conservation of SIRV2 *gp17, gp18* and *gp19* in archaeal rudiviruses and filamentous viruses**

*gp17, gp18* and *gp19* occur as a gene cluster in the genome of SIRV2 and were shown previously to be transcribed from a single promoter generating a polycistronic transcript (10). The operon organization suggests the three genes are functionally related. However, except *gp19* which was experimentally determined to possess a ssDNA endonuclease activity (21), very little is known about the function of *gp17* and *gp18*. The crystal structure of *gp17* homolog encoded by SIRV1 has been resolved (24), but no functional insight is available. Although a weak similarity between a limited part of *gp18* and bacterial ATPase domains of Lon proteases was described (15), a tertiary structure prediction using the threading program Phyre 2 (25) suggested a different function. About two thirds of *gp18* sequence
(residues 140 - 430) matched with high confidence (98.8%) to MCM homolog 2 (c3f8tA) from *Methanopyrus kandleri*, suggesting that gp18 might be a hexameric helicase.

The three genes are conserved in all rudiviruses and in one of the filamentous viruses, AFV1 (Fig. 1). The amino acid (aa) sequence similarities range between 36 - 96% for gp17 homologs, 51 - 100% for gp18 homologs and 51 – 98% for gp19 homologs. Although no significant sequence similarity was detected between SIRV2 gp19 and the other filamentous viral genomes, a putative nuclease is clearly encoded by the latter (except AFV2) and it belongs to the Cas4 superfamily similar to SIRV2 gp19 (6;26). Interestingly, a highly conserved gene encoding a putative helicase is found upstream of the putative nuclease gene in all filamentous viral genomes. Upstream of the putative helicase gene is another highly conserved gene encoding a 79 aa hypothetical protein showing no sequence similarity to SIRV2 gp17 (Fig. 1). It is not clear whether this small gene is functionally related to SIRV2 gp17, while homologs or analogs of SIRV2 gp18 and gp19 are present in almost all rudiral and filamentous viral genomes. Genome comparison of all the rudiral viruses and filamentous viruses using the Mutagen program (27) revealed that SIRV2 *gp17, gp18 and gp19* constitute the only conserved gene cluster in the archaeal linear viruses (Fig. S1 and Table S3). Thus, the three genes appear important for both viral families.

**gp17 is a single-stranded DNA binding protein**

*Structural features of SIRV2 gp17.* To gain insights into the functions of the gene cluster, we first examined the structure of gp17, which forms a dimer in the crystals (Protein Data Bank identifier [ID] 2X5T)(24)(Fig. S2A). Although it doesn’t show obvious structural similarity to any known domains present in Protein Data Bank (PDB), the electrostatics and the shape of the molecule indicate a DNA binding activity. It is dominated by basic (blue) residues on the concave side (Fig. S2C) and by acidic (red) residues on the convex side (Fig. S2D). The concave side fits well as a DNA straddling pocket. Two arginine side chains point down in the middle of the arch. The dimer binds a sulphate group, which may reflect its affinity towards phosphates.

The total length of gp17 is 131 residues, and structural information is missing for 38 residues at the C-terminus. In line with this lack of structural information, analysis of the sequence with protein disorder predictors revealed a potential intrinsically disordered C-terminus of about 35 residues (28-30) (Fig. S3). Analysis of gp17 homologs encoded by
other rudiviruses and AFV1 revealed intrinsically disordered C-terminus of similar size (data not shown), indicating the importance of the disorder in this domain for the function of the protein.

**ssDNA binding activity of gp17.** gp17 was amplified from the SIRV2 genome and cloned into *E. coli* vector pET30a. The C-terminally His-tagged protein was purified from *E. coli* to homogeneity (Fig. S4) and tested for binding activity to different DNA substrates. As shown in Fig. 2A, gp17 binds to substrates that are either ssDNA or dsDNA containing a single or double flaps, whereas no binding to the blunt ended dsDNA was detected with the same range of protein concentrations. The same result was obtained when ssDNA and blunt ended dsDNA were mixed with equal molar concentration in the reaction, where almost all ssDNA, but no or very little dsDNA, were shifted in the presence of 130 nM gp17 (Fig. 2B). At higher concentrations of gp17, no free ssDNA is available, and the protein exhibited binding to the dsDNA, albeit with a much lower affinity. At a gp17 concentration of 3.9 µM, a significant amount of the dsDNA still remains unshifted, indicating that the affinity of gp17 towards ssDNA is at least 30 times higher than to dsDNA.

**A few positively charged residues forming a U-shaped binding channel on the gp17 dimer are crucial for its ssDNA binding activity**

To identify essential elements of the ssDNA-binding domain of gp17, we first aligned the sequences of gp17 homologs and identified 3 fully conserved positively charged residues, R60, K61 and K82 (Fig. S5). R60 and K61 are located in a loop at the central cleft of the concave side, while K82 is found on the surface of the convex side (Fig. 3A). The three residues were mutated individually into alanine and the binding affinity of the mutant proteins to ssDNA was compared to that of the wild-type gp17. Whereas the K82A variant exhibited a similar level of binding affinity as the WT gp17, a 2 and a 5 fold reduction in binding affinity was observed, respectively, for the R60A and the K61A variants. Interestingly, simultaneous mutation of R60 and K61 to alanine abolished almost completely its ssDNA binding activity (Fig. 3B).

Four other positively charged residues, R24, K27, K29 and R33, are relatively conserved in the rudiviruses (Fig S5), and the corresponding residue of SIRV2 R33 in AFV1 ORF135 (K32) is also positively charged. Therefore the four residues were mutated to alanine either individually or simultaneously. While the double mutant R24K27 and the triple
mutant R24K27K29 showed very little or only a mild reduction in the binding affinity (Fig. S6), the R33 mutant demonstrated the most profound effect within the tested single mutants, with an 8 fold drop of the binding activity observed (Fig. 3B and 3C).

The above experiments demonstrated that R33, R60 and K61 are important for the ssDNA binding whereas R24, K27, K29 and K82 are less or not important. By examining the location and the orientation of the residues, it is obvious that the side chains of the former residues point to the central cleft and those of the latter residues point to the outer surface (Fig. 3A). It appears that the residues important for DNA binding (R33, R60, K61) form a positively charged and U-shaped structure in the gp17 dimer, thus straddling on ssDNA and causing bending of the ssDNA (Fig. 3D).

Within the U-shaped path another relatively conserved residue, H54, has a positive charge (Fig. S5) and was thus mutated to alanine to test its possible contribution to ssDNA binding. In line with its charge, conservation and location in the protein, H54 appears also important for ssDNA binding, as the H54A mutant demonstrated a 4 fold drop of binding activity compared with the WT protein (Fig. 3B, 3C and 3D).

**Purification, refolding and stability of gp18**

To characterize gp18 biochemically, a certain amount of soluble protein was needed. As gp18 couldn’t be cloned into *Sulfolobus* (see below) due to its toxicity and as recombinant expression in *E.coli* resulted in the formation of inclusion bodies, a denaturation and refolding strategy was employed to purify the His-tagged gp18 from *E.coli*. Following cell lysis, the inclusion bodies were pelleted and dissolved in 8 M urea (Fig. 4A lane 3), and gp18-His was purified using Ni-NTA-agarose beads. The denatured gp18-His was then refolded in L-Arginine buffer (31). The protein appeared refolded properly as it remained soluble in the solution after 20 min incubation at 70°C (Fig. 4A lane 4). To assess the fold integrity of gp18 after recombinant production and refolding, the final preparation was subjected to structure analyses by CD spectroscopy. The far-UV CD spectrum recorded at room temperature revealed distinct negative molar ellipticity with minima at 218 and 208 nm, strongly indicating that the protein is folded with content of both α-helices and β-strands (Fig. 4B). Additionally, a temperature denaturation monitored at 220 nm showed that the protein was stable with two cooperative transitions, one with an apparent melting temperature (T_{m}^{app}) of
~60°C and a major, highly cooperative transition with a $T_{m}^{\text{app}}$ of 91°C, $\Delta H(T_m) = -469 \pm 15$ kJ/mol and $\Delta S(T_m) = -1.29 \pm 0.04$ kJ/mol (Fig. 4C). Because of the extreme stability of the protein, the post transition was not perfectly revealed at the current experimental conditions, and hence the thermodynamic parameters are approximations. However, conclusively, recombinantly produced gp18 was highly stable, cooperatively folded and with content of both $\alpha$-helices and $\beta$-strands.

The oligomerization status of the refolded gp18 was analysed by gel filtration chromatography, which resulted in the formation of a broad peak containing two "shoulders" with a total elution volume of 23.6 ml at a flow rate of 0.5 ml/min (Fig. 4D). The elute volume of the main peak (labelled 1 in Fig. 4D) was between 9.85 and 10.60 ml and those of the two "shoulders" were 8.05 ml and 12.75 ml, respectively. Assuming that gp18 has a shape and partial specific volume similar to those of standard proteins, the molecular mass of the main peak is estimated to be between 685 kDa and 458 kDa and that of the two "shoulders" to be Vo and 147 kDa, respectively, calculated from a standard linear regression equation, $K_{av} = -0.2976(\log\text{MW}) + 1.8388$ (Fig. S7). Since the molecular weight of the monomeric gp18 is 50.36 kDa, the protein was refolded as a series of oligomers from trimmers, nonamers to dodecamers. The molecular mass of the main top (560.6 kDa) ranges from nonamers to dodecamers, which might be the functional folds of gp18.

**gp18 stimulates the annealing of complementary oligonucleotides**

Since structural prediction suggested that gp18 could be a hexameric helicase such as MCM, an essential protein for the initiation and elongation phases of DNA replication (32), we performed helicase assays. However, no helicase activity was detected in spite of multiple trials with different nucleic acid substrates and varied experimental conditions with different metal ions, NTPs, and temperatures (data not shown). Surprisingly, during the helicase assays, we found that instead of unwinding dsDNA, gp18 seemed to be able to increase the dsDNA yield from two complementary oligonucleotides.

To determine whether gp18 catalyzes ssDNA annealing, a $[^{32}\text{P}]$-labelled 57-nt oligonucleotide (oligo-4 in Table S2) was first mixed with gp18. The annealing reactions were initiated by the addition of the complementary strand (oligo-5) and stopped by a 20-fold excess of the unlabeled oligo-4. After deproteinization, the reaction products were resolved on a native PAGE gel. gp18-mediated annealing was dependent upon protein
concentration, with the reaction being most efficient at 200 nM of gp18 under the experimental condition (Fig. 5A, lane 5), also supporting the oligomeric structure of the protein. Moreover, the efficiency of annealing was not changed when ATP was excluded from the reaction (Fig. 5A, lane 6), suggesting that ATP hydrolysis is not needed in the catalyzed process.

As shown in the left panel of Fig. 5B, spontaneous annealing between the two oligonucleotides occurred slowly in a time-dependent manner with only 40% of the ssDNA annealed after 4 minutes incubation (Fig. 5C). The annealing process was drastically accelerated by gp18 (right panel of Fig. 5B). The oligonucleotides were almost completely annealed to form the slow-migrating dsDNA after 4 minutes of incubation in the presence of gp18 (Fig. 5B and 5C).

**gp19 demonstrates both ssDNA endonuclease and 5´-3´ exonuclease activities**

Although previously demonstrated as a ssDNA endonuclease (21), gp19 shares sequence similarity with different CRISPR-associated Cas4 proteins (26), which possess metal-dependent endonuclease and 5´→3´exonuclease activities against ssDNA (33). We therefore examined the possible exonuclease activity of gp19 using DNA substrates of different structures. As shown in Fig. 6A, the migration of the blunt–end duplex DNA did not change upon the addition of gp19, indicating that it is not a substrate of gp19. While the 3´-flap duplex DNA remained unchanged as the blunt-ended DNA, the 5´-flap duplex DNA was cleaved with the final product having the same size of the blunt-ended DNA. It appeared that gp19 initiated cleavage from the 5´ single-strand end and stopped at the single strand and double strand junction (Fig. 6B). This indicates that gp19 has the 5´-3´ ssDNA exonuclease activity.

To confirm the ssDNA endonuclease activity of gp19, the circular ssDNA M13mp18 was tested with the same reaction buffer. As shown in Fig. 6C, the incubation with gp19 led to slow degradation of the circular ssDNA. These results confirm that SIRV2 gp19 possesses both 5´→3´exonuclease activity and endonuclease activity against ssDNA.

**Interactions between gp17 and gp18 and between gp18 and gp19**

Given that gp17, gp18 and gp19 all work on the same type of substrate, ssDNA, it appeared possible that the three proteins interact with one another. Since the removal of the last 10
residues showed little effect on the DNA binding activity of gp17 (Fig. S6), the intrinsically disordered C-terminus is possibly involved in other functions such as protein-protein interactions as demonstrated for the disordered C-terminus of bacterial SSB proteins (34). Therefore, we first tested its possible interactions with gp18 and gp19.

gp17 and its C-terminally truncated variants were expressed as GST fusion proteins and purified separately (Fig.7A). Individual GST fusion proteins were incubated with His-tagged gp18, and immobilized on GSH beads. After centrifugation, the beads were washed and boiled in SDS buffer before loading on SDS gel for Western blotting. Western hybridization using His-tag antibody revealed the presence of gp18 on the GSH beads with immobilized wild type gp17 protein, indicative of the interaction between gp17 and gp18 (Fig. 7B). However, no interaction was detected between gp18 and the two gp17 variants with 10 and 20 C-terminal residues removed, respectively (Fig. 7C). These results demonstrate that gp17 interacts with gp18 and the C-terminal disordered domain of gp17 is essential for the interaction.

The same method was applied to test possible interactions between GST-tagged gp17 and His-tagged gp19, and between GST-tagged gp19 and His-tagged gp17, none of which showed positive results (data not shown). While no interaction was detected between gp17 and gp19, the GST-tagged gp19 retained a small amount of His-tagged gp18 on the GSH beads (Fig. 7B), demonstrating a weak interaction between gp18 and gp19.

gp17 binds to two Sulfolobus host proteins

ssDNA binding proteins are essential for protecting ssDNA and recruiting specific ssDNA-processing proteins. In bacteria, SSBs were found to interact with more than a dozen different proteins involved in DNA replication, recombination and repair (34). To identify possible interactions with other proteins, gp17 was cloned into the E.coli/Sulfolobus shuttle vector pEXA2 under the control of arabinose promoter (22) and expressed in Sulfolobus. By Ni-NTA-Agarose chromatography the His-tagged gp17 was co-purified with two large proteins, of about 60 and 150 kDa, respectively (Fig. 7D). The absence of the two bands in proteins purified from the control cells transformed with empty pEXA2 supported that they were pulled-down specifically by gp17. Western blot hybridization using His-tag antibody revealed a single band with the expected size of gp17-His, and the two large bands were thus not oligomers of gp17-His (Fig.7E).
The two bands were sliced from the gel and identified by MALDI-TOF analysis. Band 1 contained a hypothetical protein encoded by SSO2277 with a theoretical mass of 57 kDa, carrying an ATPase domain. Band 2 was identified to be reverse gyrase from *S. solfataricus P2* (SSO0422) with a mass of 142 kDa (Table S4). The same procedure was repeated with SIRV2 infected transformants and revealed again the same results (data not shown). No viral proteins such as gp18 were co-purified with gp17, which could be due to low expression of *gp18*, as demonstrated previously by microarray analysis (12).

We attempted to clone *gp18* and *gp19* individually into *Sulfolobus* using pEXA2 as cloning vector. Whereas *gp18* was shown to be highly toxic and couldn’t be transformed into *Sulfolobus*, overexpression of gp19 caused growth retardation of the transformant, and no host proteins were identified to interact with gp19 (data not shown).

**DISCUSSION**

Single stranded DNA binding proteins are ubiquitous across all three domains of life and are found in many viruses playing essential roles in genome maintenance, DNA replication, recombination, repair and transcription. They can coat, protect and remove secondary structures of the ssDNA intermediates. Besides, some specific ssDNA-processing proteins are recruited and coordinated by ssDNA binding proteins during DNA metabolism pathways (35-37). In spite of high sequence, structural and functional divergence, almost all classical ssDNA binding proteins contain one of the following four structural topologies: oligonucleotide/oligosaccharide/oligopeptide-binding (OB) folds, K homology (KH) domains, RNA recognition motifs (RRMs), and whirlly domains (38). Recently a group of hyperthermophilic archaeal organisms were found to lack a classical ssDNA binding protein and instead to harbour a distinct ssDNA binding protein termed ThermoDBP (39). The ssDNA binding protein encoded by SIRV2 *gp17* differs in structure from the classical ssDNA binding proteins as well as from the ThermoDBPs, and thus constitutes a novel non-canonical ssDNA binding protein.

Single strand annealing activity has been detected in different proteins including some helicases and recombinases encoded by cellular life and by some viruses (reviewed by (40). In many of the helicases containing annealing activity, a separate protein domain distinct from the helicase domain is responsible for the annealing activity (40). Remarkably, a helicase domain-containing protein, HARP, was recently discovered to possess annealing,
but no unwinding, activity (41). HARP binds to the ssDNA binding protein RPA and anneals RPA-coated complementary ssDNA. Mutations in HARP are associated with Schimke Immuno-Osseous Dysplasia (SIOD) disease and the defects in the annealing activity of two HARP mutants correlate with the severity of the disease (41). Together with AH2, another protein with similar features (42), HARP was termed annealing helicase. In this study, the annealing activity was clearly demonstrated for the SIRV2 gp18 protein. The failure of detecting the helicase activity, which was predicted by structural modelling of the gp18 sequence, could be due to the lack of proper experimental conditions or possible mask of helicase activity by the stronger annealing activity. A third possibility is that gp18 carries no helicase activity, as demonstrated for the annealing helicases. While only the structural modelling revealed a connection between SIRV2 gp18 and a MCM helicase, a high sequence similarity to Cas3 and other helicases was clearly detected by BlastP searches of the gp18 analogues encoded in the genomes of most filamentous viruses (Fig. 1). Interestingly, the *E. coli* Cas3 was found to possess both helicase and annealing activities (43).

To better understand the function of the entire gene operon, the protein product of the third gene, *gp19*, was further characterized in this study which revealed a 5'-3' ssDNA exonuclease activity, in addition to the previously demonstrated ssDNA endonuclease activity (Fig. 6 and Garder et al., 2011b). The operonic or clustered organization of the three genes in rudi- and filamentous viruses (Fig. 1) and the observed interactions between their protein products (Fig. 7) strongly suggest their close cooperation in a same process(es) involving ssDNA. The SIRV2 genome replication study by different approaches demonstrated that SIRV2 forms ssDNA intermediates larger than a single genome size, and large concatamers are abundant during the replication process (Martinez-Alvarez et al., in preparation). This requires, first of all, abundant ssDNA binding protein to protect the ssDNA intermediates and the highly expressed gp17 (12) may fulfil this requirement. To mature into dsDNA monomers, the long ssDNA concatamers must anneal between the two complementary strands, which could be facilitated by gp18. Subsequent nicking by a ssDNA endonuclease and ligation by an unknown ligase would produce a mature dsDNA genome. Through protein-protein interactions, gp17 can recruit gp18 to facilitate ssDNA annealing whereas gp19 can be recruited by gp18 to perform the final cleavage. In support of this scenario, gp17 was found to be still present at high amount at the late stage of SIRV2 life.
cycle, together with the tail-fiber protein of SIRV2 virions (11). Thus, it is very likely that the gene operon is involved in genome maturation of SIRV2 replicative intermediates.

Another common and interesting feature shared by the rudiviruses and filamentous viruses is the presence of multiple 12 bp insertion/deletions (indels) in their genomes, revealed by sequence alignment between closely related viral genomes and between homologous genes (6;44). In the latter case where the nucleotide sequences diverged too much to be aligned, amino acid sequence alignment between homologs allowed the detection of a single or multiple 4 residue indels. Given the proposed function of Cas4 in CRISPR spacer acquisition and the fact that gp19 belongs to the Cas4 nuclease superfamily (26;33), it is possible that gp19 is involved in the generation of the 12 bp indels and the annealing activity of gp18 fits well with both insertion and deletion scenarios. Following strand-displacement replication as proposed for both AFV1 (45) and SIRV2 (19) unpublished data from Martinez-Alvarez et al.), ssDNA bubbles may arise frequently and spontaneously during genome maturation. Repair of such structures involving ssDNA binding protein (gp17), annealing protein (gp18) and ssDNA nuclease (gp19) could in principle produce either insertions or deletions.

A third possible function of the gene operon is recombination involved in general repair or replication initiation, which has been proved important for many viruses (e.g. T4 as in (46). After dsDNA unwinding by a helicase, which remains to be identified in this case, ssDNA nuclease, binding and annealing activities are all needed in the classical recombination processes (47) and the identified functions of the three proteins fit well with the scenario. In support of this, Phyre2 structural modelling of SSO2277, a Sulfolobus protein interacting with gp17 (Fig. 7D) and annotated as hypothetical, revealed a good match (99.9% confidence over half of the protein) with proteins of the family RecF, RecN, Rad50 etc (data not shown). The latter proteins are involved in recombination (48).

In conclusion, this is the first study providing the functional characterization of an entire gene operon conserved in archaeal rudiviruses and filamentous viruses. Due to low or no sequence homology with characterized proteins, the majority of archaeal viral genes remain hypothetical. This had hindered the progress of the archaeal virology field. The results from this study will therefore contribute to better understanding of the novel viruses infecting Archaea, the third domain of life. More importantly, the sequence and/or structural divergence of the three proteins from previously characterized ssDNA binding, annealing
and nuclease proteins not only add novelty to, but also provide important information for evolutionary studies of these proteins, which are nearly ubiquitous from bacteria, archaea to eukaryotes including humans.

**FUNDING**

This work was supported by the European Union Frame Work 7 program 265933. Y.G. received a stipend from China Scholarship Council.

**REFERENCES**


**TABLE AND FIGURE LEGENDS**

Table 1. Structures of the DNA substrates used in this study.

Figure 1. Organization of SIRV2 gp17, gp18, gp19 and their homologs in the genomes of archaeal linear viruses. The pattern codes are as follows:

- SIRV2 gp17 homolog;
- SIRV2 gp18 homolog or analog;
- SIRV2 gp19 homolog;
- conserved gene upstream of SIRV2 gp18 homolog in most filamentous viruses.

Figure 2. gp17 binds to ssDNA. (A) gp17 binds to DNA substrates with either a 23-nt 5’-ssDNA flap or Y-shaped double-flaps (23 nt), but not to blunt-ended dsDNA. F, free DNA; C, DNA-protein complex. (B) gp17 shows a high preference towards ssDNA than to dsDNA. The concentration (nM) of gp17 is indicated on the top of the gel.

Figure 3. Mutagenesis of gp17 revealed a U-shaped binding path for ssDNA. (A) The structure of a monomer of gp17 homolog with the conserved positive charged residues labeled in stick model. (B) Gel retardation assays using gp17 WT and mutant proteins and ssDNA. Protein concentrations are labeled on the top of each gel. DNA forms are indicated by a short line (free) or a line covered with a circle (DNA-protein complex). (C) Quantification of the ssDNA binding activity of different gp17 mutants based on the results shown in B. (D)
Binding path of ssDNA on gp17. The residues contributing to ssDNA binding are labeled in sticks.

Figure 4. Characterization of the refolded recombinant gp18. (A) Purification and refolding of gp18 from inclusion bodies. Lane 1, supernatant of E.coli cells expressing gp18; lane 2, pellet of the lysate; lane 3, pellet protein dissolved in 8 M urea; lane 4, supernatant of gp18 after purification, refolding, heating at 70°C for 20 min and centrifugation. (B) Far-UV CD spectrum of refolded gp18. (C) Temperature denaturation of gp18 followed at 220 nm. (D) Gel-filtration chromatographic analysis of the purified and refolded gp18 protein. The main peak and the two shoulders are labeled.

Figure 5. gp18 stimulates annealing of the complementary oligonucleotides. (A) Concentration-dependent enhancement of oligonucleotide annealing by gp18. The 32P-labeled 57-mer oligo-4 (1nM) and the complementary oligo-5 (1.2 nM) were incubated in the absence (lane 1) or presence (lanes 2 to 6) of gp18. gp18 concentrations were indicated on the top of the gel. The presence (lanes 2 to 5) or absence (lane 6) of ATP is also indicated. (B) Time course of gp18-enhanced single-strand annealing. Left panel, annealing without gp18; right panel, annealing in the presence of 200 nM gp18. (C) Quantification of annealed DNA in the absence or presence of gp18. The percentages were calculated based on the intensities of bands in B.

Figure 6. Nuclease activities of SIRV2 gp19. (A) Selective cleavage of DNA substrate with a 5´ ssDNA flap by gp19. (B) Gradual cleavage of ssDNA. (C) Endonuclease activity of gp19. The circular ssDNA of M13mp18 was incubated at 50 °C with or without the addition of 0.5 μM gp19, and the incubation time is given on top of the gel.

Figure 7. Interactions between gp17, gp18 and gp19 and between gp17 and Sulfolobus host proteins. (A) Schematic presentation of GST-tagged gp17 mutants. (B) Pull-down assays by GST affinity chromatography. The purified and refolded His-tagged gp18 (labeled as P for prey) was incubated with GST-tagged gp17 or GST-tagged gp19 on GSH column for 1 h. After washing with PBS buffer, the GSH beads were boiled in the SDS loading buffer and loaded for SDS-PAGE, and the interacting protein was detected by anti-His antibody. GST protein was used as negative control. Positive controls for Western blotting were carried out using the input His-tagged gp18. (C) Interaction between GST-tagged gp17 mutants and His-tagged gp18. Pull-down assays were performed as described in B. (D) Identification of Sulfolobus proteins interacting with His-tagged gp17 overexpressed in sulfolobus sofataricus P2. Three fractions of eluted proteins from the negative control cells containing the empty pEXA2 vector (lanes 1 to 3) and from the gp17 transformant (lanes 4 to 6) were tested. The
identified proteins are indicated at the right side. (E) Western blot hybridization of negative control (Lane1) and gp17 protein elution (Lane 2).

Table S1. Details of the primers used in this study.

Table S2. Sequences of the oligonucleotides used as substrates in this study.

Table S3. The location, gene length and functions of the conserved gene cluster in all linear viruses.

Table S4. Mass spectrometric peptide mapping and sequencing analysis of the two pulled down proteins.

Figure S1. Genome comparison of all the rudiviruses and filamentous viruses using the Mutagen program. Conserved gene clusters are labeled with red square. Homologs are color-coded whereas white rectangles represent ORFs without homologues.

Figure S2. Structure of SIRV1 ORF131_{2-96} (PDB identifier [ID] 2X5T)(24). (A) Dimer structure of SIRV1 ORF131_{2-96} coloured in deep teal and violet purple for the two monomers. (B) Secondary structure elements of the monomer are labeled in different colors. (C) and (D), A surface representation shown on the concave side and convex side of the dimer, indicating the electrostatic potential of the putative binding interface.

Figure S3. (A) Probability of disordered gp17 aa: two different programs (IUpred and PONDR) were used for the prediction, both revealed disorder at the C-terminus. (B) Percentage of helicity of gp17.

Figure S4. Purification of SIRV2 gp17. Protein gp17 expressed and purified to homogeneity from E.coli. Lane 1-4, four elution fractions from Ni-NTA-agarose beads.

Figure S5. Alignment of SIRV2 gp17 and its homologs. Identical residues are labeled as *, and conserved positive charged residues are shaded red. Red arrows indicate the position of the β sheets, blue bars indicate the position of α helices. Residues mutated to alanine in this study are marked with black dots and numbered accordingly.

Figure S6. Gel retardation assays showing the binding of gp17 WT and some of the mutant proteins to ssDNA.

Figure S7. Standard linear regression curve. The column Superdex 200 HR 10/30 was calibrated with proteins of known molecular masses: Thyroglobulin, Bovine (669 kDa);
Apo ferritin, Horse Spleen (443 kDa); β-Amylase, Sweet Potato (200 kDa) and Alcohol Dehydrogenase, Yeast (150 kDa).
### Table 1. Structures of the substrates used in this study

<table>
<thead>
<tr>
<th>Name</th>
<th>Structure</th>
<th>Oligonucleotides*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substrate A: Blunt end duplex</td>
<td>📊</td>
<td>1+2</td>
</tr>
<tr>
<td>Substrate B: 5′-ssDNA flap duplex</td>
<td>📊</td>
<td>1+3</td>
</tr>
<tr>
<td>Substrate C: 3′-ssDNA flap duplex</td>
<td>📊</td>
<td>1+4</td>
</tr>
<tr>
<td>Substrate D: Y-shaped duplex</td>
<td>📊</td>
<td>4+5</td>
</tr>
</tbody>
</table>

*: The sequences of the oligonucleotides are provided in Table S2
Figure 1

SIRV1,2 << ...... >>
ARV << ...... >>
SRV << ...... >>
AFV1 << ...... >>
SIFV << ...... >>
AFV3-8 << ...... >>
AFV9 << ...... >>
Figure 2

A

B

gp17 (nM) 0 33 66 130 260

gp17 (nM) 0 33 66 130 260

gp17 (nM) 0 33 66 130 260

- C2

- C1

- F

- F

- C

- F

68
Figure 3

A

B

gp17 wt (nM)

0 16 33 66 133 266 400 533 666 1200

B

gp17 K82A (nM)

0 16 33 66 133 266 400 533 666 1200

B

gp17 R60A (nM)

0 16 33 66 133 266 400 533 666 1200

B

gp17 K61A (nM)

0 33 66 133 266 400 533 800 1200 2000

B

gp17 H54A (nM)

0 33 66 133 266 400 533 800 1200 2000

B

gp17 R33A (nM)

0 133 266 400 533 733 933 1200 1466 2000

B

gp17 R60A K61A (nM)

0 133 266 400 533 733 933 1200 1466 2000

C

Bound ratio %

0 20 40 60 80 100

Protein (nM)

D

R33

H54

K61

R60
Figure 5

A

<table>
<thead>
<tr>
<th></th>
<th>ATP</th>
<th>Time (min)</th>
<th>gp18 (nM)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>15</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>15</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>15</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>15</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>15</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>15</td>
<td>200</td>
</tr>
</tbody>
</table>

B

<table>
<thead>
<tr>
<th>gp18 (nM)</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>200</td>
<td>0</td>
</tr>
<tr>
<td>200</td>
<td>0.5</td>
</tr>
<tr>
<td>200</td>
<td>1</td>
</tr>
<tr>
<td>200</td>
<td>2</td>
</tr>
<tr>
<td>200</td>
<td>4</td>
</tr>
</tbody>
</table>

C

Graph showing the annealing percentage over time with different ATP conditions and gp18 concentrations.
Figure 6

A

<table>
<thead>
<tr>
<th>Time (min)</th>
<th>gp19 (0.5 µM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
</tr>
<tr>
<td>10</td>
<td>+</td>
</tr>
</tbody>
</table>

B

<table>
<thead>
<tr>
<th>Time (min)</th>
<th>gp19 (0.5 µM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>5</td>
<td>+</td>
</tr>
<tr>
<td>8</td>
<td>+</td>
</tr>
<tr>
<td>10</td>
<td>+</td>
</tr>
</tbody>
</table>

C

<table>
<thead>
<tr>
<th>Time (min)</th>
<th>gp19 (0.5 µM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>M</td>
</tr>
<tr>
<td>15</td>
<td>+</td>
</tr>
<tr>
<td>30</td>
<td>+</td>
</tr>
<tr>
<td>45</td>
<td>+</td>
</tr>
<tr>
<td>60</td>
<td>+</td>
</tr>
</tbody>
</table>
Figure 7

A

GSTgp17-Ⅰ (131 aa)
GSTgp17-Ⅱ (121 aa)
GSTgp17-Ⅲ (111 aa)

B

Input
GST+P
GSTgp17-Ⅰ+P
GSTgp19+P

C

M
Input
GST
GST+P
GSTgp17-Ⅰ
GSTgp17-Ⅱ+P
GSTgp17-Ⅲ

D

pEXA2-control
pEXA2-gp17

M 1 2 3 4 5 6
170kDa
130kDa
70kDa
55kDa
25kDa
15kDa

Reverse gyrase
Sso2277
gp17

E

1 2
gp17
### Table S1. Details of the primers used in this study

<table>
<thead>
<tr>
<th>Expression Host</th>
<th>Tag</th>
<th>Protein Name</th>
<th>Oligo Nr.</th>
<th>Sequence 5’-3’</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.coli</td>
<td>C-HIS tag</td>
<td>gp17 wt</td>
<td>1</td>
<td>Fw: GGCAGAAACCATATGGCCTCATTAAAACAATATAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gp17(1-121aa)b</td>
<td>3</td>
<td>Rv: CGTACTCGAGTTATTTTTCCTCGATTCTTCTT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gp17 K82Ab</td>
<td>4</td>
<td>1Rv: GCCATGGCCTCTAGAAATTCAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gp17 K32Ab</td>
<td>6</td>
<td>1Fw: GAATTCTAGAGGCGTATGAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gp17 R60Ab</td>
<td>8</td>
<td>1Rv: CCAATTGTTTCCGGAAATATTCT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gp17 K61Ab</td>
<td>10</td>
<td>1Rv: CCAATTGCGCTCTGAAATTATT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gp17 H54Ab</td>
<td>12</td>
<td>1Fw: CATGAGGTAATCGAGTCAGAA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gp17 R33Ab</td>
<td>14</td>
<td>1Rv: GCGATAATGTTGACATTTAAAATG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gp17 R60A K61Ab</td>
<td>16</td>
<td>1Rv: GCCGAAATGGGAATTAAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gp17 R24A K27Ab</td>
<td>18</td>
<td>1Fw: GCCGATTATTCGATAAAAAGGTATAAAAGGAGATTTGAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gp17 R24A K27A K29Ab</td>
<td>20</td>
<td>1Rv: TCTTTTTATACCCGCTATCGCAAAATCGC AGACGC</td>
</tr>
<tr>
<td></td>
<td>N-GST tag</td>
<td>gp18</td>
<td>22</td>
<td>Fw: CATTTGTTCCCATATGAGTGAAACCAACAACTATTG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gp19</td>
<td>23</td>
<td>Rv: CGTACTCGAGCCATCTCTTAAATGTCTAAATAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSTgp17-I</td>
<td>26</td>
<td>Fw: ATCGGGATCCGCTATTAAAACAATATAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSTgp17-IIc</td>
<td>28</td>
<td>Rv: CGTACTCGAGTTATTTTTCCTCCTTCTT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSTgp17-IIIc</td>
<td>29</td>
<td>Rv: CGTACTCGAGTTATTTTTCCTCCTTCTT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSTgp19</td>
<td>30</td>
<td>Fw: GCCGAGTCCGAATATGAGATTGAGATCTATAAAGAGAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>31</td>
<td>Rv: GCCGCTCGAGTTAAAAAGTGATATAATGCATTTTGGTTT</td>
</tr>
<tr>
<td></td>
<td>C-HIS tag</td>
<td>gp17</td>
<td>34</td>
<td>Fw: CATTTGTTCCCATATGAGTGAAACCAACAACCTATTG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gp18</td>
<td>36</td>
<td>Rv: TATTTTATAGCGGCGCCCAACTCTCCTCCTCAGTTT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gp19</td>
<td>38</td>
<td>Fw: CATTTGTTCCCATATGAGTTAAATATGAGAGATC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSTgp19</td>
<td>39</td>
<td>Rv: CTCTTCTACCGGCCCTTTAAAAGGTGATATAATGCATTTT</td>
</tr>
</tbody>
</table>

a The forward primer used for this construction is oligo-1.
b For this construction, the front part fragment was amplified using oligo1 and 1Rv primer, and the rest part fragment was amplified using 1Fw and oligo2. Then the two fragments were used as template to amplify the whole construction by oligo-1 and oligo-2.
c The forward primer used for this construction is oligo-26.
Table S2. Sequences of the oligonucleotides used as substrates in this study

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence (5’ to 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>oligo-1</td>
<td>CGTACTCGAGTTATTGTCCATGTCTAGCTCTTC</td>
</tr>
<tr>
<td>oligo-2</td>
<td>GAAGAGCTAGACATGGAACAAATAACTCGAGTACG</td>
</tr>
<tr>
<td>oligo-3</td>
<td>GTTATTGCATGAAAGCCCGCTGGAAGAGCTAGACATGGAACAATAACTCGAGTACG</td>
</tr>
<tr>
<td>oligo-4</td>
<td>GTCAGTCCAAAAGTACATTATTCGCTAAGCTCAGGTTATTGTCCATGTCTAGCTCTTC</td>
</tr>
<tr>
<td>oligo-5</td>
<td>GAAGAGCTAGACATGGAACAAATAACTCGAGTACGTTATTGCATGAAAGCCCGCTG</td>
</tr>
</tbody>
</table>
**Table S3.** The location, gene length and functions of the conserved gene cluster in all linear viruses

<table>
<thead>
<tr>
<th>Strain</th>
<th>Gene1</th>
<th>Length (aa)</th>
<th>Function</th>
<th>Gene2</th>
<th>Length (aa)</th>
<th>Function</th>
<th>Gene3</th>
<th>Length (aa)</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFV3</td>
<td>gp10</td>
<td>79</td>
<td>Hypothetical protein</td>
<td>gp09</td>
<td>593</td>
<td>Putative helicase</td>
<td>gp08</td>
<td>203</td>
<td>Putative nuclease</td>
</tr>
<tr>
<td>AFV6</td>
<td>gp10</td>
<td>79</td>
<td>Hypothetical protein</td>
<td>gp09</td>
<td>593</td>
<td>Putative helicase</td>
<td>gp08</td>
<td>203</td>
<td>Putative nuclease</td>
</tr>
<tr>
<td>AFV7</td>
<td>gp05</td>
<td>79</td>
<td>Hypothetical protein</td>
<td>gp04</td>
<td>593</td>
<td>Putative helicase</td>
<td>gp03</td>
<td>203</td>
<td>Putative nuclease</td>
</tr>
<tr>
<td>AFV8</td>
<td>gp07</td>
<td>79</td>
<td>Hypothetical protein</td>
<td>gp06</td>
<td>593</td>
<td>Putative helicase</td>
<td>gp05</td>
<td>203</td>
<td>Putative nuclease</td>
</tr>
<tr>
<td>AFV9</td>
<td>gp11</td>
<td>79</td>
<td>Hypothetical protein</td>
<td>gp10</td>
<td>602</td>
<td>Putative Holiday junction branch migration helicase</td>
<td>gp08</td>
<td>203</td>
<td>Putative nuclease</td>
</tr>
<tr>
<td>SIFV</td>
<td>SIFV-08</td>
<td>79</td>
<td>Hypothetical protein</td>
<td>SIFV-07</td>
<td>601</td>
<td>Putative helicase</td>
<td>SIFV-06</td>
<td>232</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>AFV2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>gp15</td>
<td>425</td>
<td>Hypothetical protein</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AFV1</td>
<td>gp14</td>
<td>135</td>
<td>Hypothetical protein</td>
<td>gp15</td>
<td>426</td>
<td>Hypothetical protein</td>
<td>gp17</td>
<td>223</td>
<td>CRISPR-associated Cas4-like protein</td>
</tr>
<tr>
<td>ARV1</td>
<td>gp12</td>
<td>134</td>
<td>Hypothetical protein</td>
<td>gp16</td>
<td>443</td>
<td>Hypothetical protein</td>
<td>gp17</td>
<td>207</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SRV</td>
<td>SRV-ORF138</td>
<td>138</td>
<td>Hypothetical protein</td>
<td>SRV-ORF440</td>
<td>440</td>
<td>Hypothetical protein</td>
<td>SRV-ORF199</td>
<td>199</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SIRV2</td>
<td>gp17</td>
<td>131</td>
<td>Hypothetical protein</td>
<td>gp18</td>
<td>436</td>
<td>Hypothetical protein</td>
<td>gp19</td>
<td>207</td>
<td>Single strand nuclease</td>
</tr>
</tbody>
</table>
Table S4. Mass spectrometric peptide mapping and sequencing analysis of the two pulled down proteins

<table>
<thead>
<tr>
<th>Sample name</th>
<th>Protein found in database</th>
<th>GI-number</th>
<th>MW</th>
<th>Score</th>
<th>Seq. cov.</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>yang1</td>
<td>hypothetical protein SS02277 [Sulfobolus solfataricus P2]</td>
<td>gi</td>
<td>15899044</td>
<td>57127</td>
<td>372</td>
<td>45%</td>
</tr>
<tr>
<td>yang2</td>
<td>reverse gyrase [Sulfobolus solfataricus P2]</td>
<td>gi</td>
<td>15897352</td>
<td>142207</td>
<td>550</td>
<td>38%</td>
</tr>
</tbody>
</table>

**Quality control standards included in the analysis**

<table>
<thead>
<tr>
<th>Standard</th>
<th>Protein found in database</th>
<th>GI-number</th>
<th>MW</th>
<th>Score</th>
<th>Seq. cov.</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 pmol transferrin</td>
<td>transferrin [Homo sapiens]</td>
<td>gi</td>
<td>115394517</td>
<td>79190</td>
<td>886</td>
<td>61%</td>
</tr>
<tr>
<td>62 fmol BSA</td>
<td>Chain A, Crystal Structure Of Bovine Serum Albumin</td>
<td>gi</td>
<td>367460260</td>
<td>58416</td>
<td>363</td>
<td>37%</td>
</tr>
</tbody>
</table>
Figure S1
Figure S2
Figure S3

A

Disorder

Residue

iUpred

PCNDR

B

% helicity

Residue
Figure S4

![Image of a gel showing protein bands at 25KDa, 15KDa, and 10KDa, with an arrow pointing to a band labeled SIRV2 gp17.]
Figure S6

<table>
<thead>
<tr>
<th>gp17 wt</th>
<th>gp17(R24A K27A)</th>
<th>gp17(R24A K27A K29A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(nM)</td>
<td>(nM)</td>
<td>(nM)</td>
</tr>
<tr>
<td>0  33  66  130  260</td>
<td>0  33  66  130  260</td>
<td>0  33  66  130  260</td>
</tr>
</tbody>
</table>

gp17(1-121aa)

(nM) 0  33  66  130  260
Figure S7

\[ y = -0.2976x + 1.8388 \]
\[ R^2 = 0.9814 \]
Manuscript II

Genome-wide binding profile of two transcription regulators of Sulfolobus solfataricus

Yang Guo, Xu Peng
In preparation
Genome-wide binding profile of two transcription regulators of
*Sulfolobus solfataricus*

Yang Guo and Xu Peng *

Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, 2200 CPH N. Denmark

* To whom correspondence should be addressed. Tel: +45 35322018; Fax: +45 35322128; Email: peng@bio.ku.dk
**Abstract**

Two transcription regulators sso2474 and sso10340 from *Sulfolobus solfataricus* P2 were differently regulated upon SIRV2 infection. A method similar as Chromatin immunoprecipitation combined with subsequent high-throughput sequencing (Chip-seq) was applied in this study to get into the gene composition of the two protein regulons *in vivo*. Mapping of the sequencing data with *Sulfolobus solfataricus* P2 and SIRV2 genomes demonstrated that sso2474 binds with a high affinity to virus genome, whereas sso10340 mainly binds to the host DNA. A total of 27 enriched host DNA fragments extracted from sso10340-DNA complex appeared as potential binding targets, most of which are genes involved in energy metabolism, transport, translation and amino acid metabolism. The genome-wide binding profiles presented here reveal two different kinds of regulon conditions and contribute to the knowledge expansion of the transcription regulations upon virus infection.
Introduction

Viruses infecting the organisms of Archaea, the third domain of life, comprise the most diverse and previously unsuspected virion morphotypes (Pina et al., 2011). In the last few years, a substantial effort was made to explore the functions of the hypothetical proteins and the viral infection life cycles (Kessler et al., 2004; Oke et al., 2011; Gardner et al., 2011; Bize et al., 2009). To date, several virus-host systems have become promising models providing a great opportunity for studying virus-host interactions.

One of the best-studied viruses in hyperthermophilic archaea is SIRV2 (*Sulfolobus islandicus* rod-shaped virus 2), isolated from an acidic hot spring in Iceland, belongs to *Rudiviridae* and share a common ancestry with the family *Lipothrixviridae*. Studies by transmission electron microscopy showed that SIRV2 virions specifically recognized the pilus-like filaments on the host cell surface to get adsorption (Quemin et al., 2013). On the other hand, two gene clusters, cluster sso3138 to sso3141 and cluster sso2386 and sso2387 identified from the SIRV2 resistant *Sulfolobus* mutants were confirmed responsible for the virus entry (Deng et al., 2014), providing first insights into its entry process. Unlike most archaeal viruses, infecting host cells with a `carrier state`, this linear non-enveloped double-stranded DNA (dsDNA) SIRV2, together with TTV1 and STIV, are lytic viruses (Bize et al., 2009; Ortmann et al., 2008; Zillig et al., 1996). The virions released from the host cell through a unique mechanism, which involves the formation of pyramid-like protrusions, transecting the cell envelope and S-layer. At the end of the infection stage, this seven isosceles triangular faces pyramid opens up, allowing mature virions to escape from the cell (Bize et al., 2009; Quax et al., 2011). To gain better insights into the biology of virus, life cycle and their effect on the host, microarray analysis to determine the transcriptional responses of the host and the virus during the infection process could be very efficient and had been successfully applied to three archaeal viruses, the fusellovirus SSV1, the icosahedral virus STIV and the *Rudiviridae* virus SIRV2 (Frols et al., 2007; Ortmann et al., 2008; Okutan et al., 2013).

What we focused in this work is to investigate the host genes regulation upon SIRV2 infection. As the previous study revealed that a total of 148 host genes differently responded, and among these genes, two transcription regulators sso2474 and sso10340 were up and
down regulated, respectively. It is raised an interesting question as to how these two proteins regulate the corresponding genes upon the virus infection stress. Are they global regulators or just regulate their own promoters? A method similar with chip-seq was applied to this study for detection of the DNA binding sites in vivo. Combined with the gene expression analysis, we can get a first insight into the transcription regulation network between virus and host cells.

Materials and methods

Sulfolobus cultivation and plasmid construction

sso2474 and sso10340 fragments were amplified from Sulfolobus solfataricus P2 genome by PCR, digested with NdeI and NotI and inserted into the similarly digested sulfolobus/E.coli shuttle vector pEXA3 (He et al., 2014), allowing the expression of Histagged gp17 under the control of arabinose promoter. The constructed plasmid, as well as the empty plasmid pEXA3, were then electroporated individually into the uracil deficient competent cells (Deng et al., 2009). Single colonies of the transformants were inoculated into test tubes containing 5 ml SCV (basal medium supplemented with 0.2% sucrose, 0.2% casamino acids and 1% vitamin solution) (Deng et al., 2009), and incubated in an Innova 3100 oil-bath shaker. Large-scale culturing was performed in ACV medium (0.2% D-arabinose was substituted for sucrose) with Erlenmeyer flasks of long necks. When the culture OD$_{600}$ reached to 0.8, it was infected by SIRV2 at about m.o.i of 10. The cells were collected after 2.5 h virus post infection.

E. coli cells cultivation and plasmid construction

The coding sequences of sso2474 was amplified by PCR from Sulfolobus solfataricus P2 genome, digested with NdeI and XhoI and subsequently inserted into a similarly digested pET-30(a) (Novagen) expression vector. E.coli BL21 CodonPlus cells were transformed with individual plasmid construct and a single clone transformant was inoculated in LB medium containing 30 µg/ml kanamycin and 25 µg/ml chloramphenicol. At an optical density (OD$_{600}$) of 0.4, IPTG (0.5 mM) was added to the culture and the cells were further cultured at 25°C for 12 hours.

Protein purification

The purification of His-tagged sso2474 and sso10340 either from sulfolobus or E.coli was carried out as follows, the harvested cell pellets were lysed in lysis buffer (50 mM Tris-Hcl pH 8.0 , 300 mM NaCl , 1 mM EDTA, 1% Triton X100 and 1 mM PMSF) by sonication, and different sonication time (4,6,8,10 min) was detected to minimum the size of the DNA
fragments bound by the proteins. Then the lysate were cleared by centrifugation at 10000× g for 20 min. Supernatant was then incubated with Ni-NTA-agarose beads (Qiagen, Germany) for 1 h at room temperature. Beads was washed three times with washing buffer (50 mM Tris-Hcl pH 8.0, 300 mM NaCl, 40 mM Imidazol) and protein-DNA complex were eluted with elution buffer (50 mM Tris-Hcl pH 8.0, 300 mM NaCl, 250 mM Imidazol). The purity of the protein was evaluated by 12.5 % SDS-PAGE and staining with PAGE blue (Sigma Aldrich, UK), and the amount of DNA in the samples were detected on 0.7% Agarose gel and stained with GelRed (Biotium).

DNA extraction and high-throughput sequencing

The eluted protein-DNA complex in solution were diluted with one volume of water and treated with RNase A at room temperature for 1h. The deproteination was carried out by incubation with 2 mg/ml Protease K at 50 °C for 2 h, and 65 °C for 8 h. Then the DNA was extracted with phenol/chloroform/isoamyl (25:24:1) mixture solution, and finally was precipitated and concentrated by ethanol precipitation. Sequencing libraries with an average fragment size of 350 bp were prepared according to protocol of the ion plus fragment library kit, and sequenced in the Ion PGM™ Sequencer (Life Technology).

Reads mapping and Peak detection

The quality filtered reads were treated and aligned to genome Sulfolobus solfataricus P2 as well as SIRV2 using Bowtie, the ultrafast memory-efficient short read aligner, to align sequenced sets of short DNA reads to large genomes (Satoh and Tabunoki, 2013), and then the enriched peaks were visualized using Artemis (Carver T, etal. 2012), allowing for up to two errors per reads (insertion, deletion and/or mismatch).

Real-time Quantitative PCR

qPCR reactions were performed in 10 μL mixtures containing 5 μL iQ SYBR Green Supermix (Bio-Rad, Cat. No. 170-8880), 1 mM primers and around 1 ng total DNA. Separate reactions were prepared for detection of reference gene and virus specific and Sulfolobus solfataricus-host specific amplicons. The mixtures were prepared in duplicates in 96-well microliter PCR plates (Bio-Rad Laboratories), sealed with an adhesive cover (Bio-Rad Laboratories) and worked on the CFX96 Real-Time Detection System (Bio-Rad Laboratories) following this uniform cycling parameters: Initialization (95°C for 3 min) was followed by the denaturation of the strands (95°C for 10 sec), annealing of the primers to the template (55°C for 10 sec), elongation of the primers by the DNA polymerase (72°C for 15 sec). The cycle from denaturation to elongation were repeated 40 times. Thereafter, the final elongation step was performed at 95°C for 10 sec. At last, a melting temperature gradient
with 0.5°C increasing increment from 65–95°C for 5 sec was used to confirm the specificity of the primer sets. Besides, the possibility of unspecific amplification products and contamination was checked by using a non-template control (NTC). Furthermore, a positive control was used with a known amount of template. qPCR data were analyzed with the Bio-Rad CFX manager software, which allows for the immediate determination of the cycle threshold (Ct), melting curves and quantification of samples.

**Motif analysis**

For de novo motif discovery within the significantly enriched DNA fragments, their genomic sequences were submitted to MEME (Bailey and Elkan, 1994). Parameters were set to search for zero or one palindromic motif of 16 bp width per sequence.

**DNA band shift assays**

50 nM of the ssDNA (oligo1-CGTACTCGAGTTATTGTCCATGTCTAGCTCTTC) and blunt-ended dsDNA (which was formed by annealing the oligo1 with its complementary oligonucleotide) were incubated for 20 min at 50°C with increasing concentrations of SSO2474 (0-2.0 μM) in 20 μl DNA-binding buffer (10 mM Tris-Cl, pH 8.0, 100 mM KCl, 2 mM DTT, 10% glycerol). The samples were loaded onto 12% (v/v) acrylamide gel and electrophoresed in 0.5 × TBE buffer for 1 h 50 min. Following electrophoresis, the gels were stained with SYBR® Gold (Life Technologies) and scanned by Typhoon FLA 7000 (GE Healthcare Life Science).

**Results**

**Function prediction and high conservation of the two host transcription regulators**

The transcription machinery in Archaea has drawn a lot of interests due to its bacterial-like regulators and eukayote-like basal factors (Bell et al., 2001). Although possessing unique genome structure, half of the transcription factors (TFs) identified in archaeal genomes share at least one homolog with bacterial genomes (Perez-Rueda and Janga, 2010). Therefore, structural or sequence similarities to the well-studied protiens could provide efficient and direct information for the study of the unknown archaeal proteins.

BlastP search of sso2474 revealed a putative HTH (Helix-turn-Helix) motif located between 20 and 70 amino acid residues, and a conserved domain matching MarR-2 (multiple antibiotic resistance) family proteins was also detected at the same location. The genome sequence similarities in the database suggest that sso2474 belong to the TrmB-like (transcriptional regulator of the maltose system) family, and it has a 36% identity and 65% sequence similarity to transcription regulator TrmB in *Sulfolobus archaeon AZ1* and a 36%
identity and 61% similarity to transcription regulator TrmB in strain Acidianus hospitalis. Most of the TrmB proteins are able to bind DNA using a HTH motif as DBD (DNA–binding domain). The DBD is located at the N-terminal region and a mutational analysis revealed that it is essential for binding (Maruyama et al., 2011). However, the tertiary structure prediction of sso2474 using the threading program Phyre 2 (Kelley and Sternberg, 2009) suggested a high confidence (99.6%) to MarR-like family transcription regulator with a 93% alignment coverage. The crystal structures of many MarR family proteins from bacterial and archaeal species were solved, and they reveal a common architecture with a characteristic winged helix domain for DNA binding. Although sequence identities between these homologs is less than 20%, they all possess the same core fold (Nichols et al., 2009). The protein sso2474 was conserved in Sulfolobus, Acidianus and Metallophaera species of Sulfolobales as well as euryarchaeotal Halobacteria species (Fig. S1).

The down-regulated gene, sso10340, encoding a 10 kDa protein, has a high identity with truncated variant of Lrp/AsnC-family. Sequence alignment between sso10340 and E.coli Lrp (leucine-responsive regulatory) protein, one of the best characterized Lrp family proteins, revealed that the sso10340 protein aa sequence matched well with the C-terminal amino acid effector domain of the Lrp protein (Fig. S3). The structure prediction of sso10340 by Phyre2 also suggested a high confidence (99.9%) to the STS042 protein from Sulfolobus tokodaii 7. STS042 was identified as a stand-alone RAM (regulation of amino acid metabolism) module protein, which has homologies with the C-terminal domain of Lrp/AsnC-family proteins (Miyazono et al., 2008). Search results among the DNA database indicated that Lrp/AsnC-family proteins distribute among many bacterial and most archaea. Sso10340 has homologues in crenarchaeotal sulfolobales species as well as bacterial species (Fig. S2).

**DNA extraction from protein-DNA complex and high-throughput sequencing**

A method similar to Chip-seq was used to gain further insights into the two proteins regulon. sso2474 and sso10340 were cloned into the E.coli-Sulfolobus shuttle vector pEXA2 under the control of arabinose promoter, with a His-tag in the C-terminus, respectively (Gudbergsdottir et al., 2011). In order to detect their binding sites in both the host and virus genome, the cells were infected with SIRV2 at a m.o.i of 10 after the expression of the target protein was induced for 15 h. The cells were collected after 2.5 h post virus infection.

By Ni-NTA-Agarose chromatography the protein-DNA complex were purified. Proteins were detected in SDS-PAGE gel and DNA bound by these proteins was run on the agarose gel. As shown in Fig. 1A, sso2474 was purified to homogeneity, with a single band detected in SDS-gel. The DNA extracted from sso2474 exhibited hundreds of folds higher yields
than the control DNA, which was purified by Ni-NTA-Agarose beads from the cells transformed with an empty plasmid. It seems that sso2474 showed a really high affinity to DNA. The SDS-PAGE and western blot analysis showed that the Lrp-like protein, sso10340, exhibit a range of oligomeric states including dimers, octamers and decamers even after SDS treatment (Fig. 1B), resembling the Lrp/AsnC family proteins which form a range of multimeric species in solution (Brinkman et al., 2003; Leonard et al., 2001). There are also significantly more DNA from sso10340-DNA complex than from the control.

The purified DNA-protein complex was firstly treated with RNase A to remove the contaminated RNA, and the deproteination was carried out by incubation with protease K. The target DNA fragments of each sample were finally extracted using phenol/chloroform extraction and ethanol precipitation, with an average size of 300-500 bp. Then the prepared sample was sequenced using ion torrent next-generation sequencing. Of the sequenced 363-393 thousand reads, 347-362 thousand reads was uniquely mapped with either *sulfolobus solfataricus* host genome DNA or SIRV2 virus DNA (Table 1). It is interesting that 91.7 % of the mapped reads from sso2474 are aligned with virus genome, whereas 92.04 % reads from sso10340 belong to the host genome, indicating that sso2474 has a high affinity to virus genome and sso10340 specially regulate the host genes.

It is surprising that almost all the DNA extracted from sso2474 was aligned to virus genome. In order to validate whether it is due to the high amount of virus genome present in the cell, we checked the copy number ratio between host and virus genome by Real-Time PCR (qPCR). The infected cells (the same one for sequence) were collected and washed 3 times to remove the virus on the cell surface, and the total DNA from the infected *Sulfolobus* cells was extracted. One set of primers belong to the *Sulfolobus solfataricus* TFB-II were designed to detect the host genome copy numbers and the primers amplifying the SIRV2 coat protein were designed to check the virus copy numbers in the same DNA sample. Whereas, the data in Table 2 showed that there was average 0.6 virus entered in one host cell after 2.5 h post infection, excluding the possibility that the high coverage of viral sequence reads was due to a high copy number of the virus present in the infected cells. Thus, we conclude the sso2474 preferentially binds the viral DNA.

**Detection of the enriched DNA fragment and genome-wide binding profile of the two proteins**

To identify the DNA-enriched regions, we use Bowtie, the ultrafast memory-efficient short read aligner, to align sequenced sets of short DNA reads to large genome *Sulfolobus solfataricus* P2 as well as SIRV2 (Satoh and Tabunoki, 2013). And the genomic locations of
the peaks was identified and visualized by Artemis, an integrated platform to analyze high-throughput sequence-based experimental data (Carver et al., 2012).

**Protein Sso2474 binds to virus DNA with low specificity**

Although only 3.76 % DNA extracted from protein sso2474 can be aligned to the host genome, there is still a specific binding peak showing up in the map (Fig. 2A), located upstream and inside of sso2474, indicating that the protein was regulated by itself (peak 10 in Fig 2A). In contrast, an average of 5400 reads were aligned to SIRV2 genome, which was more than 500 folds than that to the host genome. They were demonstrated as mountain shape with wide peaks covering the whole virus genome (Fig 2B). Even so, some potential specific binding sites were marked with numbers, the binding regions were amplified by PCR, and the gel mobility shift assays were carried out for validation.

Protein sso2474 was firstly expressed and purified from *Sulfolobus solfataricus* P2 (Fig 1A). However, DNA specifically bound by the protein cannot be removed by either DNaseI or PEI (Phenylethyleneimine) and formed an extremely high background. Then we set out to express the sso2474 protein in *E.coli* inserted into the pET-30a vector with a C-terminal His tag and purified by Ni^{2+}-affinity chromatography. It was expressed soluble in high amount and SDS-PAGE analysis of the purified protein revealed a pure major band with a molecular weight of approximately 15kDa (Fig. 3A).

Firstly the 11 enriched fragments and a negative control fragment were amplified by PCR from SIRV2 genome, and the electrophoretic mobility shift assay (EMSA) of the recombinant sso2474 from *E.coli* with the target fragment were carried out. The results showed that this protein bound all the DNA fragments with no specificity (data not shown). Since the binding region of sso2474 cannot be detected, another possibility is that this protein prefers to bind ssDNA rather than dsDNA, some single-strand DNA binding proteins bind DNA in a non-sequence specific way (Dickey et al., 2013). To verify whether sso2474 is a single-strand binding protein, an EMSA experiment with equal molar ratio of ssDNA and dsDNA substrate mixture were performed.

The concentration of the protein was increased from 0.1 μM to 2.0 μM. The samples were deposited in a 12% acrylamide gel and was run in 0.5 x TBE buffer. It is demonstrated that sso2474 preferred to bind dsDNA than ssDNA. As it is shown in Fig. 3B lane 3, the band representing dsDNA began to shift while the amount of ssDNA kept the same. When the protein binds all dsDNA in the sample, and there is still more protein left, it begins to bind the ssDNA. When the protein concentration increased to 1.6 μM (Fig 3B, Lane 7), almost all the substrates formed complex and no free DNA left. The result demonstrated a clear image that sso2474 showed more fold affinity to dsDNA than to ssDNA.
Protein sso10340 bound the host genome at several regions

Compared with sso2474, the genome-wide binding profile of sso10340 with host genome was well mapped showing a dozen of binding sites (Fig 4 A). However, the reads corresponding to viral sequences were randomly aligned with virus genome, similar to the control sample (Fig 4 B). Only regions exhibiting more than 2-fold enrichment in CHIP DNA versus input DNA were considered to be bound significantly to sso10340. A total of 27 genomic regions, scattered in the genome, was identified. The various functions of the genes that these binding peaks overlapped or closest to were summarized, most of which participate in amino acid metabolism, energy metabolism, biosynthesis and transport (Table 3).

Additionally, the fragments were grouped into four categories according to their location with respect to open reading frames (Fig 4 C). As we observed that 41% (upstream and intragenic but upstream) of the 27 regions localize to the upstream regulatory region of the corresponding gene, and 46% of them fell within the coding region. The small left peaks (13%) were found in the regions locating both the downstream of the neighbored genes.

As half of the binding region fell into the upstream area of the corresponding gene, a binding profile of 14 genomic regions near promoter area were zoomed in and analyzed in detail (Fig. S4). The binding genomic fragment was amplified by PCR with an average size of 150bp. The protein sso10340 was purified from Sulfolobus, and an EMSA screen of these regions was performed to verify whether these targets regions also interact with purified protein in vitro. However, no binding was observed by protein sso10340 in vitro.

Motif analysis for sso10340 binding site

The protein sso10340 binding motif was defined by enriched oligonucleotide sequences within bound regions. The sequences of these DNA fragments were submitted to the motif-based sequence analysis tool MEME-ChIP (Machanick & Bailey, 2011. http://meme.nbcr.net.) to detect conserved DNA motif. The most suggested motif was shown in Fig. 5 B.

This 11 bp, an imperfect palindromic sequence was present in 96% of all binding regions and have a similarity with the known motif of PPARG (Peroxisome proliferator activated receptor gamma) (MA0066.1) (Fig 5 A). PPARgamma binds as heterodimer composing of members of the retinoid X receptor family (RXR) and PPRE (PPAR response elements), which had a direct repeat of two half sites of 5´-AGGTCA-3´ separated by one nucleotide (Fig 5 A).

Discussion

95
The archaeal transcription regulation possessed the eukaryotic-like basal transcription machinery and bacterial-like regulators that distinguished them from the other two domains (Koonin and Galperin, 1997; Grabowski and Kelman, 2003). Bindings of the TBP (TATA-box binding protein) and TFB (transcription factor B) to TATA box and BRE (TFB response element) in the promoter region are critical to the transcription initiation of archaeal genes (Bell et al., 1999). How the bacterial-type transcriptional regulators regulate the eukaryotic-like transcription machinery in archaea, especially on the virus infection, are still need to be elucidated.

The protein sso2474 showed an amino acid sequence similarity with TrmB family proteins, which were found in all three domains of life, containing all three kinds of possible TF combinations- repressors, activators or both. In archaea, most of the TrmB family proteins were spread in the kingdom Euryarchaeota, only some exist in Crenarchaeota (Maruyama et al., 2011). No matter the best studied TrmB proteins in Thermococcales *P. furiosus* of Euryarchaeota or the research on TrmB family protein MalR of *S. acidocaldarius* in Crenarchaeon, most of documented TrmBs seem to function as controlling diverse sugar transporters or different genes of sugar metabolism, such as maltose and glucose processing, as well as genes involved in other metabolisms (Kanai et al., 2007; Lee et al., 2008; Reichlen et al., 2012; Wagner et al., 2014). However, in this study, the binding map of the protein sso2474 indicated that this protein did not show a significant response on regulating the genes related with sugar metabolism (except sso2474 itself) to activate or repress the corresponding gene for its healthy maintenance.

On the other hand, both the conserved domain and the structure prediction revealed that sso2474 belongs to the MarR (multiple antibiotic resistance) family transcription regulators. MarR family proteins constitute a diverse group of transcription regulators that modulate the expression of genes encoding proteins involved in the metabolic pathways, stress responses, virulence and degradation or export of harmful chemicals such as antibiotics, organic solvents (White et al., 1997), oxidative stress agents (Ariza et al., 1994), and house disinfections (McMurry et al., 1998). It seems that this mar locus is involved in the mechanism that the stains used to resist the lethal effects of a wide range of toxic agents.

*E. coli* MarR was the firstly described MarR family regulator and its homologs are widely distributed in both bacterial and archaea. MarR, as a component of the marRAB locus in *E. coli*, is a repressor to its own operon and MarA is a transcription activator that can active the operon and regulates the expression of proteins important to the multiple antibiotic resistance (Alekshun and Levy, 1997). In many strains, constitutive expression of MarA makes a contribution to maintenance of the resistance to antibiotics and other environmental hazards, and the *marR* deletion mutant or the inactivation of MarR will result the increased
expression of MarA (Alekshun and Levy, 1999; Barbosa and Levy, 2000). To date, no research revealed that TrmB family or MarR family proteins bind specifically to virus genome upon virus infection. If sso2474 is more similar to a MarR family protein, the strain would activate the expression of corresponding proteins to resist the exposure to the virus infection, and the sso2474 could be in a way like binding to virus genome to hinder the process of transcription. Based on this hypothesis, the growth retardation to SIRV2 was compared between sso2474 overexpressed strain and the wild-type strain, and no difference was observed from the growth curve (data not show). The above experiment indicated that this protein probably is not involved in inhibiting the growth of virus. Or perhaps there is a difference between viral and host DNA, e.g. modification, so sso2474 could specifically binds to viral DNA. However, the binding mechanism of this protein and its possible interacted partners involved are needed to be further identified. It will be intriguing to detect the phenotype changes of sso2474 mutant strain upon SIRV2 infection, comparing to wild type strain.

The downregulated transcription regulator sso10340 showed a similarity to the C-terminal domain of Lrp/AsnC family proteins (leucine-responsive regulatory protein). Most of the experimentally characterized archaeal transcriptional regulators belong to this family, and it is a family that globally and specifically regulates genes. These family members can be found in both bacteria and archaea but not in eukarya (Brinkman et al., 2003). The Lrp family proteins typically have a 15 kDa molecular weight for the monomer with an N-terminal wHTH domain and a C-terminal Amino Acid Metabolism (RAM) domain. The RAM possesses a αβ sandwich fold and possibly involved in effector recognition and oligomerization of the protein subunits (Thaw et al., 2006). Actually, proteins that only possess the RAM domain are frequently observed in the genomes of many organisms. They are defined as a novel ligand-binding domain or stand-alone RAM-domain (SARD) proteins involved in regulation of amino acid metabolism (Ettema et al., 2002). Although many of them were crystalized and structurally determined, the functions of these proteins remain not clear and still need to be elucidated (Miyazono et al., 2008; Nakano et al., 2006). The failure of detecting any DNA binding by sso10340 is possibly due to lack of binding conditions or lack of DNA binding activity. It is possible that sso10340 recognizes an effector and interacts with a DNA binding protein or a transcription regulator to achieve its regulatory role in vivo. Indeed, sequences close or in the coding region of a DNA binding protein (sso2626) and a transcription regulator (sso2827) were detected in this study (Table 3). Whether sso10304 interacted with the two proteins still need to further confirmed.
Reference


Table and Figure Legends

**Table 1.** Sequencing and mapping data with *Sulfolobus solfataricus* P2 and SIRV2 genomes

**Table 2.** The average virus copy number in each host cell for 2.5 h post infection of SIRV2

**Table 3.** The products as well as their COG Functional Category of target genes binding by sso10340

**Figure 1.** Purification of protein sso2474, sso10340 as well as negative control from *Sulfolobus solfataricus* P2. (A) Purified protein elution samples of sso2474 and negative control was subjected to 12.5% SDS-PAGE gel to detect the purity of the protein (a), and 0.7% Agarose gel to detect the DNA amount bound by protein(b). (B) Purified protein elution samples from sso10340 and negative control was subjected to SDS-PAGE gel to detect the the purity of the protein (a), and Agarose gel to detect the DNA amount(b). Western blot analysis of the purified sso10340 (c).

**Figure 2.** Genome-wide distribution of sso2474 binding regions on *Sulfolobus solfataricus* P2 genome (A) and SIRV2 viron genome (B). These experiments were performed using DNA extracted from purified sso2474 and negative control. Data were analyzed and visualized as ‘Material and Methods’ section. The genome coordinates (in bp) are given on the x-axis, and y-axis represents the sequenced reads aligned on the genome. And the sharp peak marked in arrow in (A) locates in the region of protein sso2474.

**Figure 3.** (A) The purified sso2474 fractions (L1-L6) from *E.coli* were analyzed on a SDS-PAGE gel. The single major band around 15kDa represented the protein sso2474. (B) EMSA assay with dsDNA and ssDNA mixture as substrate. The concentration of the protein was increased from 0.1 μM to 2.0 μM. The length of oligos was 35 bp. The dsDNA and ssDNA were mixed with equal molar ratio (25 nM: 25 nM).

**Figure 4.** Genome-wide distribution of sso10340 binding regions on *Sulfolobus solfataricus* P2 genome (A) and SIRV2 viron genome (B) as well as classification of sso10340 binding regions with respect to genomic organization (C). (A) and (B), These experiments were performed using DNA extracted from purified sso10340 (red bar) and control (empty plasmids) (green bar)-input chip. The genome coordinates (in bp) are given on the x-axis, and y-axis represents the sequenced reads aligned on the genome. Taget gene detected in vivo are indicated. (C) ChIP-enriched regions are indicated by black horizontal bars, whereas ORFs are depicted by horizontal arrows. Binding regions ranging from – 500 bp to +100 bp relative to translation start site was considered to be upstream of a transcription unit or intragenic but upstream. ChIP-enriched regions that are exclusively located in gene coding regions or partial in downstream of the gene were identified as intragenic part. For binding in the regions belong to both downstream of neighbored gene was classified to intergenic region.

**Figure 5.** Weblogo of the motif detected with MEME-chip. For MEME motifs, the discovered motif Logo (B) from submitted bind-region sequences is shown aligned with the most similar JASPAR motif Logo (A) and the similarity is significant (E≤0.05).
Table 1. Sequencing and mapping with *Sulfolobus solfataricus* P2 and SIRV2

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sequenced reads</th>
<th>Mapped reads</th>
<th>Mapped reads with p2 genome</th>
<th>Alignment rate with p2 genome</th>
<th>Mapped reads with Sirv2 genome</th>
<th>Alignment rate with Sirv2 genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input control</td>
<td>393537</td>
<td>362185</td>
<td>359559</td>
<td>91.37%</td>
<td>2626</td>
<td>0.67%</td>
</tr>
<tr>
<td>sso2474</td>
<td>363814</td>
<td>347347</td>
<td>13677</td>
<td>3.76%</td>
<td>333670</td>
<td>91.70%</td>
</tr>
<tr>
<td>sso10340</td>
<td>388533</td>
<td>359145</td>
<td>357598</td>
<td>92.04%</td>
<td>1547</td>
<td>0.4%</td>
</tr>
</tbody>
</table>

Table 2. The average virus copy number in each host cell for 2 h post infection of SIRV2

<table>
<thead>
<tr>
<th>Time (h)</th>
<th>Host SQ Mean</th>
<th>Host Std.Dev</th>
<th>Virus SQ Mean</th>
<th>Virus Std.Dev</th>
<th>Virus copy number/host chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>5.287E+07</td>
<td>1.80E+06</td>
<td>3.158E+07</td>
<td>8.22E+06</td>
<td>0.6</td>
</tr>
</tbody>
</table>
Table 3. The products as well as their COG Functional Category of target genes binding by sso10340

<table>
<thead>
<tr>
<th>Target name</th>
<th>Genomic coordinate peak start</th>
<th>Genomic coordinate peak stop</th>
<th>Motif binding location</th>
<th>Log2 CHIP DNA/input DNA</th>
<th>Gene function</th>
<th>COG Functional Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>doxC</td>
<td>40520</td>
<td>41150</td>
<td>Inside the gene</td>
<td>1.82</td>
<td>Terminal oxidase</td>
<td>Energy production and conversion</td>
</tr>
<tr>
<td>sso1214</td>
<td>1053350</td>
<td>1054300</td>
<td>Inside the gene</td>
<td>3.41</td>
<td>Carbonic anhydrases; Anaerobic dehydrogenases; molybdopterin oxidoreductase molybdopterin-binding protein</td>
<td>Inorganic ion transport and metabolism</td>
</tr>
<tr>
<td>sso1580</td>
<td>1427200</td>
<td>1428050</td>
<td>Inside the gene</td>
<td>2.21</td>
<td>Carboxy-terminal oxidase</td>
<td>Transcription</td>
</tr>
<tr>
<td>sso2826</td>
<td>2587400</td>
<td>2588210</td>
<td>Inside the gene</td>
<td>2.34</td>
<td>Conserved,truncated variant of Lrp/AsnC-family;RNA polymerase and transcription factors</td>
<td>RNA processing and modification</td>
</tr>
<tr>
<td>pacS</td>
<td>2411320</td>
<td>2412430</td>
<td>Inside the gene</td>
<td>3.39</td>
<td>Cation transporting ATPase</td>
<td>Transcription</td>
</tr>
<tr>
<td>sso3189</td>
<td>2935150</td>
<td>2935710</td>
<td>Inside the gene</td>
<td>2.21</td>
<td>Aminopeptidase</td>
<td>Transcription</td>
</tr>
<tr>
<td>adh-11</td>
<td>2469950</td>
<td>2470580</td>
<td>Inside the gene</td>
<td>1.96</td>
<td>Alcohol dehydrogenase</td>
<td>Transcription</td>
</tr>
<tr>
<td>sso2153</td>
<td>1977900</td>
<td>1978650</td>
<td>Inside the gene</td>
<td>2.66</td>
<td>Archaeal putative transposase ISC1217; pfam04693</td>
<td>Transcription</td>
</tr>
<tr>
<td>sso12210</td>
<td>2965830</td>
<td>2966580</td>
<td>Inside the gene</td>
<td>2.07</td>
<td>Transcription regulator</td>
<td>Translation</td>
</tr>
<tr>
<td>sso10340</td>
<td>2189085</td>
<td>2189390</td>
<td>Inside the gene</td>
<td>2.31</td>
<td>Conserved,truncated variant of Lrp/AsnC-family;RNA polymerase and transcription factors</td>
<td>Transcription</td>
</tr>
<tr>
<td>sso2626</td>
<td>2391100</td>
<td>2392050</td>
<td>Inside the gene</td>
<td>2.07</td>
<td>Conserved,truncated variant of Lrp/AsnC-family;RNA polymerase and transcription factors</td>
<td>Transcription</td>
</tr>
<tr>
<td>sso2827</td>
<td>2588200</td>
<td>2589300</td>
<td>Inside the gene</td>
<td>2.07</td>
<td>Conserved,truncated variant of Lrp/AsnC-family;RNA polymerase and transcription factors</td>
<td>Transcription</td>
</tr>
<tr>
<td>rps26E</td>
<td>502000</td>
<td>502750</td>
<td>Inside the gene</td>
<td>2.49</td>
<td>30S ribosomal protein S26e</td>
<td>Translation</td>
</tr>
<tr>
<td>rp144c</td>
<td>907520</td>
<td>908380</td>
<td>Inside the gene</td>
<td>1.88</td>
<td>50S ribosomal protein L44e</td>
<td>Translation</td>
</tr>
<tr>
<td>murp</td>
<td>395515</td>
<td>396010</td>
<td>Inside the gene</td>
<td>1.78</td>
<td>ATPases involved in chromosome partitioning</td>
<td>Transcription</td>
</tr>
<tr>
<td>sso2730</td>
<td>2485375</td>
<td>2485920</td>
<td>Inside the gene</td>
<td>1.87</td>
<td>Possibly ATPase of the AAA superfamily</td>
<td>Transcription</td>
</tr>
<tr>
<td>metS</td>
<td>491920</td>
<td>492385</td>
<td>Inside the gene</td>
<td>2.48</td>
<td>Metionine-tRNA ligase, RNA methyltransferase, DNA methyltransferase, rRNA</td>
<td>RNA processing and modification</td>
</tr>
<tr>
<td>sso1044</td>
<td>903240</td>
<td>904070</td>
<td>Inside the gene</td>
<td>2.96</td>
<td>Methionine-tRNA ligase, RNA methyltransferase, DNA methyltransferase, rRNA</td>
<td>RNA processing and modification</td>
</tr>
<tr>
<td>sso1288</td>
<td>1114815</td>
<td>1115458</td>
<td>Inside the gene</td>
<td>2.16</td>
<td>short last area hits to a zoo of molecules associated with cell wall/cell membrane</td>
<td>Defense mechanisms</td>
</tr>
<tr>
<td>sso3150</td>
<td>2902540</td>
<td>2903225</td>
<td>Inside the gene</td>
<td>3.19</td>
<td>Putative membrane protein</td>
<td>Defense mechanisms</td>
</tr>
<tr>
<td>sso3178</td>
<td>2923630</td>
<td>2924485</td>
<td>Inside the gene</td>
<td>2.10</td>
<td>Putative membrane protein</td>
<td>Defense mechanisms</td>
</tr>
<tr>
<td>sso2749</td>
<td>2502810</td>
<td>2503615</td>
<td>Inside the gene</td>
<td>2.37</td>
<td>Oxidoreductase,oxidative-stress response multicopy, similar to sso1886;proteases</td>
<td>Defense mechanisms</td>
</tr>
<tr>
<td>sso2037</td>
<td>1850715</td>
<td>1851580</td>
<td>Inside the gene</td>
<td>3.06</td>
<td>Oxidoreductase,oxidative-stress response multicopy, similar to sso1886;proteases</td>
<td>Defense mechanisms</td>
</tr>
<tr>
<td>sso1758</td>
<td>1592730</td>
<td>1593280</td>
<td>Inside the gene</td>
<td>2.20</td>
<td>Oxidoreductase,oxidative-stress response multicopy, similar to sso1886;proteases</td>
<td>Defense mechanisms</td>
</tr>
<tr>
<td>sso0927</td>
<td>787505</td>
<td>788400</td>
<td>Inside the gene</td>
<td>2.09</td>
<td>Oxidoreductase,oxidative-stress response multicopy, similar to sso1886;proteases</td>
<td>Defense mechanisms</td>
</tr>
<tr>
<td>sso0438</td>
<td>380250</td>
<td>381045</td>
<td>Inside the gene</td>
<td>2.45</td>
<td>Oxidoreductase,oxidative-stress response multicopy, similar to sso1886;proteases</td>
<td>Defense mechanisms</td>
</tr>
<tr>
<td>sso1984</td>
<td>1796200</td>
<td>1796975</td>
<td>Inside the gene</td>
<td>2.98</td>
<td>Oxidoreductase,oxidative-stress response multicopy, similar to sso1886;proteases</td>
<td>Defense mechanisms</td>
</tr>
</tbody>
</table>
Figure 1

A

![Image of gel electrophoresis with molecular weight markers and bands labeled PEXA3 and sso2474-V.]

B

![Image of gel electrophoresis with molecular weight markers and bands labeled PEXA3 and 10340-v.]

(c)

![Image of gel electrophoresis with molecular weight markers and bands labeled PEXA3 and sso10340.]

- 10kDa
- 15kDa
- 27kDa
- 55kDa
- 70kDa

1000bp

octamer
dimer

ss10340
Figure 3
Figure 4
Distribution of the genomic locations of 10340 binding sites

- 46% Intragenic
- 26% Intragenic but upstream
- 15% Upstream
- 13% Intergenic
Query protein sequence:

mqgkseismp dgrvadvfnv vkflyglsdr dieilkllik sqssltmeei sselnitksv vnks
ilnlek knivikekve sskkgrrayt yrvdvnyltr klvtdldqli kdlkvkiadv igiqiekt

Genomes without synteny:

- Acidilobus_saccharovorans_345_15_uid51395
- Caldisphaera_lagunensis_DSM_15908_uid183486
- Aeropyrum_camini_SY1_JCM_12091_uid222311
- Aeropyrum_pernix_K1_uid57757
- Desulfurococcus_fermentans_DSM_16532_uid75119
- Desulfurococcus_kamchatkensis_1221n_uid59133
- Desulfurococcus_mucosus_DSM_2162_uid62227
- Ignicoccus_Hospitalis_KIN4_1_uid58365
- Ignisphaera_aggregans_DSM_17230_uid18575
- Staphythermus_hellenicus_DSM_12710_uid45893
- Staphythermus_marinus_F1_uid58719
- Thermogadus_1633_uid67488
- Thermosphaera_aggregans_DSM_11486_uid48993
- Hyperthermus_butylicus_DSM_5456_uid75775
- Thermofilum_1910b_uid20133
- Caldivirga_maquilingensis_IC_167_uid58711
- Pyrobaculum_1860_uid82379
- Pyrobaculum_aerophilum_IM2_uid57727
- Pyrobaculum_islandicum_DSM_4184_uid58635
- Pyrobaculum_neutrophilum_V24_UID37717
- Archaeoglobus_fulgidus_DSM_4304_uid51875
- Archaeoglobus_profundus_DSM_5631_uid43493
- Archaeoglobus_sulfaticallidus_PW70_1_uid201033
- Archaeoglobus_vanishii_SN6_uid56269
- Halalkalicoccus_mukohataei_DSM_12286_uid59107
- Haloferax_thermautotrophicus_ATCC_51293_uid51877
- Halobacterium_NRC_1_uid57755
- Halobacterium_salinarum_R1_uid61571
- Haloferax_volcanii_DS2_uid46845
- Halomicrobium_mukohataei_DSM_12286_uid59107
- Halorhabdus_tiamatea_SARL4B_uid201033
- Halorhabdus_utahensis_DSM_12940_uid58719
- Halorubrum_lacusprofundi_ATCC_49239_uid51877
- Haloterrigena_turkmenica_DSM_5511_uid43501
- Halovivax_ruber_XH_70_uid184819
- Natrialba_magadii_ATCC_43099_uid46245
- Natronobacterium_gregoryi_SP2_uid74439
- Natronococcus_occultus_SP4_uid48993
- Natronomonas_moolapensis_8_8_11_uid190182
- Natronomonas_pharaonis_DSM_2160_uid58435
- Methanobacterium_AL_21_uid63623
- Methanobrevibacter_AKNM_uid60167
- Methanobrevibacter_ruminantium_M1_uid45897
- Methanocaldococcus_FS406_22_uid42499
- Methanocaldococcus_IN_1_uid74199
- Methanocaldococcus_vulcanius_M7_uid41131
- Methanococcoides_burtonii_DSM_6242_uid58023
- Methanocorpusculum_labreanum_Z_uid58785
- Methanoculleus_marisnigri_JR1_uid58561
- Methanoplanus_petrolearius_DSM_11571_uid58023
- Methanoregula_formicicum_ST04_uid169620
- Methanosphaerula_palustris_E1_9c_uid59199
- Methanosalum_zhilinae_DSM_4017_uid68249
- Methanopyrus_kandleri_AV19_uid57755
- Pyrococcus_abyssi_GE5_uid62903
- Pyrococcus_furiosus_DSM_3638_uid57873
- Pyrococcus_ST04_uid167261
- Pyrococcus_yayanosii_CH1_uid68281
- Thermococcus_sibiricus_MM_739_uid59399
- Methanocella_paludicola_SANAE_uid42887
Sulfolobus tokodaii
uid57807
Score: 44.97

Sulfolobus acidocaldarius
SUSAZ
uid232254
Score: 44.97

Sulfolobus acidocaldarius
N8
uid189027
Score: 44.97

Sulfolobus islandicus
Y
G
57
14
uid58923
Score: 80.54

Sulfolobus islandicus
M
14
25
uid58849
Score: 81.21

Sulfolobus islandicus
HVE10
4
uid162067
Score: 81.21

Sulfolobus islandicus
REY15A
uid162071
Score: 81.21

Sulfolobus islandicus
LAL14
1
uid197216
Score: 81.21

Sulfolobus islandicus
L
D
8
5
uid43679
C1
Score: 81.88

Sulfolobus islandicus
L
S
2
15
uid58871
Score: 81.88

Sulfolobus solfataricus
P2
uid57721
Score: 81.88

Sulfolobus islandicus
Y
N
15
51
uid58825
C1
Score: 81.88

Sulfolobus acidocaldarius
Y N 15 51
uid58828
Score: 81.88

Sulfolobus acidocaldarius
P2 uid57721
Score: 43.69

Metallosphaera cuprina
Ar 4 uid66237
Score: 43.69

Acidiphilus hospitalis
W1 uid6875
Score: 39.53
Thermoplasmatales

- Natronomonas pharaonis
  DSM 2160 udd58435 C7
  Score: 15.56

- Halorhodoid utahensis
  DSM 12340 udd9189 C7
  Score: 15.23

Methanococcoides infernus
ME udd48803
Score: 15.23

- Methanococcoides FS406
  22 udd45490 C9
  Score: 15.23

- Methanosarcina barkeri
  Fusaro udd57715 C7
  Score: 15.23

Thermoplasmales archaean
BRNA1 udd18530
Score: 15.23

- Methanococcus maripaludis
  S2 udd58035
  Score: 15.23

- Methanothermococcus IH1
  udd51535
  C1
  Score: 15.23

- Haloferax volcanii
  DS2 udd58829
  Score: 15.23

- Methanococcus maripaludis
  C5 udd59847
  Score: 15.23

- Methanococcus maripaludis
  C6 udd58947
  Score: 15.23

- Methanococcus maripaludis
  C7 udd59847
  Score: 15.23

- Methanococcus maripaludis
  X1 udd70279
  Score: 15.23

- Halorhabdus utahensis
  DSM 12940 udd59189
  Score: 15.23

- Methanothermococcus DH1 udd51535 C7
  Score: 15.23

- Methanococcus maripaludis
  S2 udd8033
  Score: 15.23

- Natronalba magadii
  ATCC 43099 udd146245 C7
  Score: 14.97

- Methanococcoides
  DSM 2561 udd57713 C7
  Score: 14.97

- Methanoplanus petrolearius
  DSM 11571 udd2695
  Score: 14.97
Query protein sequence:
maevryi1 vsttvgkeme vadakkvsq viradpvyge ydvvveak ssddkviy eir
npnlr tvilivm

Genomes without synteny:
Staphylothermus_hellenicus_DSM_12710_uid45893
Pyrococcus_fontis_Kam940_uid4162201
Pyrobaculum_calidifontis_JCM_11548_uid58787
Pyrococcus_islandicum_DSM_4184_uid58635
Thermoproteus_tenax_Kra_1_uid74443
Archaeglobus_fulgidus_DSM_4304_uid57717
Archaeglobus_sulfaticallidus_PM70_1_uid201033
Haloarcus_hispansc_AYCC_33960_uid72475_C1
Haloarcus_hispansc_N601_uid230920_C1
Haloferax_mediterranei_KTOC_33500_uid167315_C1
Haloarcula_borinquense_DSBM_11551_uid54919_C1
Halomicrobium_mukohataei_DSM_12286_uid59107_C1
Halopiger_xanadensis_N4_6_uid68105_C1
Haloterrigena_turkmenica_DSM_551_uid43501_C1
Natrinema_J7_uid171337_C1
Natrinema_pellirubrum_DSM_15624_uid74437_C1
Natronobacterium_gregoryi_SP2_uid74439
Natronococcus_occultus_SP4_uid184863_C1
Methanobacterium_MBL1_uid231690
Methanobacterium_SWAN_1_uid67359
Methanobrevibacter_ADMA_uid206516
Methanobrevibacter_ruminantium_M1_uid45857
Methanosphaera_stadtmanae_DSM_3091_uid58407
Methanothermus_fervidus_DSM_2088_uid60167
Methanoterris_igneus_K015_uid67321
Methanocellulosa_arvoryzae_MB450_uid61623
Methanocellulosa_conradii_HZ224_uid157911
Methanocellulosa_paludicola_SANAE_uid42887
Methanocorpusulum_labreanum_Z_uid58785
Methanoecillus_burgensis_MS2_uid71377
Methanoregula_formicicum_DSM_8023_uid184406
Methanospirillum_hungatei_JF11_uid58181
Methanococcoides_burtonii_DSM_6242_uid58023
Methanobrevibacter_psychrophilus_K15_uid177925
Methanosarcina_acetivorans_C2A_uid57879
Methanosarcina_mazei_Tuc01_uid190185
Picrophilus_torridus_DSM_9790_uid58041
Figure S3

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lrp</td>
<td>MVDSKRPGK</td>
<td>DLDRIDRNIL</td>
<td>NELQRDGRIS</td>
<td>NVELSKRVGL</td>
<td>SPTPCLERV</td>
</tr>
<tr>
<td>ssol0340</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clustal C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lrp</td>
<td>RLRQGFIQG</td>
<td>YTAALLNPYL</td>
<td>DASLLVFEI</td>
<td>TINRGAPDF</td>
<td>EQFTAVQKL</td>
</tr>
<tr>
<td>ssol0340</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clustal C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lrp</td>
<td>EEIQECHLVS</td>
<td>GDFDIILKTR</td>
<td>VPDMSAYRKL</td>
<td>LGETLLALPG</td>
<td>VNDTRTYVVM</td>
</tr>
<tr>
<td>ssol0340</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clustal C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lrp</td>
<td>EEVQSNSRLV</td>
<td>IKTR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssol0340</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clustal C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure S4

1 Sso 0438 peak area(380600-380750) nr 220

AAATAGTTACACTTATCTCCTAAAGCATTGGTTTCTGCTC
tACGTAAGGCTGTCTTTCAAAGCCCTCT
ATTTCC

GTTCCCAAATCTACTTATACCCCCAAAACCTTTTAGACAATCATATTAACTTTACATTTTGACCCCATTTA
ATAGTATCTATACCTTACCTTTACGATCCACTAGAGCCAAATACCTATCTCTTTTTAGTGCTATG
ACC

Primer
EMSASso0438 S  5 CAAAGCCTCTATTCCGTTC  3 TM 53.8
EMSAsso0438 A  5 CGGTAGAGTCTCACATAGCA  3 TM 55.7 168bp
2 Sso 0570 peak area (502350-502600) nr192 (with rps26E together transcription)

GTTGTGATCAATGTGGTGCTAG
AGTACCAGAGGATAAGGGGATGATGTAAC
GCTTCT
CTAGCATCTGAAATTAGAAAGAGGATGGCAAATATTGCTAGATATCCTGTAACTAAG
TGTTACTGTGTAATTG
TGCG
ATTTTTGGGTATTATTAAGATAAGCAGAAATGAGAGAAAGCAAAAAGCTCGTTTAAGATAGGCT
TTTAAA
C
CTTTAG
TCAGAATATG7GATGAAGCTAGCTTTTTCTGAATTATTATCTTTGCACAAATTGAAGATTTTTCTCTTTGATAGCTGCTATTAG

Primer

EMSAasso0570 S 5 GTTGTGATCAATGTGGTGCTAG 3  TM 57.4
EMSAasso0570 A 5 CGCACAATTCACACAGTAACA 3  TM 57.5  158bp
3 Rp144e peak area (907700-908000) nr 195

TATATAATGATCAGGGTTAATGAGGGAGATGTTTTGGTTTTACCTATGAGAGAA

G

GCTG

GAAGGACA

GAGAAGATATGAGAG

AAAGAATATTGGATATGGAAGTAAAAGAAAACCAGAACAGAAGAGATTTGCAAAAGTT

Primer

EMSArp144e S 5 GCAAGTTATTTAACTCCCTATAAG 3 TM 53.2
EMSArp144e A 5 CTCTCATATCTCT CTGTCCTTC 3 TM 55.2 168bp
4 Sso 1758 peak area (1592900-1593150) nr 213

Primer

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
<th>TM</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMSAsso1758 S</td>
<td>5 GAATGATCCTGGAATGCTTTC</td>
<td>54.5</td>
<td>171bp</td>
</tr>
<tr>
<td>EMSAsso1758 A</td>
<td>5 GGAAGAAATATAAAGCATTAG</td>
<td>55.2</td>
<td></td>
</tr>
</tbody>
</table>
5 Sso 1879 peak area (1692550-1692750)

GGTTATCCACCTATGGATACCTTGTAACGATAATATTAAATATTGAGGAGCTGT CAGACGCATTGCGAAGTTT
AGCTGAGTACTTACGCTCCATAATGAATTAACAGAATAAGGTATAGAAGCGGAATG AATAATGGCTAACTTTA
TCACCTCAATAC

ATTAAATCTATATTGTGACATTTGACGACTTAAATAA GTTAATCAGAGAGAAACTTAGCG TAGAAACGTACCCT
TATCAAAAAGTACATCAG

Primer

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
<th>TM</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMSAsso1879 S</td>
<td>5' CAGACGCATTGCGAAGTTT 3'</td>
<td>56.5</td>
<td>165bp</td>
</tr>
<tr>
<td>EMSAsso1879 A</td>
<td>5' CGCTAAGTTTCTCTGATTAAC 3'</td>
<td>54.9</td>
<td></td>
</tr>
</tbody>
</table>
6 Rpom-2 peak area (1731850-1732040) nr 105.8 (with sso1913 together: transcription)

Primer

<table>
<thead>
<tr>
<th>EMSArpom-2 S</th>
<th>5 GATCAGAATACCCACACTTAC 3</th>
<th>TM 53.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMSArpom-2 A</td>
<td>5 GAGTTAGTATGGTCCGCA 3</td>
<td>TM 54.3</td>
</tr>
</tbody>
</table>

134bp
7 Sso1984 peak area (1796450-1796650) nr 363

Primer

EMSAso1984 S  5' GGTCAAGCAATAATATAATGGAATACAGATCTCTGTGGGATTAAATACATAGTAGAATGCAATAGTTCAATACCAAATAGTTCAAT

EMSAso1984 A  5' CCTACCCATTCAATCTTAATTGCTGATGGTTATGGGGTTTTGCCGTTAATGGAATGTAACAATAAATGGACAGACGTTATTACGTGATAGATGGAGATAAAATAAACCCAGCGGACATCGCGGATTTTTCACATACTGGAGATGGGTAAACAACTTC

FPsso1984 S  5' GCTTAAAGGGTTTGGTTTTTCCCTACCCATTCAATCTTAATTGCTGATGGTTATGGGGTTTTGCCGTTAATGGAATGTAACAATAAATGGACAGACGTTATTACGTGATAGATGGAGATAAAATAAACCCAGCGGACATCGCGGATTTTTCACATACTGGAGATGGGTAAACAACTTC

FPsso1984 A  5' GAAGTTGTTTATCCCAAGTGGATTTAATGGAATGTAACAATAAATGGACAGACGTTATTACGTGATAGATGGAGATAAAATAAACCCAGCGGACATCGCGGATTTTTCACATACTGGAGATGGGTAAACAACTTC
8 Sso 2102 peak area (1923900-1924100) nr 143

GAATCGAAATATGTAAGCAATAGTTTCAGTTATTATGTTGACTATTAGCTGAAGGAG
ATGTGACGACTGCTCTGGCAGCTTAGCTAAAAGTGTGCCATGTGACGAAGCTGTGGAACAGTTTGCTTAATATTTTGTGTGGAT
ATATGTATACCTTCTTTGATGAACTACTTCAACTTATATATATCTCAATAGGAGTAGCGACTG
CATGTGGACTAGGTGGGTAGTATCTATGTCAAAAGGCTGGTCTGTGTTGCTAATAGATTTTTTATAG
ATATGTGGTTATAAATAGCTAAAAAGGAGGGATAGAA
ATATATGAGAAAAAGGACAGAAAATAGAAGAAATAAGGATTATCTCTTTTTAAGTGGCTTGACACAATTATTATT
TATGATTTTGCCCTT

Primer

| EMSAsso2102 S | 5 GTGGACTATTAGCTGAAAGGAG 3 | TM 54.6 |
| EMSAsso2102 A | 5 CAGTCGCTACTCTTTATTGAGATA 3 | TM 56.5 | 171bp |
9 Sso 2626 peak area (2391350-2391550) nr 167.4

ATGAGATAAACTAAACAATTGAGAGATGAAGCTGACAAATTA CAACAACCCCTTAAGAAGTATATT GGGCTGTGTC

ATAGGTGCGACAGGGCTACTTTTG (TATA) TTTATGATCTAAGAAGGATGGTGCCTTACTATCGAATAGATGCAATACC GTTATCAATAAAATTAAGATCAAATACCCGGACGAATTATTTGAGAGGTGACCTTAGAAGATGCAATACC

Primer

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
<th>TM</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMSAsso2626 S</td>
<td>5 CAACAACCCCTTAAGAAGTATATT 3</td>
<td>54</td>
</tr>
<tr>
<td>EMSAsso2626 A</td>
<td>5 CTCCCCCTCTAAACTGCTTCAAT 3</td>
<td>57</td>
</tr>
</tbody>
</table>
Pacs

10 Pacs peak aera (2411750-2411950) nr 1080

Primer

EMSApacs S 5 GTCTTATATTGACTTTTCTCATCTAC 3 TM 54.0
EMSApacs A 5 GCTTGCTGG ATAAAAAGAAAA 3 TM 54.4

FPpacs S 5 CTCTGAAACTCCCTTAGACAATG 3 TM 56.4
FPpacs A 5 CAAACTGTGGATGGATTTA 3 TM 54.3
**Adh-11**

11 Adh-11 peak area (2470250-2470400) nr 120

ATAAAAAATACCTTCCATTCTTACCTTTATTTGCTATTAGTTATTTATGTTCCGATGAAGTATATATA

AGACCTATCCCAATTAAGATTTCCATTCCTAAAGGATATGGTGCTGCCATTGTATCCCCGACCATCTTAGGAAAGAAGAAAGTTACCATTAGTATTTCAAC

Primer

| EMSAadh-11 S | 5 GAGATATGCTTGGTGACTTCC 3 | TM 56.0 | 160bp |
| EMSAadh-11 A | 5 CCTAAGATGGTGGGATAC 3 | TM 55.8 |
12 Sso 3178 peak area (2923900-2924050) nr 206

TTCTTTGAAAGCTTATCACATGTAGTCCAGTTCTCTATAGTTCTCTATCTCTCCTTTTCTTAGCAGTCATATTG
ACTTATGAAATTATAAACAGACGTCTTTTTGGAATAAATGGGTTAAGTTTTGCTCACTTTCTACTTTTAAGAGAGAG
TAAATTTAAATTTTCTTTTTAAATTTTTGTIAATGCCTTGAGTTCTAAATAGTATA
TAGTTAAAAAGAGAGGGCTGAGAAATGAAATCTTTCTAGCTTCATTTTCTATTAAAGCAAGGTGGG
CATTATCGTAGCTGAGAAACATAGTCCAGTCC

Primer

| EMSAsso3178 S | 5' GCTCTATCACATGTAGTCCAG 3' | TM 55.1 |
| EMSAsso3178 A | 5' GCATAGAACCTCAAAGCATTAG 3' | TM 55.2 |

188bp
13 Sso3189 peak area (2935250-2935550)

AAGTAATATATAGAAGATATCTCTATCATACATCTTTGGCAATGGATAAGTAAATTTAAATAGGGGAATTTAGAAGA
TAGCGT

CAGCCACAATACTCCTCCTTGCTTCATTAATAGGCTTCATCTTCGCTATACCACAACTGTTAATTTATGTTATTTTAAC

Primer

EMSAsso3189 S  5 GGGGAATTTAGAAGATAGCGT  3 TM 54.5
EMSAsso3189 A  5 CAGCCACAATACTCCTCCTT  3 TM 56.2

183bp
14 Sso 12210 peak area (2966200-2966400)

Primer

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
<th>TM</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMSasso12210 S</td>
<td>5'CTTTCTACAAAAAGGATTCATCTC 3'</td>
<td>55.6</td>
</tr>
<tr>
<td>EMSasso12210 A</td>
<td>5'GGTAAACACGAGTACGCTCTC 3'</td>
<td>54.5</td>
</tr>
</tbody>
</table>
Area  505950-506100  151bp

CTCCTAAAAATAGCGAACCG

CCAAGGTTCTTCTGTTGCTTCCCTCTATAACATAA

CGGTATAGGGTC

Primer

EMSA no binding S  5 CTCCTAAAAATAGCGAACCG 3  TM 53.5
EMSA no binding A  5 GACCCATACCGGATATTACG 3  TM 54.5  [160bp]
Unveiling Cell Surface and Type IV Secretion Proteins Responsible for Archaeal Rudivirus Entry

Ling Deng, Fei He, Yuvaraj Bhoobalan-Chitty, Laura Martinez-Alvarez, Yang Guo and Xu Peng


Updated information and services can be found at: [http://jvi.asm.org/content/88/17/10264](http://jvi.asm.org/content/88/17/10264)

**SUPPLEMENTAL MATERIAL**

Supplemental material

**REFERENCES**

This article cites 32 articles, 11 of which can be accessed free at: [http://jvi.asm.org/content/88/17/10264#ref-list-1](http://jvi.asm.org/content/88/17/10264#ref-list-1)

**CONTENT ALERTS**

Receive: RSS Feeds, eTOCs, free email alerts (when new articles cite this article), [more»](http://jvi.asm.org/content/88/17/10264#ref-list-1)

Information about commercial reprint orders: [http://journals.asm.org/site/misc/reprints.xhtml](http://journals.asm.org/site/misc/reprints.xhtml)

To subscribe to another ASM Journal go to: [http://journals.asm.org/site/subscriptions/](http://journals.asm.org/site/subscriptions/)
Unveiling Cell Surface and Type IV Secretion Proteins Responsible for Archaeal Rudivirus Entry

Ling Deng, Fei He, Yuvaraj Bhoobalan-Chitty, Laura Martinez-Alvarez, Yang Guo, Xu Peng
Archaea Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark

*Sulfolobus* mutants resistant to archaeal lytic virus *Sulfolobus islandicus* rod-shaped virus 2 (SIRV2) were isolated, and mutations were identified in two gene clusters, cluster sso3138 to sso3141 and cluster sso2386 and sso2387, encoding cell surface and type IV secretion proteins, respectively. The involvement of the mutations in the resistance was confirmed by genetic complementation. Blocking of virus entry into the mutants was demonstrated by the lack of early gene transcription, strongly supporting the idea of a role of the proteins in SIRV2 entry.

To date, relatively few archaeal viruses have been characterized, and most of those that have been characterized infect acidothermophilic members of the order *Sulfolobales*. Despite their limited number of around 50 species, they exhibit considerably greater morphological diversity than the more extensively characterized bacteriophages, about 95% of which show head-tail morphologies. Archaeal viruses, in contrast, exhibit fusiform shapes, often with one or two tails, bottle shapes, bearded-globular forms, and a wide variety of rod-like and filamentous morphotypes which often carry small terminal appendages (1–3). This morphological diversity suggests that the archaeal viruses may employ a variety of mechanisms to enter their hosts, but current insights into entry mechanisms are limited to an OppA transporter protein, Sso1273, possibly providing a receptor site for the Acidianus two-tailed virus (ATV) in *Sulfolobus solfataricus* P2 (4). And very recently, microscopic studies suggested that *Sulfolobus islandicus* rod-shaped virus 2 (SIRV2) enters the host cell by attaching and moving through a pilus-like filament; however, the nature of the structure and the identity of the involved proteins remain elusive (5).

*Sulfolobus solfataricus* P2 is an acidothermophilic crenarchaeon that can host a wide range of archaeal viruses, many of which are propagated stably (1, 3, 6). Moreover, few of the viruses appear to induce cell lysis, possibly reflecting a need to minimize contact with the harsh hot acidic environment. However, recent studies have identified a few viruses that can enter a lytic phase, including the *Sulfolobus* turreted icosahedral virus (STTV), the two-tailed fusiform (ATV), and, more recently, the rudivirus SIRV2 (7–9).

SIRV2 is classified in the family *Rudiviridae* together with other well-characterized viruses, including SIRV1 (10, 11), ARV1 (12) and SRV1, (13), all of which are rod shaped and lack an envelope, and their genomes consist of linear double-stranded DNA with covalently closed ends (10, 14, 15). In a recent microarray analysis of *S. solfataricus* infected with SIRV2, we demonstrated that the viral genes were activated at different times and that mainly stress-response host genes and those implicated in vesicle formation were downregulated (16). The results also illustrated that SIRV2 infection at a multiplicity of infection (MOI) of 30 resulted in growth inhibition of *S. solfataricus* 5E6 (16). In the present experiment, the culture was infected at a lower MOI (≤1) which also led to a growth retardation, but the infected culture could enter the exponential-growth phase at 80 h postinfection (p.i.) (Fig. 1A).

The surviving cells (named 5E6R) appeared to be resistant to SIRV2 because, in contrast to the sensitive 5E6 strain, no growth...
retardation was observed when 5E6R was diluted and infected with SIRV2 at the same MOI (Fig. 1).

In order to manipulate the SIRV2-sensitive *S. solfataricus* 5E6 strain genetically (17), 10 pyrEF mutants, labeled Sens1 to Sens10, were isolated from Gelrite plates containing 5-fluoroorotic acid (5'-FOA). Their mutation sites in the *pyrEF* gene region were identified by a combination of PCR amplification, restriction digest analysis, and sequencing (17). All the mutations were shown to result from transposon insertions, either IS elements or miniature inverted terminal repeat elements (MITEs), and the insertions occurred in the coding sequences or within the single promoter (Fig. 2A). These results are consistent with the previous reports demonstrating high transposition activity in *S. solfataricus* and its contribution to chromosomal plasticity (18–20). Following the procedure described above, SIRV2-resistant cultures were generated for each of the *pyrEF* mutants. Single colonies were then produced from the cultures by streaking onto Gelrite plates to yield the purified resistant strains Res1 to Res10. The stability of the transposon insertions in the *pyrEF* genes was tested for each of the 10 *pyrEF* mutants (Sens1 to Sens10) and their corresponding SIRV2-resistant colonies (Res1 to Res10) by growing them in rich media containing uracil (17) for 3 days without transfer, prior to total DNA extraction and PCR amplification of the *pyrEF* regions. Each transposon insertion appeared to be stable, because no wild-type PCR bands were observed, except a weak wild-type band produced in Sens2, consistent with the undetectable reversion rates for *Sulfolobus* transposons recorded earlier (19, 20). Since Res2 did not generate the wild-type band, the extra PCR product in Sens2 was probably due to a minor contamination of the colony by wild-type cells (Fig. 2B).

Sens1, Sens3, Sens7, and Sens8 were selected for a transformation test because they carried different transposons located at different insertion sites (Fig. 2A). Shuttle vector pEXA was used for transformation (21), and water was used in the negative control. While Sens7 and Sens8 appeared unstable after electroporation, Sens1 and Sens3 yielded transformants without colony formation in the negative control. Thus, we focused on Sens1 and its resistant mutants for further studies of SIRV2 susceptibility.

The SIRV2-resistant cells were enriched directly from the SIRV2-sensitive culture; therefore, the only selective pressure appeared to occur either upon SIRV2 infection or during virus-induced cell lysis. Moreover, since the active clustered regularly interspaced short palindromic repeat (CRISPR) loci were all lost from the 5E6 host strain (16), the residual CRISPR loci E and F, which lack the spacer acquisition *cas* genes, were unlikely to be responsible for the resistance (21, 22). Therefore, we inferred that resistance arose as a result of mutated host genes that are important for the SIRV2 life cycle. To identify such mutations, the genomes of strains Sens1 and Res1 were sequenced by the use of a HiSeq 2000 sequencer, yielding about 200-fold coverage. The sequencing reads of both strains were aligned with the genome sequence of *S. solfataricus* P2 (23) using the R2R program (24) to identify mismatches as well as insertions and deletions. Mutations to the P2 genome in the sequenced genomes of Sens1 and Res1 were then compared manually. Only one mutation was detected and constituted a single insertion of *IS1078* into Res1 but not Sens1. The insertion was localized in *sso3139*, a gene encoding a conserved hypothetical protein lying within an operon (Fig. 3A).

Next we tested whether other resistant strains also carried mutations in *sso3139* or in other genes of the same operon by employing a primer pair (5'-GCTACGCTTCTAAACACCTAATCTG and 5'-CGAAAAGTTCTTCACTTACGT) designed to amplify the whole operon region. After PCR amplification, restriction digestion, and sequencing, another 5 strains were shown to contain mutations at different locations within *sso3139* or the adjacent *sso3140* (Fig. 3A). Interestingly, all the 6 mutations were produced by *ISC1078* insertion (Fig. 3A) and appeared to be stably maintained (Fig. 3C).

To identify possible mutations in the other 4 resistant strains, genome resequencing followed by PCR analyses of relevant genes was performed (using primers 5'-GAGTCTGGGGAAAATCGGT and 5'-TGCGCATGTAAACCTAATTCTT). These revealed IS element insertions in *sso2387* of Res2 and Res10 (Fig. 3B). *Sso2387* in Sens2 and Sens10 contains 577 amino acids (aa) but only 283 aa in the sequenced *S. solfataricus* P2 genome (23). An analysis of the sequences around *sso2387* in Sens2 and Sens10 revealed that it is a partial gene that resulted from an IS element insertion (Fig. 3A). In *S. solfataricus* P2, the sequence downstream of the IS was 308 nucleotides (nt), whereas in Sens2 and Sens10, only 195 nt were present. The absence of the 113 nt downstream of the IS resulted in a premature stop codon, which targets the encoded Sso2387 protein to the membrane, thereby preventing its secretion.

The frequently observed mutations in cluster *sso3139* and

![FIG 2](https://jvi.asm.org/figures/figure2.png) Analysis of the *pyrEF* mutants derived from *S. solfataricus* 5E6. (A) Types and insertion sites of transposons inserted in the *pyrEF* gene region of different *pyrEF* mutants. (B) PCR amplification of the *pyrEF* region from Sens1 to Sens10 and from Res1 to Res10. wt, wild type.
sso3140 and cluster sso2386 and sso2387 in the resistant strains strongly suggest that the two gene clusters are important for the SIRV2 life cycle. To confirm the implication of the mutations in the gained resistance, genetic complementation was performed for the mutated genes. As described above, Sens1 appeared stable during genetic manipulation, and we thus selected Res1 for complementation of sso3139 mutation. For complementation of mutations in the other gene cluster, Res1B, carrying an ISC1234 insertion in sso2387 (Fig. 3B), was isolated from SIRV2-infected Sens1. Res1 cells were transformed with vector pEXA2 containing sso3139, and Res1B cells were transformed with vector pEXA2 containing sso2386 and sso2387. After SIRV2 was added into the cultures, growth retardation occurred in the complemented cells, while the noncomplemented culture, transformed with the empty vector, showed a growth rate similar to that of the uninfected culture (Fig. 4A and B). Further, Southern hybridization (17) using a probe derived from the SIRV2 inverted terminal repeats (ITR) detected signals only from the complemented cells (Fig. 4C and D) and the multiple hybridized bands were consistent with ongoing replication (Fig. 4E) (10, 25). The absence of SIRV2 signal in the resistant strains indicates a defect in the virus life cycle.

To gain insights into the functions of the two gene clusters, the protein sequences of the genes were firstly analyzed by the use of program TMHMM (http://www.cbs.dtu.dk/services/TMHMM-2.0/) for the possible presence of transmembrane helices. Sso3138, Sso3139, and Sso3140 were predicted to be primarily located extracellularly (see Fig. S1 in the supplemental material), correlating with a previous prediction of the presence of class III signal peptides at their N termini (26). Among these, Sso3140 was confirmed to be a membrane-associated protein in a proteomics study (27). In contrast to the other 3 proteins, Sso3141 was predicted to contain two transmembrane helices, one at the N terminus and the other at the C terminus, while the sequence between them was presumed to be located intracellularly. Therefore, it appears that the proteins encoded in the operon form a membrane-associated cell surface structure and may function as a receptor for SIRV2. Moreover, it was demonstrated recently that Sso2386 carries multiple transmembrane helices and that Sso2387 constitutes an ATPase associated with a type IV secretion system, and they were designated AapF and AapE, respectively (28). Further, homologs of both are essential for the formation of the adhesive type IV pilus of S. acidocaldarius (28). The association with the cell membrane of proteins encoded by both gene clusters strongly indicates their involvement in the entry process of SIRV2.

The failure of viral entry into Res1 and Res1B cells was further confirmed by reverse transcription-PCR analysis of one of the early genes, ORF131a (17). RNA extracted from cells taken at 15 min p.i. was DNase I treated and reverse transcribed (SuperScript II reverse transcriptase; Invitrogen). PCR performed on the cDNAs detected ORF131a only from Sens1 cells, while the positive-control sso0446 (tfb-1) gene was detected in all the 3 strains (Fig. 4F). This strongly supports the conclusion that the proteins encoded by the two gene clusters are involved in SIRV2 entry. A likely scenario is that gene cluster sso3138 to sso3141 encodes a surface receptor for SIRV2 and that gene cluster sso2386 and sso2387 is involved in the secretion of the receptor components.

Except in Escherichia coli, very few virus receptors are known in the domains of Bacteria and Archaea (29). The primary receptors for E. coli filamentous phages are pili which retract toward the cell surface, bringing the phages to the secondary receptor located in the periplasm (30). Linear archaeal viruses, including rudiviruses, have been observed to attach to pili (5, 31, 32). Future work is needed to determine the association of the two identified gene

FIG 3 Different mutations in the SIRV2-resistant strains and their stability. (A) Transposon insertions in sso3139 and sso3140. (B) Mutations in sso2386 and sso2387. (C) PCR amplification of the mutation region from different resistant strains.
clusters with the structure of pili. To our knowledge, this is the first work providing genetic and biochemical evidence for a possible receptor system in archaeal virus entry.

ACKNOWLEDGMENTS

We thank Roger A. Garrett for critically reading the manuscript. This work is supported by a European Union FP7 grant (265933).

REFERENCES


