EXPRESSION, PURIFICATION AND PRELIMINARY FUNCTIONAL CHARACTERIZATION OF THE *ACIDIANUS* TWO-TAILED VIRUS TnpB ENDONUCLEASES

A Thesis Submitted to

The Graduate School of

İzmir Institute of Technology

in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

In Molecular Biology and Genetics

by

Arwa Saleh BURGEIA

July 2024

İzmir

We approve the thesis of **Arwa Saleh BURGEIA**

Examining Committee Members:

Assist. Prof. Dr. Muse OKE

Department of Molecular Biology and Genetics, İzmir Institute of Technology

---

Assoc. Prof. Dr. Ergin SAHIN

Department of Biology, Dokuz Eylül Üniversitesi

Assist. Prof. Dr. Hümeyra TAŞKENT

Department of Bioengineering, İzmir Institute of Technology

08 July 2024

Assist. Prof. Dr. Muse OKE

Supervisor, Department of Molecular Biology and Genetics, İzmir Institute of Technology

---

Prof. Dr. Özden YALÇIN ÖZUYSAL

Head of the Department of

Molecular Biology and Genetics

Prof. Dr. Mehtap EANES

Dean of the Graduate School

ACKNOWLEDGMENT

I would like to start by thanking my supervisor Assist. Prof. Dr. Muse OKE for his continuous support and patience throughout my MSc journey. I would like to express my appreciation to him for everything he has taught me in the laboratory and outside. It was a great honor and privilege to have had the chance to work closely with him.

I would also like to thank TÜBITAK for funding this project, and YTB for funding my studies.

I would like to thank all of the undergraduate and graduate students who have worked on this project hand in hand with me. Special thanks and appreciation to Ceren and Elif, who have constantly showed up and dedicated much of their time and energy to this project.

I am also very thankful for the constant support I have received from my friends Gamze, Ersa, Marry, Şüheda and Azra.

Last but not least, my endless appreciation goes to my mother, Salma, for her continuous financial, emotional and academic support. I would like to also thank my sisters Ala, Esra and Elaf and my aunt Samar and her daughters Leenah and Tala for their support.

ABSTRACT

EXPRESSION, PURIFICATION AND PRELIMINARY FUNCTIONAL CHARACTERIZATION OF THE *ACIDIANUS* TWO-TAILED VIRUS TnpB ENDONUCLEASES

TnpB is an RNA-guided endonuclease encoded in the IS200/605 transposon family. It forms a complex with a self-encoded ωRNA that directs TnpB to an appropriate target DNA sequence for activity. DNA cleavage at the target site promotes homologous recombination thus ensuring propagation of IS200/605 elements. Although IS200/605 elements are abundant in prokaryotic genomes, they are rarely found in viruses. However, the *Acidianus* two-tailed virus (ATV), which possesses 72 genes, encodes four TnpB proteins (gp10, gp40, gp43 and gp68). In this study, bioinformatics analysis of all ATV *tnpB* genes demonstrated that they were all solo *tnpB* genes and therefore could be classified as part of the IS1341 group. TnpB proteins are characterized by having a DED catalytic site motif and a C-terminal C4 zinc finger motif. Although the ATV gp10 protein retains the DED active site motif found in most TnpB proteins, the other ATV TnpB proteins possess a different amino acid instead of the Glu residue. Additionally, the ATV TnpB proteins possess the C4 zinc finger motif except for ATV gp40, which possesses three cysteine residues. All four ATV tnpB genes were cloned and expressed in the heterologous *Escherichia coli* host. The gp40 and gp68 proteins were purified using Ni^{2+}-NTA and gel filtration chromatography. Although gp68 appears to bind to DNA, there is insufficient evidence for RNA binding. Cleavage assays revealed that gp68 demonstrated nonspecific DNA nickase and cleavage activities. The significance of this study and the broader implications regarding the possible role of TnpB in virus survival are discussed.

ÖZET

ACIDIANUS İKI KUYRUKLU VIRÜSÜ TnpB ENDONÜKLEAZLARININ EKSPRESYONU, SAFLAŞTIRILMASI VE ÖNCÜL FONKSIYONEL KARAKTERIZASYONU

TnpB, IS200/605 transpozon ailesinde kodlanan bir RNA kılavuzlu endonükleazdır. TnpB, kendisi tarafından kodlanan bir ωRNA ile kompleks oluşturur ve bu RNA, TnpB'yi aktivite için uygun hedef DNA dizisine yönlendirir. Hedef DNA'nın bölgesinin kesilmesi homolog rekombinasyonu teşvik eder ve böylece IS200/605 elementlerinin yayılmasını sağlar. IS200/605 elementleri prokaryot genomlarında yaygın olsa da virüslerde nadiren bulunur. Bununla birlikte, 72 geni bulunan *Acidianus* iki kuyruklu virüsü (ATV), dört TnpB proteini kodlar (gp10, gp40, gp43 ve gp68). Bu çalışmada, tüm ATV *tnpB* genlerinin biyoinformatik analizi, bunların yalnızca *tnpB* genleri olduğunu gösterdi ve dolayısıyla IS1341 grubunun bir parçası olarak sınıflandırılabildi. TnpB proteinleri, bir DED katalitik bölge motifine ve bir C-terminal C4 çinko parmak motifine sahiptir. ATV gp10 proteini çoğu TnpB proteini gibi DED aktif bölge motifini korurken, diğer ATV TnpB proteinleri Glu kalıntısının yerine farklı bir amino asit taşır. Ayrıca, ATV TnpB proteinleri ATV gp40 dışında C4 çinko parmak motifine sahiptir; ATV gp40 üç sistein kalıntısına sahiptir. Dört ATV *tnpB* geni de heterolog *Escherichia coli* konak hücresinde klonlandı ve ekprese edildi. Gp40 ve gp68 proteinleri Ni2+-NTA ve jel filtrasyon kromatografisi kullanılarak saflaştırıldı. Gp68'in DNA'ya bağlandığı, ancak RNA bağlamak için yeterli kanıt olmadığı görüldü. Kesim deneyleri, gp68'in spesifik olmayan DNA nikaz ve kesim aktiviteleri gösterdi. Bu çalışmanın önemi ve TnpB'nin virüslerin hayatta kalmasındaki olası rolü ile ilgili geniş çaplı etkileri tartışılmıştır.

Dedicated to my mother,

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1. IS200/605 Transposon Family

Insertion sequence (IS) elements occur exclusively in genomes of archaea and bacteria and are among the most abundant mobile genetic elements in nature. They typically encode a transposase that recognizes inverted terminal repeats at both ends of the element and mobilizes the element via double-stranded DNA intermediates. Insertion of the element at the target site usually results in duplication at the target site. These elements are currently grouped into 28 families primarily on the basis of the catalytic residues of their transposases, architecture and homology of entire elements, and transposition mechanisms (Siguier et al., 2015). One of the most studied IS families is the IS200/605 family, which consists of 153 members. This family differs from other IS families in that they transpose through single-stranded DNA intermediates, possess sub-terminal left end (LE) and right end (RE) palindromic elements, and are inserted into the target site without duplicating the target site (He, et al. 2015).

1.1.1. Classification of the IS200/605 family

Although the IS200/605 transposon family comprises 153 distinct members, they are classified into three main groups based on the presence and organization of *tnpA* and *tnpB* genes. The *tnpA* gene product functions as the transposase and therefore catalyzes mobilization (He, et al. 2015). The *tnpB* gene product was initially proposed to regulate transposition (Pasternak et al., 2013), but recent experimental evidence suggests that the protein may function to prevent permanent loss of the IS (discussed in more detail below). The three main groups of the IS200/605 and their general features are discussed next.

A. IS200 group

This group encodes only for the *tnpA* gene. This group is mainly found in Gram-positive and Gram-negative eubacteria and in some archaea. The TnpA protein sequence alignment from different group members indicates that TnpA is conserved with some variation in the sequence and length of the C-terminus. *Salmonella typhimurium* is the founding member, but the group also includes *Escherichia*, *Shigella* and *Yersinia* species (He, et al. 2015).

B. IS605 group

This group encodes both *tnpA* and *tnpB*, and therefore, it tends to be longer in sequence. The TnpA copies of this group carry shorter C-terminal tail but are not a separate clade from the IS200 group. The *tnpA* and *tnpB* genes tend to exhibit different configurations in respect to each other (He, et al. 2015). As shown in Figure 1.1 they might be divergent (IS605, IS606), partially overlapping (IS608, ISDra2), or separate (ISSCpe2, ISEfa4).

Moreover, IS605 along with IS606 and IS608 have been identified in different strains of *Helicobacter pylori* where the IS605 has been found to be involved in the genomic rearrangements. However, the most characterized member of this group is the ISDra2 from *Deinococcus radiodurans* where *tnpA* and *tnpB* genes are found overlapping which will be discussed in more detail in the next section (Kersulyte et al., 1998).

C. IS1341 group

This group encodes only *tnpB* and can be found in multiple copies such as in the thermophilic bacterium PS3, where three copies are present. It is still unclear whether members of this group are autonomous IS elements or require TnpA from other IS elements in the same cell for transposition (He, et al. 2015).

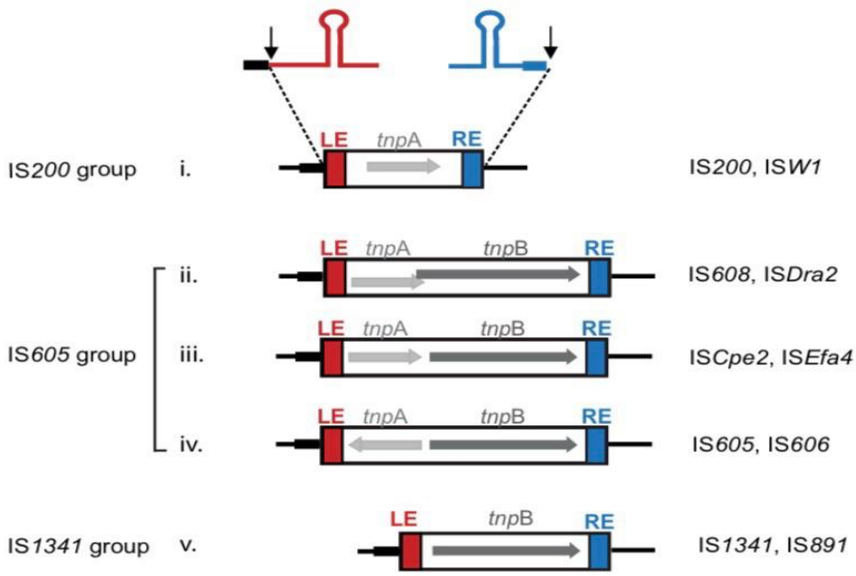


Figure 1.1. The IS200/605 transposon family organization. (i) IS200 group that contains *tnpA* gene only. (ii-iv) IS605 group that contains both *tnpA* and *tnpB* in their different organizations. (v) IS1341 group that only contains tnpB (Source: He, et al. 2015)

1.1.2. Single-stranded 'Peel and Paste' Mechanism

As mentioned above, the IS ends of IS200/605 family consist of palindromic repeat and not inverted repeats which are found in most IS elements. The transposase binds to the subterminal right-end (RE) and left-end (LE) palindromic repeats that from a hairpin structure. The binding takes place only on the top strand of the transposon and hence a single strand would be transposed. Another characteristic of this family is that it has target specificity. Most IS elements insert randomly whereas IS200/605 members insert into the 3' end of a well-defined tetra- or pentanucleotide sequence (Kersulyte et al., 2002).

D.radiodurans ISDra2 of the IS200/605 family is perhaps the best studied member of this family. It contains both *tnpA* and *tnpB* that overlap. It is flanked by the LE and RE elements. TnpA contains 140 amino acids and is an extremely small protein. It is classified as a Y1 transposase of the HUH family that contain one or two active site tyrosine residues which are the nucleophiles during DNA cleavage. It also contains a

divalent metal cofactor (Hickman et al., 2010). The peel and paste mechanism in ISDra2 is coupled with DNA replication and does not require the duplication of the target site and works as follows: TnpA excises the transposon form the lagging strand near 5' TTGAT sequence which is the pentanucleotide targeted by TnpA. This causes the formation of a circular single-stranded DNA intermediate. This is then inserted 3' of the TTGAT sequence on the acceptor joint (target) as shown in Figure 1.2a. The IS608 from H. Pylori follows the same mechanism but uses the 5'-TTAC which is a tetranucleotide sequence for recognition. (Hickman et al., 2010). Consequently, the role of TnpA is well defined.

The role of TnpB, which consists of 408 amino acids, was not well established because previous experiments demonstrated that it was dispensable for transposition, although it might possess a regulatory role (Barabas, et al. 2008). However, recent experiments have shown that TnpB functions as an RNA-guided endonuclease. Specifically, TnpB introduces a double-stranded DNA break at the donor site, which triggers homology-directed repair (Figure 1.2b). Thus, TnpB triggers the homologous recombination mechanism that restores the transposon back to the original site and therefore retains the transposon. TnpB is guided to the target DNA sequence at the donor site by a non-coding RNA sequence referred to as omega RNA (ωRNA). Subsequently, TnpB recognizes a specific sequence upstream of the target sequence at the donor site referred to as the transposon-adjacent motif (TAM) and then catalyzes a double-stranded DNA break upstream the TAM sequence (Altae-Tran et al., 2021; Karvelis et al., 2021). Details of the requirement for ωRNA and TAM in the catalytic mechanism of TnpB are discussed next.

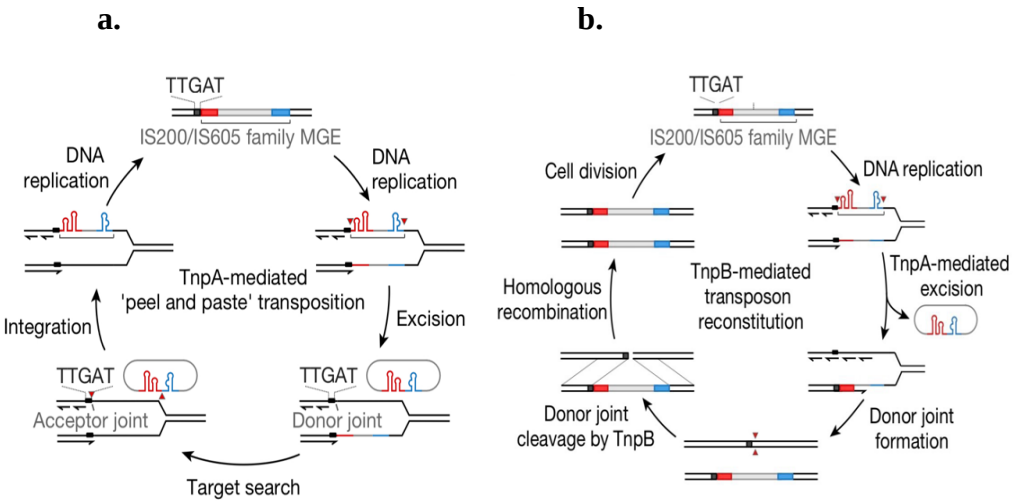


Figure 1.2. (a) TnpA-mediated 'peel and paste' transposition. (b) A proposed function for TnpB in the transposition process which results in the retention of the transposon (Source: Karvelis, et al. 2021)

1.2. Catalytic mechanism of TnpB

The ωRNA sequence is encoded in the same locus as *tnpB* gene. Its 5' end is located at the 3' end of the tnpB gene and its 3' end is located within the right end element of the transposon. The ωRNA sequence can be divided into two parts: a 5'-end 150-nt sequence referred to as the scaffold sequence, and a 3'-end 16-nt sequence referred to as the guide sequence. Expression of both TnpB and ωRNA results in the formation of the TnpB-ωRNA complex. TnpB cleaves its own mRNA at the 5' end to form the mature ωRNA (Nety et al., 2023). Following TnpA-mediated transposition, the TnpB-ωRNA complex binds to the target DNA sequence primarily by sequence complementarity of the guide sequence and the target DNA sequence. Upon recognition of the TAM sequence located 5' to the target DNA sequence, TnpB cleaves the target DNA sequence in a reaction catalyzed by the catalytic motif situated in its active site. The TAM sequence is identical to the DNA sequence at the left end of the transposon and is typically AT-rich. The TnpB catalytic motif consists of one glutamic acid and two aspartic acid residues, thus forming the so-called DED motif (Karvelis, et al. 2021, Altae-Tran, et al. 2021). These active site residues are situated in the RuvC domain of TnpB as exemplified by the ISDra2 transposon shown in Figure 1.3a.

Two cryo-EM structures of the ISDra2 TnpB-ωRNA-target DNA complexes have been determined (Nakagawa et al., 2023; Sasnauskas et al., 2023). These structures reveal mechanistic details of the macromolecular interactions and provide valuable insights into the catalytic mechanism. As shown in Figure 1.3b, the overall architecture of ISDra2 TnpB consists of two lobes referred to as the recognition (REC) and nuclease (NUC) lobes. The REC lobe is further divided into the wedge (WED) (residues 1 to 12 and 117 to 176) and REC (residues 13 to 116) domains and the NUC lobe comprises the RuvC (residues 185 to 326 and 360 to 375) and target nucleic acid binding (TNB; residues 327 to 359) domains. The WED domain has an oligonucleotide-binding fold and consists of seven-stranded β-barrel that is flanked by an α-helix. The REC domain is composed of four α-helices and inserted between the β1 and β2 strands of the WED domain. The RuvC domain, which possesses the RNaseH fold, consists of a five-stranded mixed β-sheet flanked by four α-helices. The TNB domain is located between the β5 strand and α4 helix and contains the C4-type zinc finger (ZF) motif. Although most TnpB proteins contain the C4-ZF motif, the ZF motifs are either absent or degenerate in some TnpB proteins, thus suggesting that they play supplementary roles such as improving protein stability (Altae-Tran et al., 2023).

The cryo-EM structures also revealed the nature of the interactions between TnpB and the nucleic acids. The binding sites of the nucleic acids are organized such that the TAM sequence binding site is located between the WED and REC domains, and the ωRNA-target DNA heteroduplex is located between a central channel formed by the REC and RuvC domains. TnpB forms a complex with ωRNA after processing the RNA molecule as described above (Figure 1.3c). The TNB domain facilitates the unwinding of DNA and loading into the RuvC active.

Most of the TnpB characterization studies have focused on the bacterial systems. However, a recent study reported characterization of TnpB from the acidothermophilic archaeon, *Sulfolobus islandicus* REY15A (SisTnpB1; Xu et al., 2023). In S.*islandicus*, *tnpB* occurs either alone (solo) or associated with *tnpA*. Relative to the ISDra2 tnpB, they have been found to from a distinct branch, thus indicating their evolutionary differences. A conserved AT-rich TAM sequence has been identified at the 5’ left end of the IS elements consisting of both *tnpA* and *tnpB* and the solo *tnpB*. The ωRNA has been identified and was found to be conserved except for the last 20 nucleotides at the 3’ end

which are predicted to be the guide sequence. Furthermore, it has been shown that SisTnpB1 can cleave both dsDNA and ssDNA (Xu, et al. 2023).

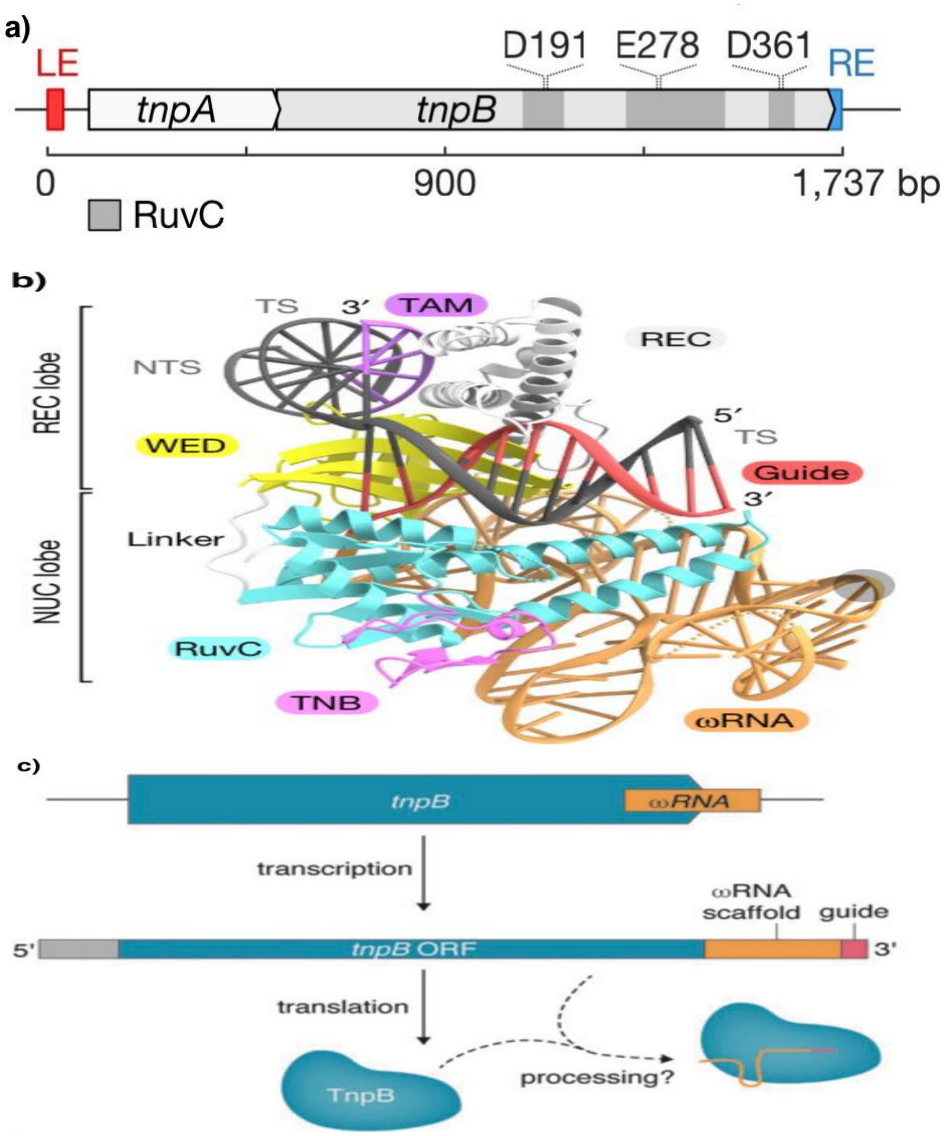


Figure 1.3. a) The ISDra2 transposon, which consists of *tnpA* and *tnpB* genes and flanked by LE and RE partially palindromic sequences. b) Structure of ISDra2 TnpB protein. c) TnpB processing its own ωRNA to yield an active RNP complex (Source: Nety, et al. 2023 & Nakagawa et al., 2023)

1.2.1. Relationship with CRISPR-Cas System

TnpB have been classified under a new class of RNA-guided systems named OMEGA (Obligate Mobile Element-guided Activity). This class is composed of an RNA-

guided endonuclease protein such as TnpB or IscB and a noncoding RNA named the ωRNA. Bioinformatics analyses had suggested that the Cas9 effector protein of the type II CRISPR-Cas system may have evolved from IscB. Similarly, it has also been proposed that the type V CRISPR-Cas effector proteins such as Cas12 may have evolved from TnpB (Chylinski et al., 2014). Further evidence to support the predicted evolutionary relationships are now available due to the similarities between the structures and functions of the Cas effector proteins and the OMEGA proteins. Specifically, for TnpB and Cas12, the organization of the primary and tertiary structures are similar (Figure 1.4). The RuvC domains are homologous, and the REC/WED domain in Cas12 that recognizes the PAM (protospacer adjacent motif) on the target DNA is also found in ISDra2 TnpB where it is used to recognize the TAM sequence. The PAM sequence is required for Cas12 family proteins to initiate target recognition and R-loop formation which is equivalent to the TAM sequence required by TnpB proteins (Nakagawa et al., 2023). It has also been shown that TnpB is most similar to Cas12f and least similar to Cas12a due to the accumulation of many insertions in the protein. Figure 1.4 shows a comparison between ISDra2 TnpB, Cas12f and Cas12a proteins in regard to the domain organization, protein structure and guide RNA structures (Sasnauskas et al., 2023). As seen in Figure 1.4, the Cas12f guide RNA stem 5 is equivalent to stem 2 in the TnpB ωRNA. The stem 5 is formed by the 3'end of tracrRNA and 5'end of crRNA which can be linked to form the guide RNA in CRISPR systems. It can be seen that this link is present naturally in TnpB, which might suggest that the guide RNA got split into tracrRNA and crRNA. A proposed reason for this event taking place is that this would allow the CRISPR array to be able to accommodate different crRNAs (Sasnauskas et al., 2023).

Nevertheless, differences are present between TnpB and Cas12. For example, TnpB is monomer and requires a single guide RNA in contrast to Cas12 dimer that binds to crRAN-tracrRNA complex. TnpB protein are about 400 amino acids in size which is much smaller than Cas12 proteins that are about 1500 amino acids. (Siguier, et al. 2014, Isabel, et al. 2024). These features make TnpB a more attractive alternative for genome editing purposes.

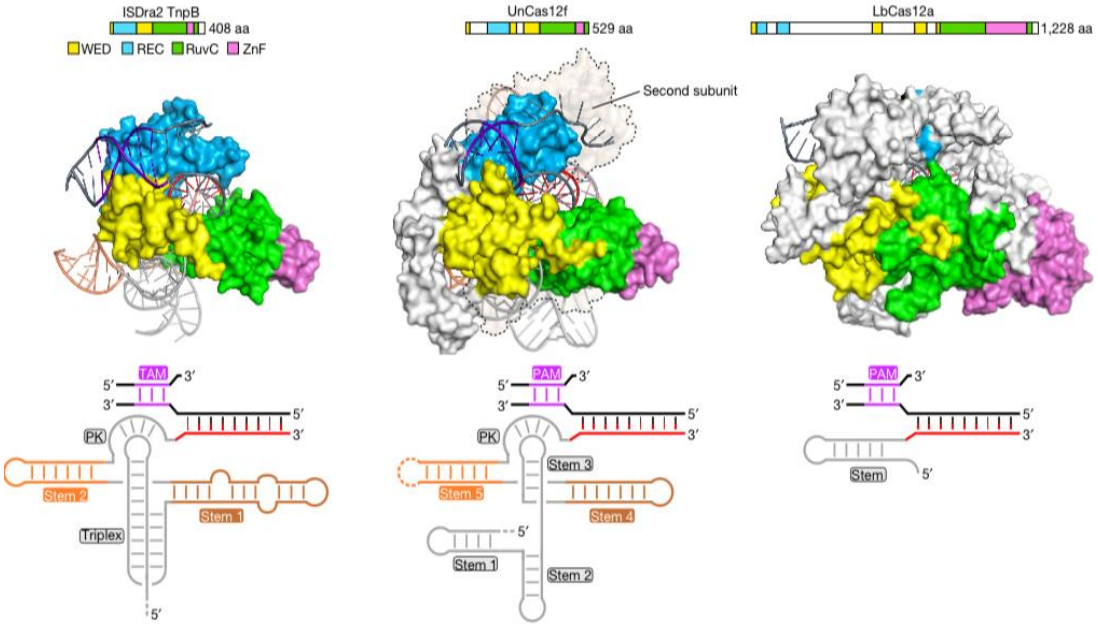


Figure 1.4. Shows the domain organization, protein structure and guide RNA of TnpB, Cas12f and Cas12a proteins (Source: Sasnauskas et al., 2023)

1.2.2. Relationship with eukaryotic transposon-encoded Fanzor (Fz) proteins

Fanzor proteins are encoded by transposable elements and are widespread in eukaryotes and in large double-stranded viruses that infect eukaryotes. They are TnpB-like proteins and share conserved amino acids residues with TnpB from the center of the protein to the C-terminus. Two groups have been identified, Fanzor1 which is encoded by at least five different superfamilies Mariner, Sola, IS4, Helitron and MuD, and Fanzor 2 which is encoded by IS-607 type transposons (Bao and Jurka, 2013). Fanzor proteins are believed to have evolved from TnpB proteins through horizontal gene transfer due to the presence of prokaryotic-eukaryotic symbiotic hosts. They are also believed to have spread with the help of eukaryotic viruses and eukaryotic symbionts. This has been demonstrated by the presence of Fanzors in both the viruses and their potential algae or mollusk host. The exact role of the Fanzor RNA-guided endonuclease activity remains unknown, however, some Fanzors have been reported to co-occur with TnpA-like HUH

endonuclease transposase. This suggests that they might play a role in the peel & paste mechanism described in section 1.1.2. Due to its small size and eukaryotic origins, Fanzors are believed to have potential in genome editing (Saito et al., 2023).

1.2.3. Genomic Editing Potential

The cleavage activity of ISDra2 TnpB has not only been demonstrated in vitro, but also in *E.coli* cells and in human cells. Using plasmid interference assays in *E.coli*, it was demonstrated that the TnpB RNP complex was able to cleave TAM flanked DNA targets on plasmids inside *E.coli* cells (Karvelis, et al. 2021). Moreover, SisTnpB1 has demonstrated very high efficiency genomic editing ability (Xu, et al. 2023). Moreover, the same TnpB was tested in human cells (HEK293T). It showed that TnpB had introduced mutations at the target sites, and hence, is able to cleave eukaryotic gDNA (Karvelis, et al. 2021).

1.3. *Acidianus* Two Tailed Virus (ATV)

1.3.1. General Features and Life Cycle

ATV is a member of the *Bicaudaviridae* family of viruses that infect hyperthermophilic archaea. They were isolated from hot acidic springs, with temperatures ranging from 87-93°C and pH of 1.5-2. ATV is unique because it can undergo morphological changes independently and outside of the host cell (extra cellular morphogenesis). After the lemon-shaped tail-less virions are extruded from the host cells, they start to develop their tails extracellularly as shown in Figure 1.5. This event occurs at temperatures close to 85°C, which is the temperature of their natural habitat (Prangishvili, et al. 2006).

The life cycle of ATV can be either lytic or lysogenic. In the lytic cycle, the virus replicates inside the host cell and then cause cell lysis, whereas in the lysogenic cycle, the viral genome is integrated into the host chromosome by the virus-encoded integrase, and

this leads to conversion of the infected cell into a lysogen. When incubating the virus with its hyperthermophilic archaea host at 85°C, it was observed that the host is converted to the lysogen. On the other hand, when the temperature was decreased to 75°C, cell lysis takes place.

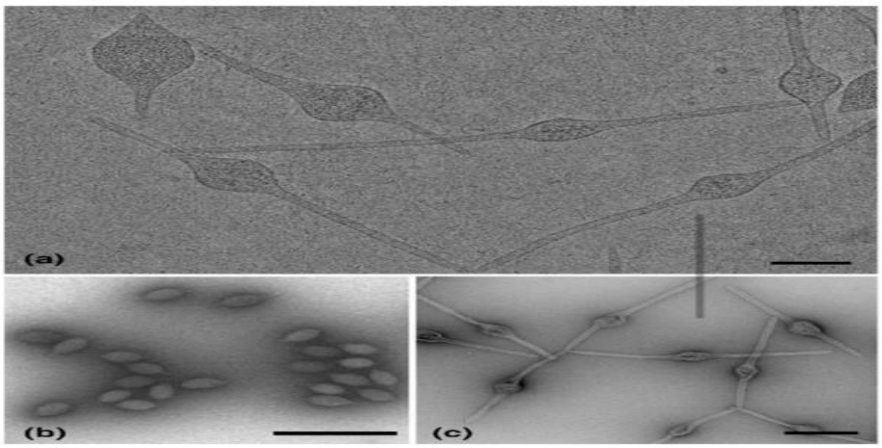


Figure 1.5. (a) Showing the ATV tail development at different stages. (b) showing the tailless visions. (c) showing the tailed visions (Source: Prangishvili, et al. 2006)

1.3.2. Genomic Properties

ATV has a circular dsDNA genome of 62,730 bp. In comparison to other thermoacidophilic viruses, ATV has a high G+C content (41.2%). The virus was predicted to encode 72 proteins and carry four putative transposable elements. It encodes 11 major structural proteins, two of which have been structurally characterized and are found to display unique folds not observed in other viruses (Prangishvili, et al. 2006). Most of the ORFs have not been annotated yet due to lack of similarity with sequences in public sequence databases. In comparison to crenarchaeal viruses, an exceptional characteristic for ATV is that it encodes four putative TnpB proteins (gp10, gp40, gp43 and gp68). The presence of these genes is extraordinary because it has never been observed in crenarchaeal viruses but rather in crenarchaeal conjugative plasmids (Prangishvili, et al. 2006). Thus, it would be interesting to determine if the putative ATV TnpB proteins function in a similar manner to their cellular counterparts.

1.4. Aim of the Study

This study aims to understand the role of the putative ATV TnpB proteins in the life cycle of the virus. Specifically, it aims to determine if features of the recently characterized OMEGA systems i.e. TnpB-mediated processing of ωRNA, formation of TnpB-ωRNA complex and TnpB-mediated cleavage of double-stranded DNA is retained by the ATV TnpB proteins. By so doing, this work would provide experimental evidence for several ATV proteins and therefore, provide much needed information for the poorly annotated ATV genome. It ıs anticipated that this work will also provide foundation for further studies and the potential adaptation of the ATV TnpB proteins for biotechnological applications particularly due to their anticipated hyperthermophilic properties. In this study, the ATV TnpB proteins will be analyzed using bioinformatics tools, and the putative *tnpB* and ωRNA genes will be cloned and expressed in the heterologous *E. coli* host. Functional assays will also be carried out to determine if the TnpB-ωRNA complexes possess DNA cleavage activities.

CHAPTER 2

MATERIALS AND METHOD

2.1. Bioinformatics analysis

ATV putative *tnpB* DNA and protein sequences were downloaded from NCBI GenBank (Accession number: NC_007409.1). BLAST search of *tnpB* genes and proteins were done using NCBI DNA and protein BLAST search. DNA sequence alignment was performed using NCBI blastn and protein sequence alignments were done using NCBI blastp and Uniprot. In order to predict the ATV TnpB ωRNA sequences, a sequence alignment of all *tnpB* sequences starting from -250 nucleotides upstream of the stop codon until the end of the integenic region was carried out. Also, RNAFold was used in order to check for hairpin structures. Prediction of TAM sequences involved sequence alignment of the 5' intergenic region of all four ATV TnpB proteins, and the first four to five uninterrupted conserved bases of the aligned sequences were considered the TAM sequences.

2.2. Design of TnpB expression constructs

Genes encoding the four ATV TnpB proteins (GenBank accession numbers: CAI59885.1 for gp10; CAI59895.1 for gp40; CAI59866.1 for gp43; CAI59875.1 for gp68) were ordered as synthetic gene constructs from GenScript. The gene constructs were designed to incorporate DNA coding sequences for maltose binding protein (MBP), A 10× histidine tag (His10) and a tobacco etch virus (TEV) cleavage sequence, with the final synthetic construct occurring from the 5' end as follows: MBP - His10 - TEV protease cleavage site – specific *tnpB* gene. The constructs were cloned into the *BamHI* and *NdeI* restriction enzyme sites of the pET42b vector. A second set of gene constructs were designed as follows: 5'-His10-*tnpB* gene-entire 3' intergenic region, and these constructs were cloned as *NdeI-XhoI* inserts into the pET30a(+) expression plasmid. Details of both sets of expression constructs are presented in Appendix C.

2.3. Transformation and glycerol stock preparation

After preparing the plasmids according to the company guidelines, the recombinant pET42b(+) plasmids were transformed into both *Escherichia coli* DH5α and *E. coli* BL21(DE3) cells and the recombinant pET30a(+) cells were transformed into DH5α and BL21-A1 cells. The transformation procedure was performed as follows: 2µL of plasmid DNA was mixed with 50µL of component cells by gently tapping the bottom of the Eppendorf tube. The mixture was incubated on ice for 30 minutes. Afterwards, heat shock treatment was performed by placing the mixture in a preheated 42°C water bath for 45 seconds. The mixture was immediately placed on ice for 3 minutes and subsequently, 200µL LB media was added to the mixture. The transformation mixture was incubated at 37°C, 200 rpm for 1 hour. 50µL of the transformation mixture was plated on one LB agar plate (supplemented with 50mg/mL kanamycin) and the remaining 200µL was plated on another. Plates were incubated overnight at 37°C and colonies were observed the next day. For preparation of 50% glycerol stocks, distinct colonies were picked from the plates and used to inoculate 5 mL LB (supplemented with 50mg/mL kanamycin) and incubated overnight at 37°C, 200 rpm. Subsequently, 500µL overnight culture was mixed with 500µL glycerol, frozen in liquid nitrogen and then stored at -80°C.

2.4. Plasmid verification

The recombinant pET30a(+) vector was subjected to restriction digest analysis to verify that the insert was cloned into the plasmid. Distinct *E. coli* DH5α colonies from LB agar plates were used to inoculate 5mL LB and 50mg/mL kanamycin and incubated at 37°C overnight. Cells were harvested by centrifugation at 5000 rpm, 4°C for 20 minutes. Plasmids were purified using GeneJet Plasmid Miniprep Kit from Thermo Fisher and then subjected to restriction digest. The restriction digest reaction was prepared by mixing 2µL of enzyme, 2µL of DNA, 5µL of rCutSmart buffer and 49µL of nuclease free water. The reaction was placed in the water bath at 37°C 1hour. The restriction enzymes used for pET30a(+) were *NdeI* and *XhoI*. Subsequently, 2× loading dye was added to the reaction

mixtures and loaded onto 1.5% agarose gel supplemented with 0.5µg/mL ethidium bromide. Electrophoresis was performed at 100V for 30 minutes. The gels were then visualized using UV transilluminator.

2.5. Protein expression

Starter cultures were prepared by picking cells from the glycerol stocks (BL21 (DE3) and BL21-A1) and adding them to 5mL LB with 50mg/mL kanamycin. The starter culture was incubated at 37°C, 200 rpm overnight. 1% of the starter culture was inoculated into 5mL (for small scale) or 500mL (for large scale) terrific broth media. The cultures were grown at 37°C, 200 rpm until the absorbance at 600 nm was between 0.6 and 0.8, at which point 0.2mM IPTG concentration was added for induction of protein expression. The temperature was decreased to 16°C and the cultures were incubated overnight. The cells where then harvest by centrifugation at 5000 rpm for 20 minutes and stored at -80°C for later use.

2.6. Protein purification

Frozen cell pellets were resuspended in lysis buffer (binding buffer: 50mM Tris pH 8, 500mM NaCl, 40mM Imidazole, 5% glycerol and 5mM b-mercaptoetdanol supplemented with a tablet of protease inhibitor cocktail) and mixed thoroughly to ensure homogeneity. Resuspended cells were lysed by sonication at a frequency of 20kHz for 30 seconds (x4 cycles). A 30 second break was given between each cycle in order to ensure that no overheating or foaming occurs. The lysate was then centrifuged at 18,000 rpm, 4°C to separate the soluble from the insoluble fraction. Proteins from the small-scale trials were purified using the Ni^{2+}-NTA affinity chromatography and proteins from large scale expression were purified using Ni^{2+}-NTA affinity chromatography followed by gel filtration chromatography. The Ni^{2+}-NTA purification buffers were prepared as follows: binding buffer (50mM Tris pH 8, 500mM NaCl, 40mM Imidazole, 5% glycerol and 5mM

b-mercaptoetdanol), elution buffer (50mM Tris pH 8, 500mM NaCl, 500mM Imidazole, 5% glycerol and 5mM β-mercaptoetdanol).

For small scale Ni^{2+}-NTA purification, the soluble fraction was mixed with 1mL Nickel beads for 30 minutes. The sample was then added to a bench column and the flow through (FT) is collected. The column is then washed twice with 500µL of binding buffer and then eluted at least three times with 150µL of elution buffer. For the MBP - His10 - TEV protease cleavage site – *tnpB* constructs, the buffers were exchanged into TEV cleavage buffer (200mM Tris, 500mM NaCl and 10% glycerol) overnight and samples were incubated with 1 unit of TEV protease.

Large scale Ni^{2+}-NTA purification was performed using ÄKTAprime fast protein liquid chromatography (FPLC) system. The sample was loaded at 0.5 mL/min onto a 1 mL HisTrap FF (Cytiva) column and eluted at 1 mL/min in a 10 mL gradient. The eluted fractions that corresponded to the peak on the chromatogram were pooled and loaded onto HiLoad 16/600 superdex 75pg (Cytiva) gel filtration (GF) column equilibrated in buffer containing 10mM Tris pH 8, 200mM NaCl, 5mM DTT and 10% glycerol. Peak fractions from the GF chromatogram were pooled and then concentrated using the Amicon® Ultra 3K device, aliquoted and stored at -80°C.

All purification samples were mixed with 2X Laemmli sample buffer, heated at 95°C for ten minutes, centrifuged for 10 minutes loaded onto 10% SDS PAGE gel. The gels were stained with Coomassie blue for 30 minutes and visualized under white light.

2.7. Target DNA design

The predicted TAM and target DNA sequences (refer to Appendix C for sequences) were sent to GenScript for cloning into pUC19 plasmid. The plasmid with the insert (hereafter referred to as pTAM) and an empty pUC19 vector were transformed into DH5α cells and the plasmids were purified with GeneJet Plasmid Miniprep Kit using manufacturer's instructions. The purified plasmid concentrations were determined using nanodrop spectrophotometer and the plasmids were then stored in -20°C.

2.8. Cleavage assays

For the cleavage assays, 5 nM RNA-protein complex was mixed with 5ng/µL plasmid in different buffers as detailed next. Buffer A consisted of 10 mM Tris-HCl buffer pH 7.5, 1 mM DTT, 1 mM EDTA, 100 mM NaCl; buffer B, 10 mM Tris-HCl buffer pH 7.5, 1 mM DTT, 10 mM $MgCl_2$, 100 mM NaCl; and buffer C, 10 mM Tris-HCl buffer pH 7.5, 1 mM DTT, 10 mM $MnCl_2$,100 mM NaCl. The reaction mixes were then divided into two and one half was incubated at 37°C and the other at 65°C in a water bath set at the appropriate temperature for 1 hour. The reactions were then stopped by adding 20 mM protease K and incubating at 37°C for 1 hour. Finally, 2× loading dye was added to the reaction mixtures and loaded onto 1.5% agarose gel. The gels were then visualized using UV transilluminator.

2.9. RNA isolation

Total RNA purification was performed using the GeneJet RNA purification kit from ThermoFisher. Starter cultures were prepared as previously described by picking cells from the glycerol stocks of BL21-A1 cells and adding them to 5mL LB with 50mg/ml kanamycin. The starter cultures were incubated at 37°C, 200 rpm overnight. 1% of the starter culture was inoculated into 5mL (for small scale). The cultures were grown at 37°C, 200 rpm until the absorbance at 600 nm was between 0.6 and 0.8, at which point 0.2mM IPTG concentration was added for induction of protein expression. The temperature was decreased to 16°C and the cultures were incubated overnight. The cells where then harvest by centrifugation at 5000 rpm for 20 minutes and the total RNA was purified using GeneJet RNA purification kit. The The elution RNA volume was 50µL, and elution was performed twice. Samples were then mixed with 2×loading dye and loaded onto 1.5% agarose gel. The gels were then visualized using UV transilluminator.

CHAPTER 3

RESULTS AND DISCUSSION

3.1. Bioinformatics analysis

3.1.1. Identification of the ATV TnpB proteins

There are currently two complete genome sequence entries (accession numbers: AJ888457 and NC_007409) for ATV on the NCBI genome database. All four ATV TnpB proteins are annotated as 'transposase' in the NC_007409 entry and as 'putative transposase' in the AJ888457 entry. Considering that the ATV genome was annotated over 18 years ago (Häring et al., 2005 & Prangishvili et al., 2006), and in light of the recent functional characterization of several TnpB proteins, an attempt was made to verify these annotations and to determine transposon family. First, a sequence alignment of the putative ATV proteins with the best-characterized TnpB endonuclease, the ISDra2 TnpB, was carried out to determine the conservation of TnpB domains. The sequence alignment revealed that although the putative ATV TnpB proteins share low sequence identities (ranging from 18.33% to 27.13%) with the ISDra2 TnpB, several key residues are conserved (Figure 3.1). Although the Asp residues of the catalytic DED motif (located in the RuvC domain), is conserved in all the ATV TnpB proteins, the Glu residue is only conserved in ATV gp10. Several studies have demonstrated that mutating any of these residues obliterates nuclease activity (Karvelis et al., 2021 & Meers et al., 2023). However, it has now been demonstrated that one of the catalytic residues can occur at an alternate position in several TnpBs and Fanzors, with the protein retaining nuclease activity (Jiang et al., 2023 & Altae-Tran et al., 2021)This rearrangement includes having an alternative glutamate located in a loop that faces the catalytic site or it could be mutated to another amino acid such as histidine. However, two or more mutations in the DED residues in the RuvC domain result in the inactivation of the protein (Wiegand et al., 2023). The observation here that none of the four ATV TnpB catalytic residues have two

or more mutated residues, it is tempting to speculate that the four ATV TnpB proteins could function as active nucleases.

From the sequence alignment it can be seen that the C4 Zinc Finger motif is conserved in all of the ATV TnpB proteins with the exception of ATV gp40, which possesses three cysteine residues instead of four. The role of the Zinc Finger motif in TnpB function is yet to be characterized. However, since several uncharacterized TnpB proteins lack this domain, it has been suggested that it is not involved in nuclease activity but instead may have a role in protein stabilization Altae-Tran et al., 2021).

Besides the highly conserved DED catalytic motif and the Zinc Finger motif, it was not possible to readily identify the other TnpB characteristic motifs using the ISDra2 sequence as template. Consequently, it was decided to examine the sequence homologies between all four ATV TnpB proteins and determine the origins of the ATV proteins particularly as TnpB has been determined to be prokaryotic in origin.

As shown in figure 3.2, The ATV TnpB proteins share between 17.82% and 86.16% sequence identity. From the identity matrix a very clear and interesting high sequence identity of 86.16% can be seen between gp43 and gp68. A nucleotide and amino acid sequence alignment of both proteins was performed as shown in figures 3.3 & 3.4. This suggests that both proteins might share a common ancestral gene. In order to determine if this represents a plausible explanation, a DNA sequence alignment of both gp43 and gp68 was carried out.

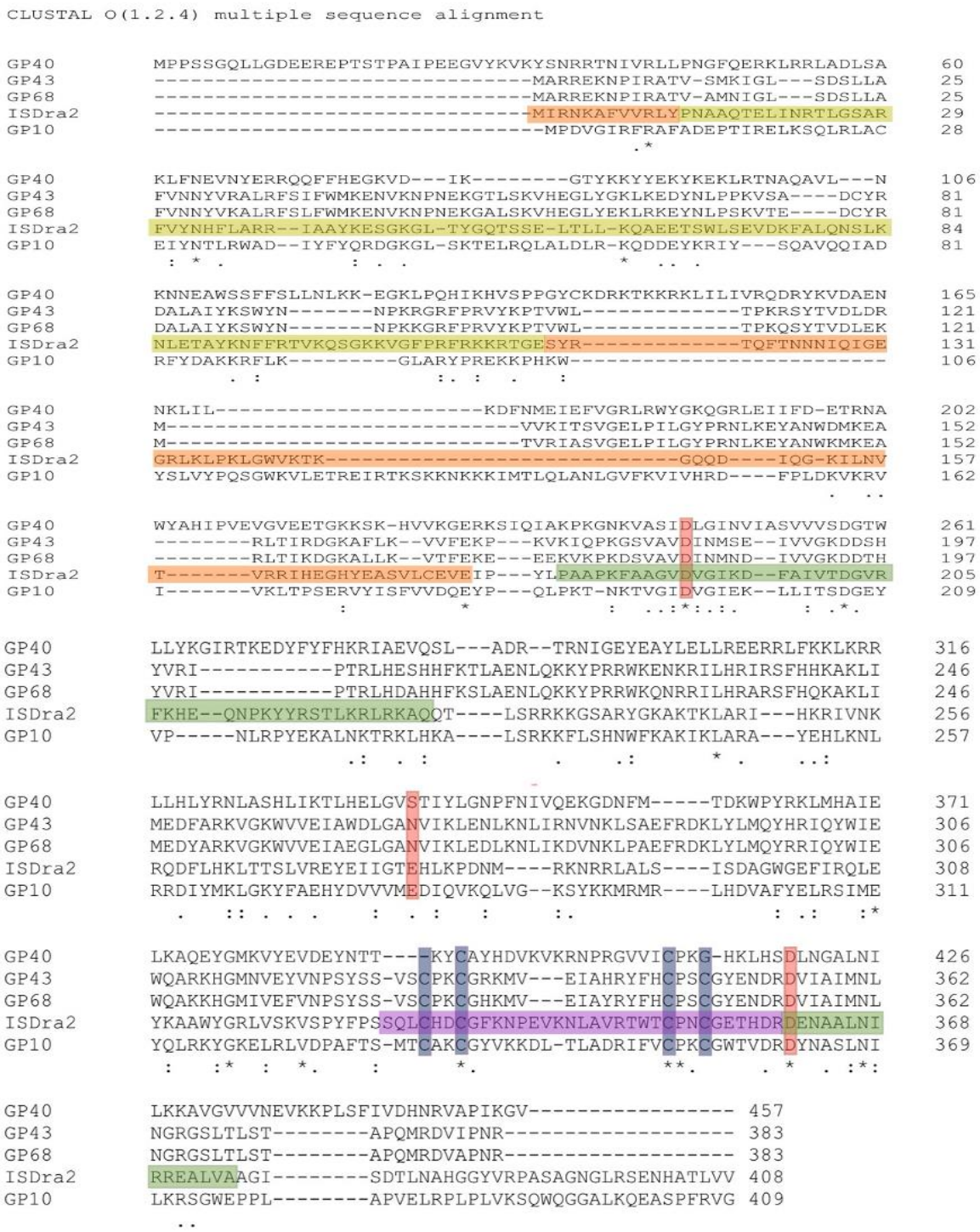


Figure 3.1. Protein sequence alignment of all four ATV TnpB proteins and ISDra2 TnpB protein using Uniprot. Active site residues (DED) are highlighted in red, the C4-type zinc finger motif in blue. The ISDra2 TnpB domains are coloured as follows: WED domain, orange: REC domain, yellow: RuvC domain in green and TNB domain in purple

Figure 3.2. Percent identity matrix for the ATV TnpB and ISDra2 TnpB sequences. The fourth column indicates the percent identity between each ATV TnpB protein and ISDra2 TnpB

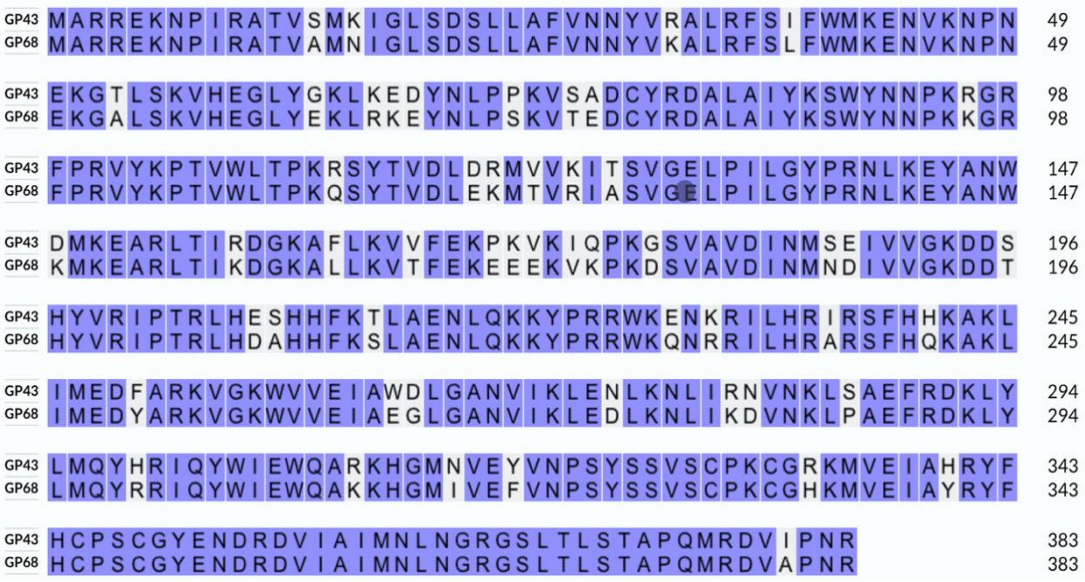


Figure 3.3. Sequence alignment using uniprot for ATV gp43 and gp68. The sequence of both proteins is highly identical (Conserved sequences are in purple)

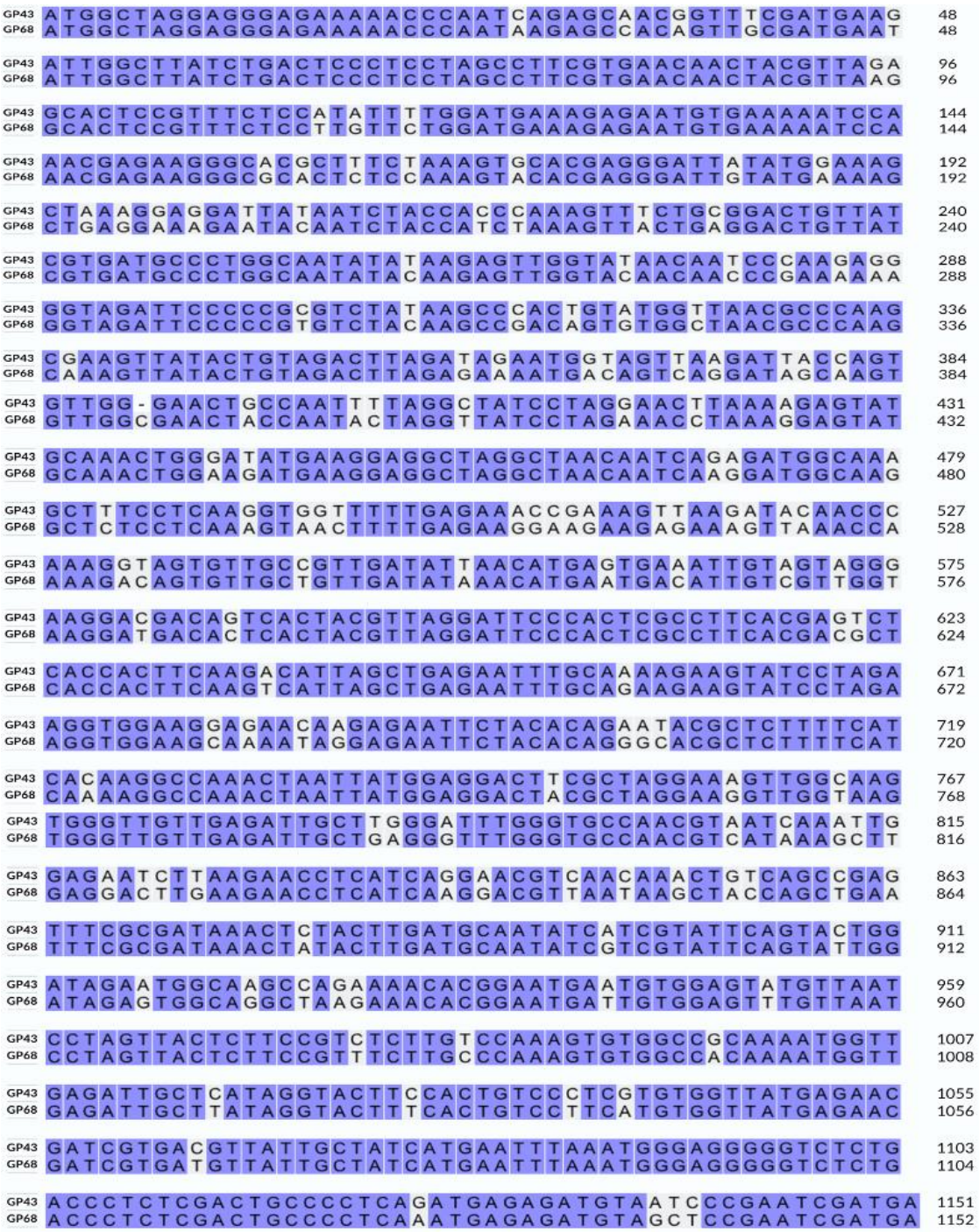


Figure 3.4. Nucleotide sequence alignment using uniprot for ATV gp43 and gp68. The sequence of both genes is highly identical (Conserved sequences are in purple)

3.1.2. Origins of the ATV TnpB proteins

The vast majority of TnpB proteins have been identified in prokaryotes and more recently in eukaryotes. Few occurrences have been observed in viruses, most notably in the viruses of unicellular eukaryotes. This observation suggests that TnpB is of prokaryotic origin. Phylogenetic analysis revealed an evolutionary link between the TnpB proteins of these viruses and those of prokaryotes and the eukaryotic Fanzors (Altae-Tran et al., 2023). Thus, to determine if the origins of the ATV TnpB proteins, a BLAST search of the proteins was carried out to determine the cellular homologs. As shown in Tables 3.1-3.4, all the ATV TnpB proteins have counterparts in the closely related *Acidianus* two-tailed virus 2 (> 87% identity) and several members of the *Sulfolobales* (>75% identity). BLAST search results for all ATV TnpB proteins can be found in Appendix B.

ATV (*Bicaudavirus pozzuoliense* species) is a member of the family *Bicaudaviridae*, which consists of viruses that infect hyperthemnophilic archaea of the order *Sulfolobales*. As previously mentioned, the virus has both lytic and lysogenic life cycles. During the lysogenic life cycle, the viral genome is integrated into the host chromosome. As the viral genome is extracted from the host chromosome, it is not uncommon for some genetic material of the host to be extracted together with the viral genome. Thus, this process might have allowed the virus to highjack some of the genetic material of the host. If this were the case, the genetic material taken up by the virus might be integrated into the genome of the virus and possibly, endow the virus with an evolutionary advantage.

Table 3.1. NCBI BLAST search results for ATV TnpB gp10 protein

Description	Scientific Name	Percent identity
transposase (ATV TnpB GP10)	*Bicaudavirus pozzuoliense*	100.00
putative transposase	*Acidianus two-tailed virus 2*	93.40
transposase [Sulfolobales archaeon]	*Sulfolobales archaeon*	81.46
RNA-guided endonuclease TnpB family protein	*Sulfolobus islandicus*	78.89
RNA-guided endonuclease TnpB family protein	*Acidianus brierleyi*	80.65

RNA-guided endonuclease TnpB family protein	*Acidianus brierleyi*	80.90
transposase	*Sulfolobus islandicus*	78.64
RNA-guided endonuclease TnpB family protein	*Sulfolobus islandicus*	78.14
transposase	*Sulfolobus islandicus*	77.89
IS200/IS605 family OrfB protein	*Sulfolobus islandicus HVE10/4*	79.65
transposase	*Sulfolobus islandicus*	78.14
RNA-guided endonuclease TnpB family protein	*Sulfolobus islandicus*	79.65
RNA-guided endonuclease TnpB family protein	*Sulfurisphaera ohwakuensis*	77.56
RNA-guided endonuclease TnpB family protein	*Sulfolobus islandicus*	78.14
RNA-guided endonuclease TnpB family protein	*Sulfolobus islandicus*	78.14
RNA-guided endonuclease TnpB family protein	*Sulfolobus islandicus*	78.39
transposase	*Sulfolobus islandicus*	77.89
RNA-guided endonuclease TnpB family protein	*Acidianus brierleyi*	80.65
RNA-guided endonuclease TnpB family protein	*Sulfolobus islandicus*	77.89
RNA-guided endonuclease TnpB family protein	*Acidianus brierleyi*	78.64

Table 3.2. NCBI BLAST search results for ATV TnpB gp40 protein

Description	Scientific Name	Percent identity
transposase (ATV TnpB GP40)	*Bicaudavirus pozzuoliense*	100.00
IS605 family OrfB family transposase	*Acidianus two-tailed virus 2*	89.28
RNA-guided endonuclease TnpB family protein	*Saccharolobus solfataricus*	85.78
RNA-guided endonuclease TnpB family protein	*Saccharolobus solfataricus*	84.90
RNA-guided endonuclease TnpB family protein	*Acidianus infernus*	85.12
transposase	*Saccharolobus solfataricus*	85.78
RNA-guided endonuclease TnpB family protein	*Saccharolobus solfataricus*	83.59
RNA-guided endonuclease TnpB family protein	*Saccharolobus solfataricus*	86.00
transposase	*Acidianus infernus*	82.39
transposase	*Acidianus infernus*	84.53
transposase	*Acidianus infernus*	85.87
RNA-guided endonuclease TnpB family protein	*Acidianus infernus*	84.86
transposase	*Caldivirga sp.*	76.47
transposase	*Candidatus Marsarchaeota G2 archaeon ECH_B_SAG-C16*	75.60

transposase	*Thermocladium sp. ECH_B*	72.49
transposase	*Thermocladium sp. ECH_B*	71.93
RNA-guided endonuclease TnpB family protein	*Saccharolobus shibatae*	70.59
Transposase	*Acidilobus sp.*	74.73
transposase, IS605 OrfB family, central region	*uncultured Acidilobus sp. CIS*	74.73
transposase	*Thermocladium sp. ECH_B*	72.21

Table 3.3. NCBI BLAST search results for ATV TnpB gp43 protein

Description	Scientific Name	Percent identity
transposase (ATV TnpB GP43)	*Bicaudavirus pozzuoliense*	100.00
IS605 family OrfB transposase	*Acidianus two-tailed virus 2*	87.73
RNA-guided endonuclease TnpB family protein [Aci	*Acidianus infernus*	85.90
RNA-guided endonuclease TnpB family protein	*Acidianus infernus*	85.12
transposase	*Bicaudavirus pozzuoliense*	86.16
RNA-guided endonuclease TnpB family protein	*Saccharolobus solfataricus*	83.29
RNA-guided endonuclease TnpB family protein	*Acidianus infernus*	83.55
RNA-guided endonuclease TnpB family protein	*Sulfolobus islandicus*	79.11
RNA-guided endonuclease TnpB family protein	*Acidianus ambivalens*	77.55
RNA-guided endonuclease TnpB family protein	*Acidianus ambivalens*	77.02
RNA-guided endonuclease TnpB family protein	*Acidianus hospitalis*	78.59
RNA-guided endonuclease TnpB family protein	*Sulfolobales archaeon*	77.34
RNA-guided endonuclease TnpB family protein	*Sulfuracidifex metallicus*	77.81
RNA-guided endonuclease TnpB family protein	*Acidianus ambivalens*	76.76
RNA-guided endonuclease TnpB family protein	*Sulfolobales archaeon*	78.39
RNA-guided endonuclease TnpB family protein	*Sulfuracidifex metallicus*	78.07
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	*Acidianus hospitalis*	76.50
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	*Acidianus hospitalis*	76.56
RNA-guided endonuclease TnpB family protein [Saccharolobus solfataricus]	*Saccharolobus solfataricus*	78.07
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	*Acidianus hospitalis*	77.34

Table 3.4. NCBI BLAST search results for ATV TnpB gp68 protein

Description	Scientific Name	Percent identity
transposase (ATV TnpB Gp68)	*Bicaudavirus pozzuoliense*	100.00
IS605 family OrfB transposase	*Acidianus two-tailed virus 2*	87.21
RNA-guided endonuclease TnpB family protein	*Saccharolobus solfataricus*	87.21
RNA-guided endonuclease TnpB family protein	*Acidianus infernus*	85.64
transposase	*Bicaudavirus pozzuoliense*	86.16
RNA-guided endonuclease TnpB family protein	*Acidianus infernus*	85.64
RNA-guided endonuclease TnpB family protein	*Acidianus infernus*	87.47
RNA-guided endonuclease TnpB family protein	*Acidianus ambivalens*	79.63
RNA-guided endonuclease TnpB family protein	*Acidianus ambivalens*	79.90
RNA-guided endonuclease TnpB family protein	*Acidianus ambivalens*	79.43
RNA-guided endonuclease TnpB family protein	*Acidianus ambivalens*	78.33
RNA-guided endonuclease TnpB family protein	*Sulfuracidifex metallicus*	78.85
RNA-guided endonuclease TnpB family protein	*Acidianus hospitalis*	78.91
RNA-guided endonuclease TnpB family protein	*Acidianus hospitalis*	80.42
RNA-guided endonuclease TnpB family protein	*Sulfuracidifex metallicus*	79.11
RNA-guided endonuclease TnpB family protein	*Sulfolobus islandicus*	78.59
RNA-guided endonuclease TnpB family protein	*Acidianus hospitalis*	77.55
RNA-guided endonuclease TnpB family protein	*Acidianus ambivalens*	78.59
RNA-guided endonuclease TnpB family protein	*Sulfuracidifex metallicus*	77.86
RNA-guided endonuclease TnpB family protein	*Acidianus hospitalis*	78.07

Next, an attempt was made to determine if the ATV TnpB proteins originated from solo tnpB genes (i.e. from IS1341 transposons) or from IS200 or IS605 transposons (which consist of both *tnpA* and *tnpB* genes) in host cells. In both the IS200 and IS605 transposons, the *tnpA* gene is always located upstream of the *tnpB* gene. The genes immediately upstream of all four ATV *tnpB* genes – gp09, gp39, gp42 and gp67 – are annotated as hypothetical proteins. Consequently, a BLAST search of the protein products of these genes was carried out to determine whether they are TnpA proteins. Results revealed that whereas cellular homologs were identified for gp39 and gp67,

homologous proteins for gp09 and gp67 were hypothetical proteins in the closely related ATV variant 1 and ATV2 (Table 3.5-3.8 and appendix B).

Table 3.5. NCBI BLAST search results for ATV gp09 protein

Description	Scientific Name	Percent identity
protease (ATV gp09)	*Bicaudavirus pozzuoliense*	100.00
putative metal-dependent protease	*Acidianus two-tailed virus 2*	89.95
hypothetical protein QIT30_gp34	*Saccharolobus solfataricus rod-shaped virus 1*	40.79
protease	*Acidianus rod-shaped virus 2*	40.43
hypothetical protein	*Sulfuracidifex metallicus*	37.07

Table 3.6 NCBI BLAST search results for ATV gp39 protein.

Description	Scientific Name	Percent identity
hypothetical protein (ATV_gp39)	*Bicaudavirus pozzuoliense*	100.00
hypothetical protein	*Acidianus infernus*	90.84
hypothetical protein	*Acidianus two-tailed virus 2*	86.03
hypothetical protein	*Saccharolobus solfataricus*	80.74

Table 3.7. NCBI BLAST search results for ATV gp42 protein.

Description	Scientific Name	Percent identity
major head protein (ATV gp42)	*Bicaudavirus pozzuoliense*	100.00
Chain Z, RNAP inhibitory protein	*Bicaudavirus pozzuoliense*	100.00
archaeal structural protein	*Acidianus two-tailed phage variant 1*	100.00
archaeal structural protein	*Acidianus two-tailed virus 2*	90.78
hypothetical protein	*Acidianus infernus*	92.41
major head protein	*Acidianus tailed spindle virus*	43.96
major head protein	*Sulfolobus monocaudavirus SMV2*	38.82
major head protein	*Sulfolobus monocaudavirus SMV1*	36.47
major head protein	*Sulfolobus monocaudavirus SMV4*	35.29
major head protein	*Sulfolobus monocaudavirus SMV3*	46.67

Table 3.8. NCBI BLAST search results for ATV gp67 protein.

Description	Scientific Name	Percent identity
hypothetical protein ATV_gp67	*Bicaudavirus pozzuoliense*	100.00
membrane protein	*Acidianus two-tailed virus 2*	88.29
hypothetical protein	*Acidianus infernus*	95.26

3.1.3. Prediction ωRNA sequences

The gene encoding ωRNA comprises the 3'end of the *tnpB* gene and several nucleotides in the intergenic region between the *tnpB* gene and the adjacent gene and is typically 150 nucleotides in length. The guide sequence, which is a specific sequence that is used to recognize the target DNA by complementation, is located at the end of the ωRNA sequence and is typically about 20 nucleotides in length (Karvelis et al., 2021 & Nety et al., 2023). In order to predict the ATV TnpB ωRNA sequences, a sequence alignment of all *tnpB* nucleotide sequences starting from -250 nucleotides upstream of the stop codon until the end of the intergenic region was carried out (Figure 3.5). As previously reported, the guide sequence would be the sequence following conserved regions since the guide sequence would differ from one sequence to another. However, the results were not inclusive. Moreover, with the same sequences, RNAFold was performed to predict the hairpin structures and guide RNA sequences (Figure 3.6).

It is worth noting that there is no confirmation that suggest that every TnpB must have an ωRNA and perform as an RNA-guided endonuclease. Exaptation can occur to TnpB protein. Exaptation is when something takes on a function that was not designed for it naturally and losses its original function (Altae-Tran et al., 2023). The study indicates that a function TnpB can take on could be a regulatory function that require binding bu not cleavage. Therefore, ATV TnpBs, which have not been thoroughly studied yet, might exhibit this feature or not.

```
GP10/40/43/68
CLUSTAL O(1.2.4) multiple sequence alignment

GP43      --------------------------------ttcagtactggatagaatggcaagccaga      29
GP68      --------------------------------ttcagtattggatagagtggcaggctaag      29
GP10      -------ccccctatagatttaaaaacttctattcacaatattctatttaatgcccgac      53
GP40      gaacacagctgtttttaacgaaactaacttcttcaaacttttataaaactgagggttaaa      60
                                         *          *  *

GP43      aaacacggaatgaatgtggagtatgttaatcctagttactcttccgtctcttgtccaaag      89
GP68      aaacacggaatgattgtggagtttgttaatcctagttactcttccgtttcttgcccaaag      89
GP10      gtagggattaggtttagggctttcgc-------agatgaacctacaattagggag----t      102
GP40      gtgggcggaa-----agggaatggat-------agtcccttatagatttttaaac-----t      104
                    *       **  *           **       *     *

GP43      tgtggccgcaaaatggttgagattgctcataggtacttccactgtccctcg---------      140
GP68      tgtggccacaaaatggttgagattgcttataggtactttcactgtccttca---------      140
GP10      taaaa--tcccagttgaggttagcgtgcgaaatttacaacaccctacgttgggcagat--      158
GP40      tcagc------aataacgagtagtaatgagcagttatgccaccctcctctggtcaactct      158
          *          * *       *           *     ***  * *

GP43      --tgtggttatgagaacgatcgtgacgttattgctatcatgaatttaaatgggaggggt      198
GP68      --tgtggttatgagaacgatcgtgatgttattgctatcatgaatttaaatgggaggggt      198
GP10      -------------------------atatattt-----c---------------------      167
GP40      taggggatgaggagcgggagccgacttctactc-----ccgcaataccggaggagggtgt      213
                                       ** *

GP43      ctctgaccctctcgactgcccctcagatgagagatgtaat---------cccgaatcgat      249
GP68      ctctgaccctctcgactgcccctcaaatgagagatgtagc---------tccgaatcgat      249
GP10      -taccaaagagatgggaaa-ggtctaagcaagactgaattgagacaactcgc-tctagat      224
GP40      ctacaaagttaaatactca-aatcgaagaaccaatatagttcgccttcttccaaatggtt      272
           *   *                ** *  *    *  *            *   * * *

GP43      gaggggaactatgaacccctgggagagaaccctca--------cccttccaaggcgggg      301
GP68      gaggggaaccctcgctgacggggaggaagtcaggcc--------ttcctaagggatggat      301
GP10      ttgagaaaacaggacgatgaata-caagagaatctattctcaagcagtacaacaaattgc      283
GP40      ttcaagaaagaaaactgaggaggttagcagacctctctgcaaagctcttcaacgaagtta      332
                **

GP43      aggaagtcagtggccc-------------------------------------      317
GP68      agccattatgtttttaaactttctatcaagattccctgggt-----------      342
GP10      agacaga-------------ttcta----------------------------      295
GP40      actatgaaaggaggcaacagttctttcacgaagggaaagtggacattaaggg      384
          *
```

Figure 3.5. Sequence alignment for all the ATV *tnpB* nucleotide sequences starting from -250 nucleotides upstream the stop codon and until the end of the intergenic region

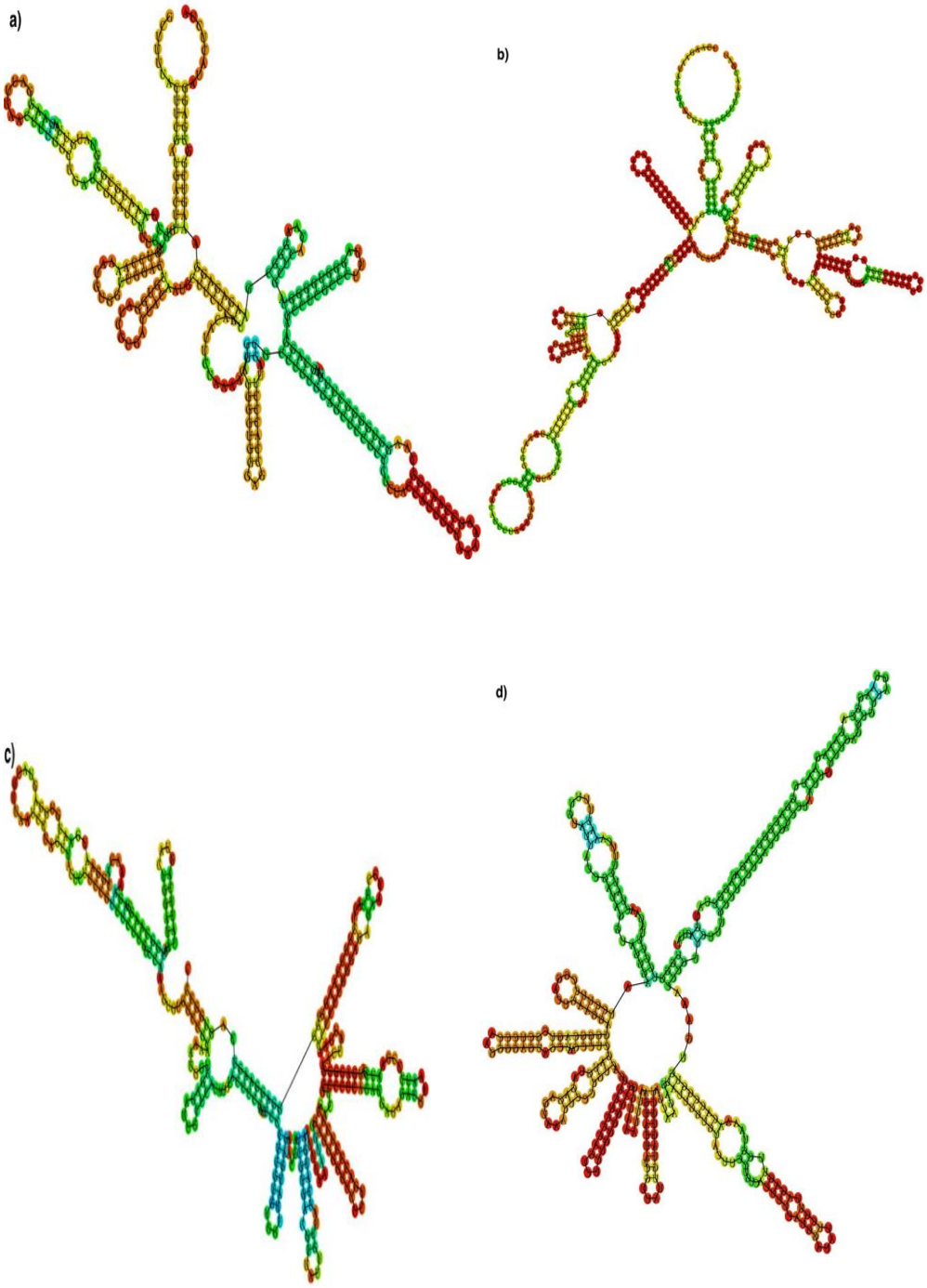


Figure 3.6. RNAFold for sequences starting from -250 nucleotides upstream of the stop codon until the end of the intergenic region of all ATV *tnpB* genes. a) sequences form gp10. b) sequences from gp40. c) sequences from gp43. d) sequences from gp68

3.1.4. TAM sequence prediction

In previous studies, the TAM sequence has been reported to match the sequence of the TnpA target site which occurs either upstream of the *tnpA* gene (for IS200 or IS605 elements) or upstream of the *tnpB* gene, if the transposon occurs without the *tnpA* gene. Mobilization of solo *tnpB* genes is predicted to require TnpA supplied in trans (Barabas et al., 2008). Prediction of TAM sequences involve sequence alignment of the 5' intergenic region of all 1S200/605 elements in any particular organism, and the first four to five uninterrupted conserved bases of the aligned sequences are considered the TAM sequences (Xu et al., 2023). When the methodology described above was used for prediction of TAM sequences in ATV, the results revealed that the sequence of the first five conserved bases is TTATA (Figure 3.7).

This sequence bears features of most well-characterized TAM sequences i.e. AT-rich (Xu et al., 2023). More pertinent to this study, this predicted TAM sequence is similar to the TAM sequences i.e. TTTAA and TTTAT for IS605 and IS1341 transposons, respectively identified in the hyperthermophilic archaean and ATV host *S. islandicus* REY15A (Xu et al., 2023). Thus, in addition to the predicted ATV TnpB TAM sequence, it was decided to include the *S. islandicus* TAM sequences in our experiments for two main reasons. First, the *S. islandicus* TAM sequences are similar enough to the predicted ATV TnpB TAM sequences and could be considered randomized sequences. Such randomized sequences are used to assess the specificity of the TAM sequence. Second, it was considered interesting to determine if the ATV TnpB proteins recognize the *S. islandicus* TAM sequence, which might suggest a role for ATV TnpB proteins in host interference. For the purpose of this study, each TAM oligonucleotide substrate sequence

was designed such that the TAM sequence is upstream an oligonucleotide sequence complementary to the guide RNA sequence.

Moreover, the target DNA was designed in a away to complement the predicted guide sequence (around 16 nucleotides) and other nucleotides derived from the 5' Intergenic end of the *tnpB* gene were added after it.

Figure 3.7. DNA sequence alignment of the 5' intergenic region of all ATV *tnpB* genes in order to predict the TAM sequence. Conserved sequences are in red.

Finally, after predicting the possible TAM and guide RNA sequences, oligonucleotides containing the 5'-TAM-guide sequence-extra nucleotides-3' were designed for the cleavage assays. Both target and non-target (reverse complement) strands were designed. Also, the same sequences were cloned into pUC19 plasmid to test the plasmid cleavage activity of the proteins, which will be referred to as pTAM. (All sequences can be found in Appendix C)

3.2. ATV TnpB Proteins' Purification

3.2.1 Plasmid verification

The recombinant pET30a(+) vector was subjected to restriction digest analysis to verify that the insert was cloned into the plasmid. The restriction enzymes used for pET30a(+) were *NdeI* and *XhoI*. The size of pET30a(+) plasmid is 5422bp and the inserts which include 10XHis-tag and the 3' Intergenic regions vary in size and are as follows: gp10 is 1304bp, gp40 is 1546, gp43 is 1251bp and gp68 is 1276bp. As seen in Figure 3.8, the insert band (lower red arrow) can be visualized, confirming the presence of the inserts.

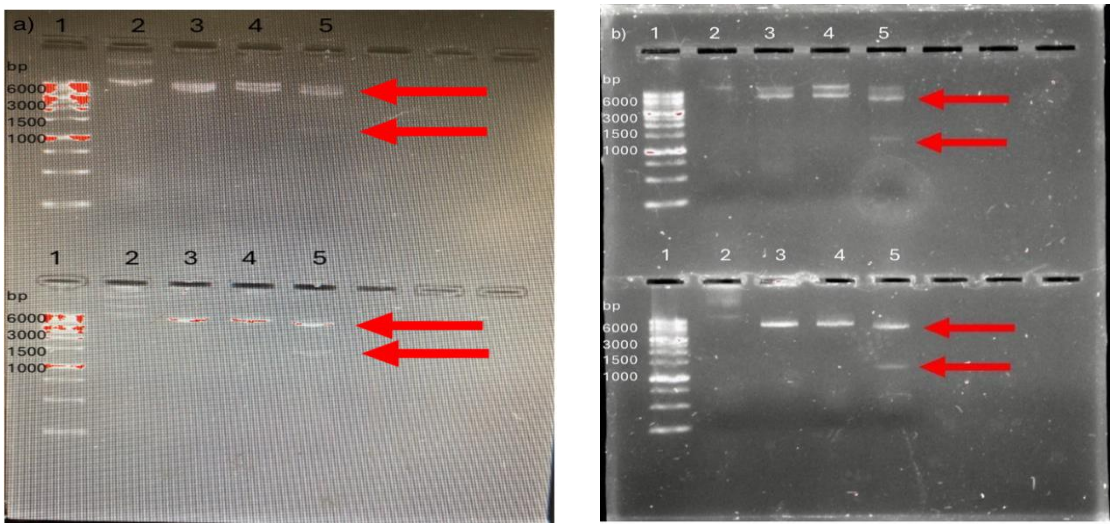


Figure 3.8. shows the restriction enzyme cleavage of pET30a(+) plasmids loaded into 1.5% agarose gel. The top red arrow indicate the plasmid and the bottom red arrow indicates the insert. a) Top gel: gp10-pET30a(+) restriction digest. The lanes indicate the following: 1. Marker, 2. Plasmid with no digest, 3. Single digest with NdeI, 4. Single digest with XhoI, 5. Double digest. Bottom gel: gp40-pET30a(+) restriction digest. The lanes indicate the following: 1. Marker, 2. Plasmid with no digest, 3. Single digest with NdeI, 4. Single digest with XhoI, 5. Double digest. b) Top gel: gp43-pET30a(+) restriction digest. The lanes indicate the following: 1. Marker, 2. Plasmid with no digest, 3. Single digest with NdeI, 4. Single digest with XhoI, 5. Double digest. Bottom gel: gp68-pET30a(+) restriction digest. The lanes indicate the following: 1. Marker, 2. Plasmid with no digest, 3. Single digest with NdeI, 4. Single digest with XhoI, 5. Double digest.

3.2.2. Expression and purification of MBP-TnpB construct

Genes encoding the four ATV TnpB proteins were ordered from GenScript with an N-terminal gene cassette consisting of MBP-His10-TEV protease recognition sequence (MBP, maltose binding protein; His10, a stretch of 10 histidine residues, TEV, tobacco etch virus) cloned into the *BamHI* and *NdeI* restriction enzyme sites of the pET42b (+) vector. All the genes were codon-optimized for expression in the *Escherichia coli* expression host. The decision to employ the MBP solubility tag was based on the observation that all other functionally characterized TnpB proteins were only successfully expressed with solubility tags such as MBP or the glutathione S transferase (GST), apparently due to the low solubility of TnpB proteins (Karvelis et al., 2021). Small-scale expression trials in 5-ml LB or TB were carried out to determine the best expression conditions prior to large scale expression. Subsequently, small scale Ni^{2+}-NTA affinity chromatography was carried out to assess the level of protein expression.

The calculated molecular weights of the TnpB proteins are as follows: gp10 (48.226 kDa), gp40 (53.287 kDa), gp43 (44.954 kDa) and gp68 (44.814 kDa). However, when expressed with the N-terminal tag as described above, the calculated molecular weights increased to 97.238 kDa, 102.299 kDa, 93.967 kDa and 93.826 kDa respectively. SDS-PAGE analysis of the small scale Ni^{2+}-NTA purification revealed low expression levels for all the proteins (Figure 3.9).

Since the MBP-His10-TEV protease recognition construct-gp40 construct displayed the highest expression level, it was decided to commence large scale expression (in 1 liter TB) and purification with the protein. SDS-PAGE analysis of Ni^{2+}-NTA affinity chromatography showed that the expression level was high (Figure 3.10). However, efforts to cleave the His tag with a commercially obtained TEV protease were not successful despite several attempts. Consequently, another strategy was designed to express and purify the ATV TnpB proteins.

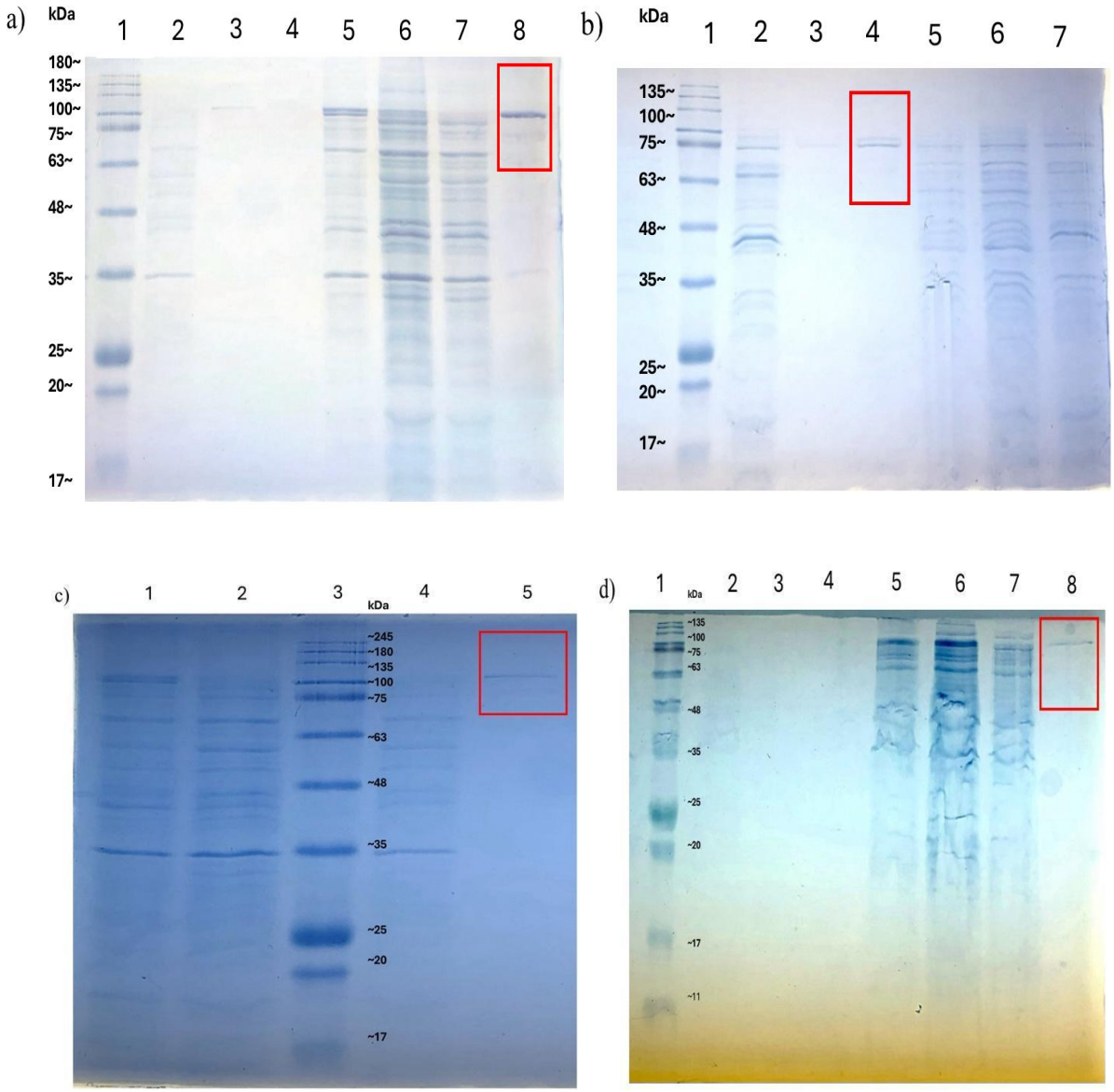
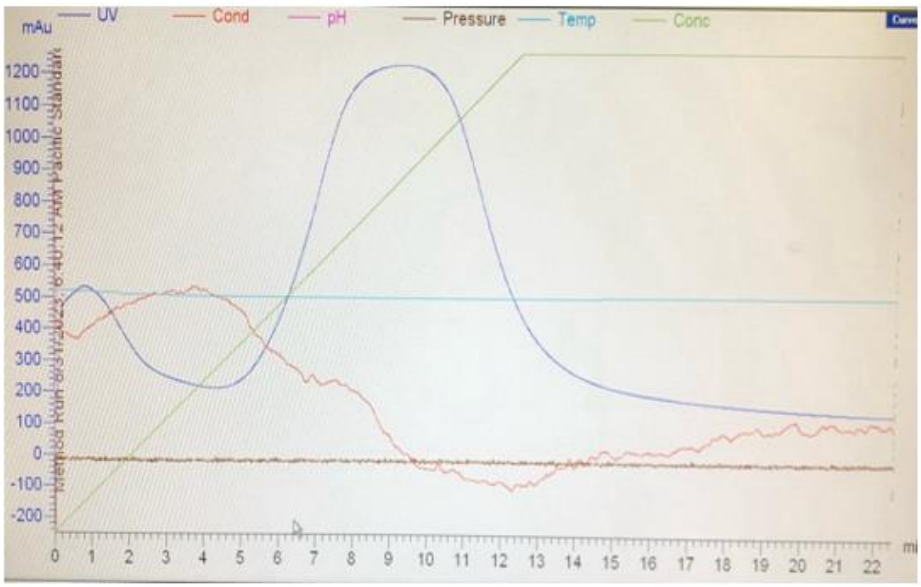


Figure 3.9. 10% SDS-PAGE gel results for small scale Ni^{2+}-NTA purifications. Red boxes indicate elution fraction. a) gp10 small scale Ni^{2+}-NTA purification. b) gp40 small scale Ni^{2+}-NTA purification. c) gp43 small scale Ni^{2+}-NTA purification. d) gp68 small scale Ni^{2+}-NTA purification.

a)



b)

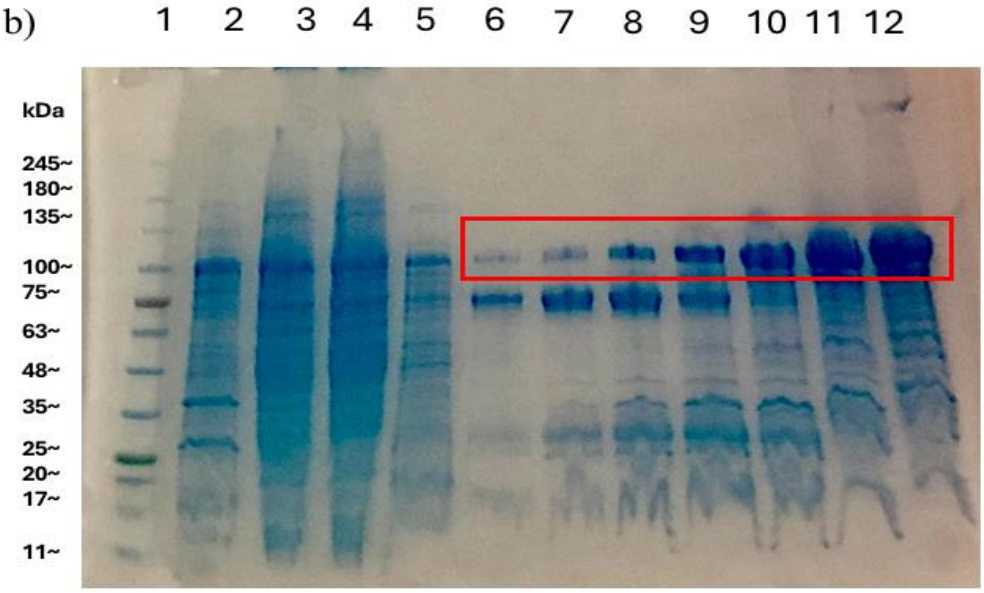


Figure 3.10. Large scale Ni^{2+}-NTA purification for gp40 (MBP construct) using FPLC. a) FPLC chromatogram. b) 10% SDS-PAGE gel result (the red box indicates the elution fractions). The lanes represent the following: 1. Marker, 2. Lysate, 3. Pellet, 4. Soluble fraction, 5. Flow through, 6-12. Elution Fractions

3.2.3. Large scale expression trials of TnpB-RNA complex

Given the difficulties with TEV protease cleavage described above, a different expression construct was designed to eliminate the need for a TEV protease. Additionally, since previous studies have successfully expressed the TnpB-ωRNA complex by including codons of the 3' intergenic region, expression constructs consisting of a His10-tag at the N-terminus and the entire *tnpB* gene plus the 3' intergenic region was designed (see Appendix C for details). This expression construct was ordered from Genscript cloned into the pET30a(+) vector and transformed into BL21-A1 competent cells. The reason pET30a(+) vector was used this time was because it appears to be the most common vector for expression of proteins from archaea and their viruses, including the *S. islandicus* TnpB protein (Xu et al., 2023 & (Feng et al., 2024). Importantly, the vector has been used to successfully express other ATV proteins in our laboratory. The decision to use BL21-A1 was because this strain of *E. coli* possesses tight regulation and is therefore suitable for expression of toxin proteins (Lee, Francklyn and Hamilton, 1987), including TnpB proteins (Karvelis et al., 2021).

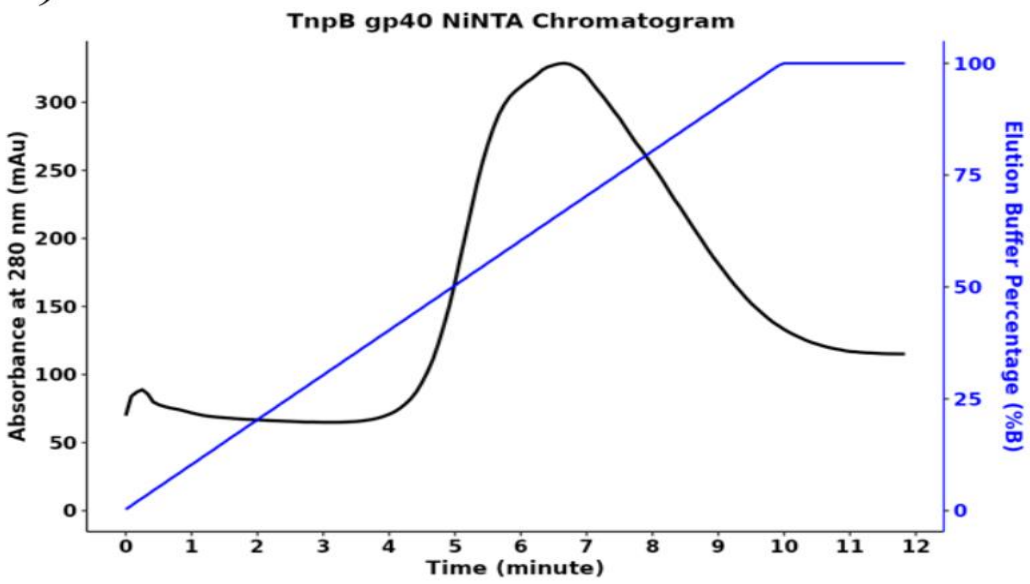
With this construct described above, the calculated molecular weights of the ATV TnpB proteins were as follows: His10-gp10 (49.596 kDa), His10-gp40 (54.658 kDa), and His10-gp43 (46.325 ida) and His10-gp68 (46.185 kDa). Following expression in 1 litre TB and subsequent purification using Ni^{2+}-NTA affinity chromatography and gel filtration chromatography, the results were analyzed by SDS-PAGE. Purification results for His10-gp40 (Figure 3.11) showed a band migrating between the 63 kDa and 75 kDa marker, which is higher than the theoretical molecular weight of 54.658 kDa. Similarly, the same observation was seen in gp68 purification (Figure 3.12) which has a theoretical molecular weight of 46.185 kDa. After concentration, the proteins were run on SDS-PAGE gels as seen in Figure 3.13.

The anomalous migratory patterns observed for both ATV proteins might occur due to several reasons. A possible explanation for the higher molecular weights observed here may be as a result of translational readthrough. This typically occurs when an aminoacyl-tRNA incorrectly responds to a stop codon, resulting in continuation of protein synthesis until another stop codon is encountered in the reading frame. Of the three naturally occurring stop codons, UAA, UAG and UGA, UAA is the most common stop

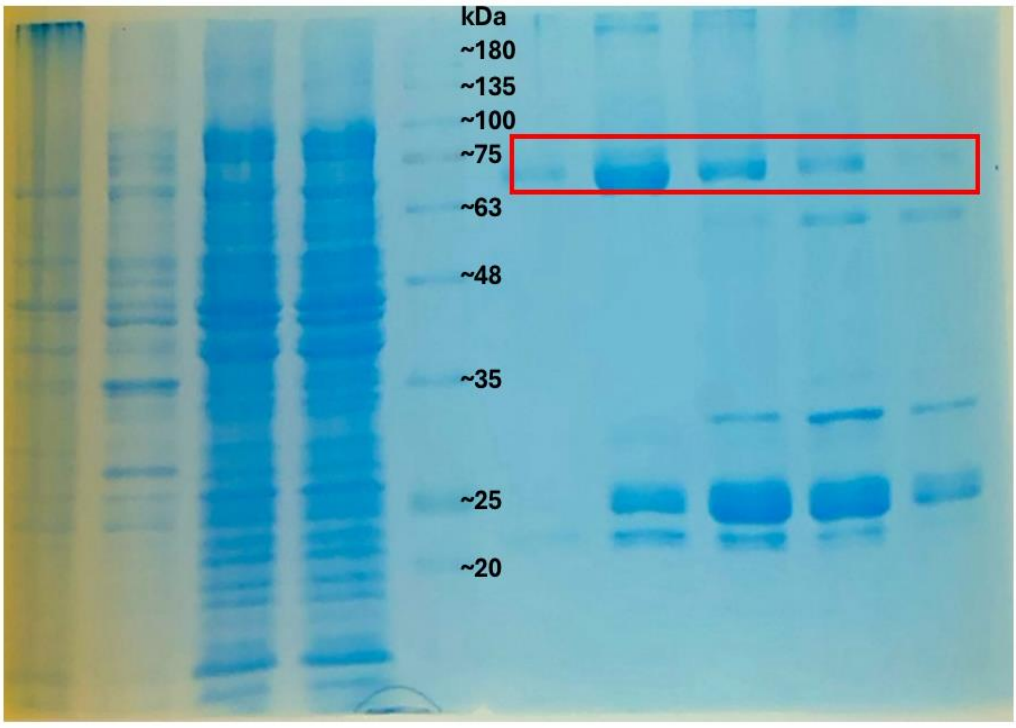
codon in *E. coli* (Sharp and Bulmer, 1988). Although the gene constructs cloned into the *E. coli* expression hosts comprised of the *tnpB* genes with their 3' intergenic sequences, the UAA stop codon was placed at the end of the *tnpB* gene. Thus, it was anticipated that protein synthesis would terminate at the stop codon despite the presence of the intergenic sequence. Attempts have been made to determine the identity and molecular mass of the proteins using mass spectrometry. However, the IZTECH Mass Spectrometry Facility has been experiencing technical problems for some time now.

Besides translational readthrough, protein-specific factors have been reported to result in anomalous migration of soluble proteins during SDS-PAGE. These include post-translational modifications (Billings et al., 1979), poor interactions with SDS molecules (Shi et al., 2012) or abundance of disordered regions or specific amino acids such as proline, basic or acidic amino acid residues (Ziemer, Mason and Carlson, 1982 & Ge et al., 2005). None of the TnpB proteins characterized till date possess any post-translational modifications and therefore, it was considered quite unlikely that the ATV TnpB proteins are modified. Protparam analysis of the amino acid components of all the ATV TnpB proteins reveal that the proline residues make up 5% or less of the total amino acid count, which is not unusual, and therefore most unlikely to account for the unusual SDS-PAGE profile. Protparam analysis also revealed that the charged amino acid residues are not excessive, especially as it was determined to be similar to other TnpB proteins. Therefore, it is quite likely that the unusual migratory pattern observed here occurs through a mechanism that is not immediately apparent. It is anticipated that mass spectrometry analysis may unravel the reason(s) for this anomaly.

a)


TnpB gp40 NiNTA Chromatogram
Absorbance at 280 nm (mAu)
Elution Buffer Percentage (%B)
Time (minute)
0
50
100
150
200
250
300
0
25
50
75
100
0 1 2 3 4 5 6 7 8 9 10 11 12

b)


1 2 3 4 5 6 7 8 9 10
kDa
~180
~135
~100
~75
~63
~48
~35
~25
~20

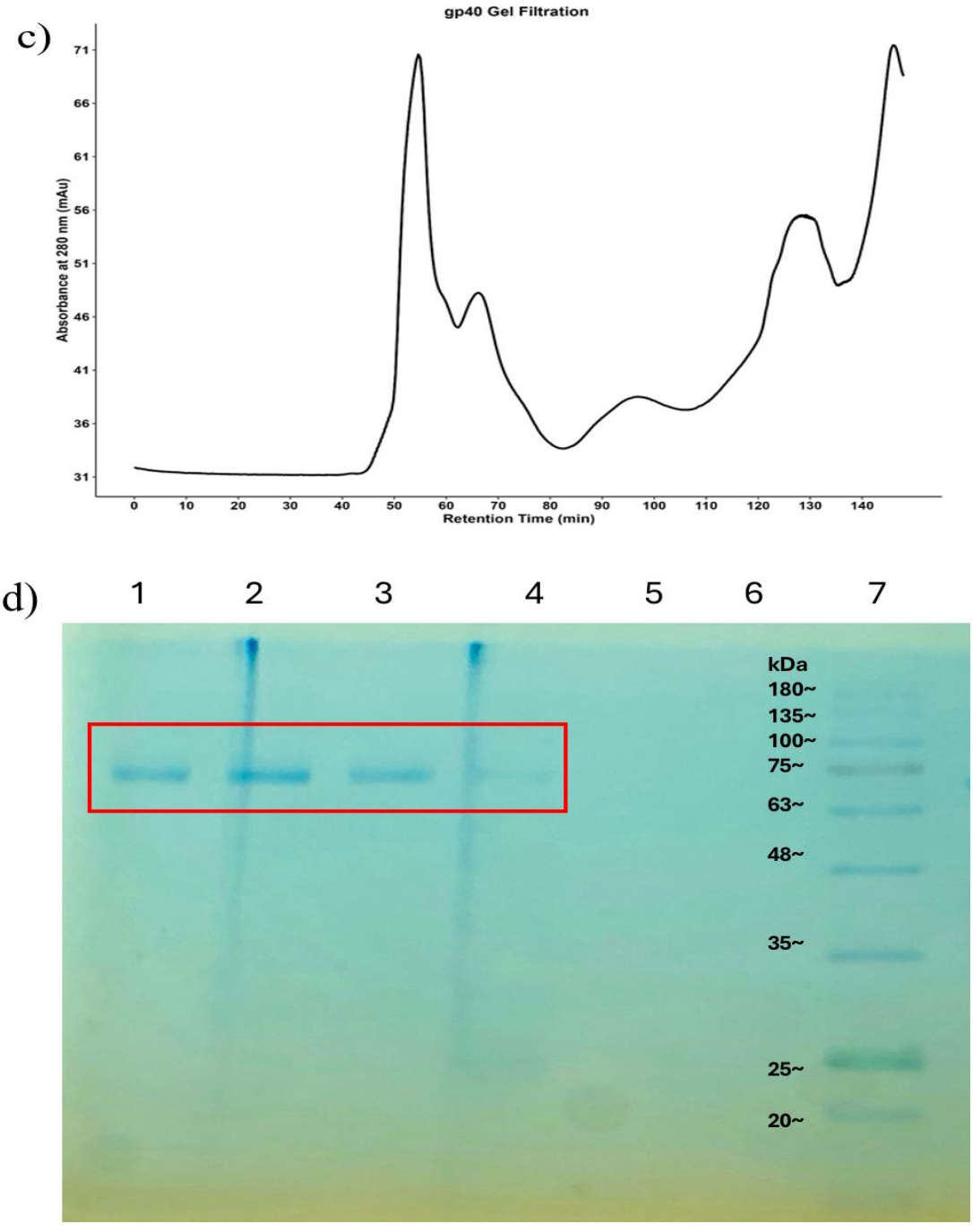
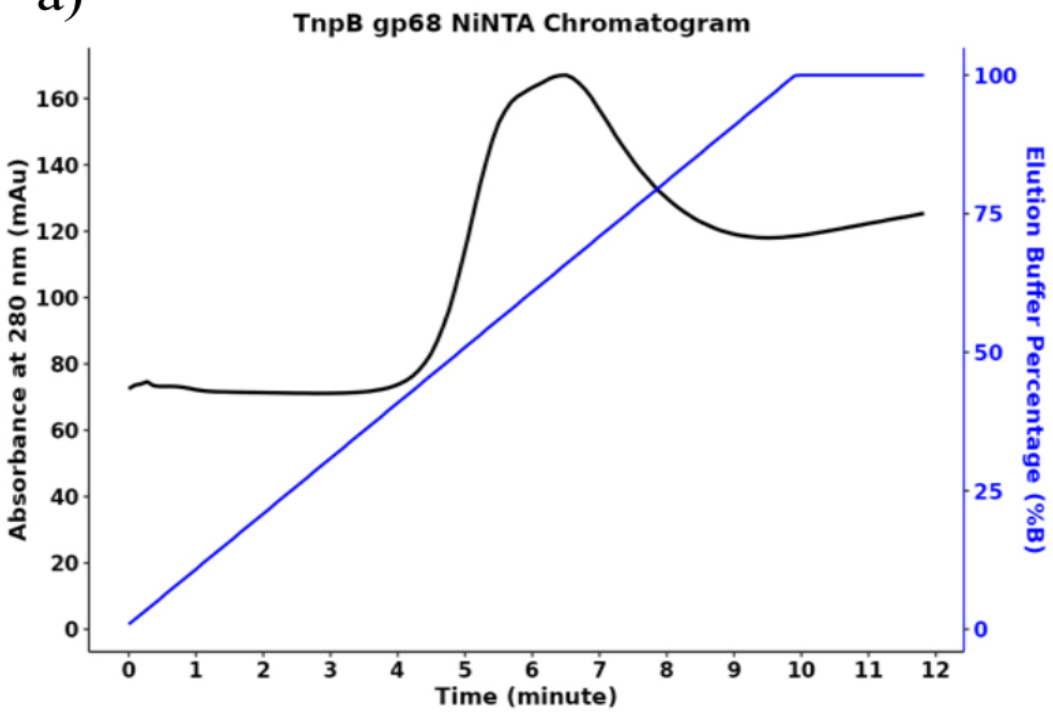
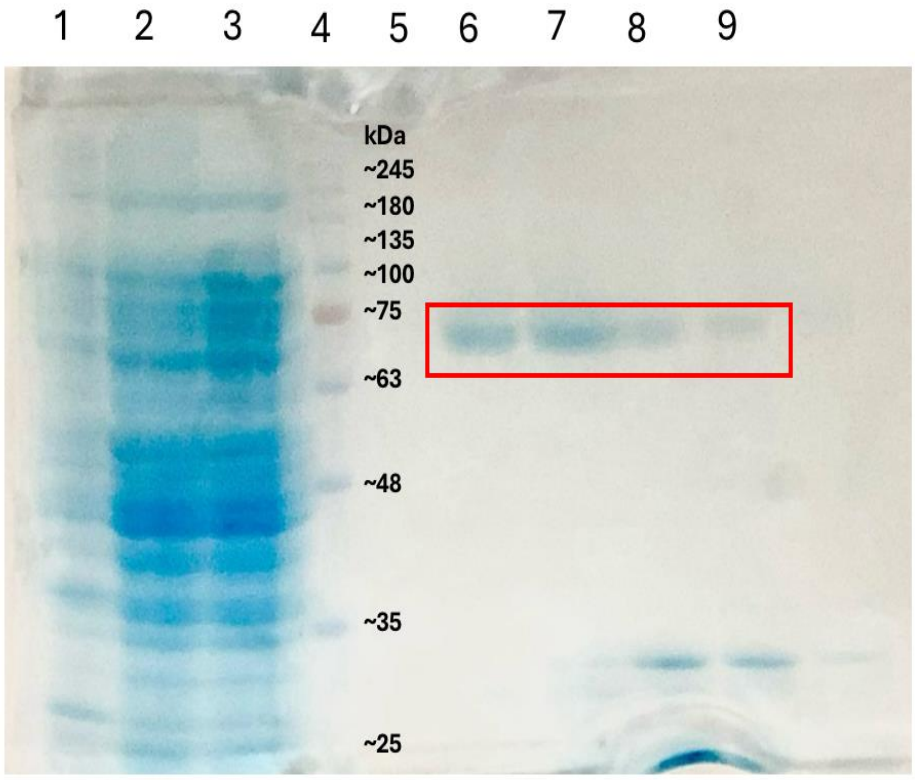


Figure 3.11. Large scale purification for gp40 (pET30a(+) construct). a) Ni^{2+}-NTA purification chromatogram. b) 10% SDS-PAGE gel results for Ni^{2+}-NTA purification (the red box indicates the elution fractions). The lanes represent the following: 1. Soluble Fraction, 2. Pellet, 3. Lysate 4. Flow through 5. Marker, 6-10. Elution Fractions. c) GF purification chromatogram. d) 10% SDS-PAGE gel results for GF purification (the red box indicates the elution fractions taken from the first peak in the chromatogram)

a)


TnpB gp68 NiNTA Chromatogram
Absorbance at 280 nm (mAu)
Elution Buffer Percentage (%B)
Time (minute)
0
20
40
60
80
100
120
140
160
0
25
50
75
100
0 1 2 3 4 5 6 7 8 9 10 11 12

b)


1 2 3 4 5 6 7 8 9
kDa
~245
~180
~135
~100
~75
~63
~48
~35
~25

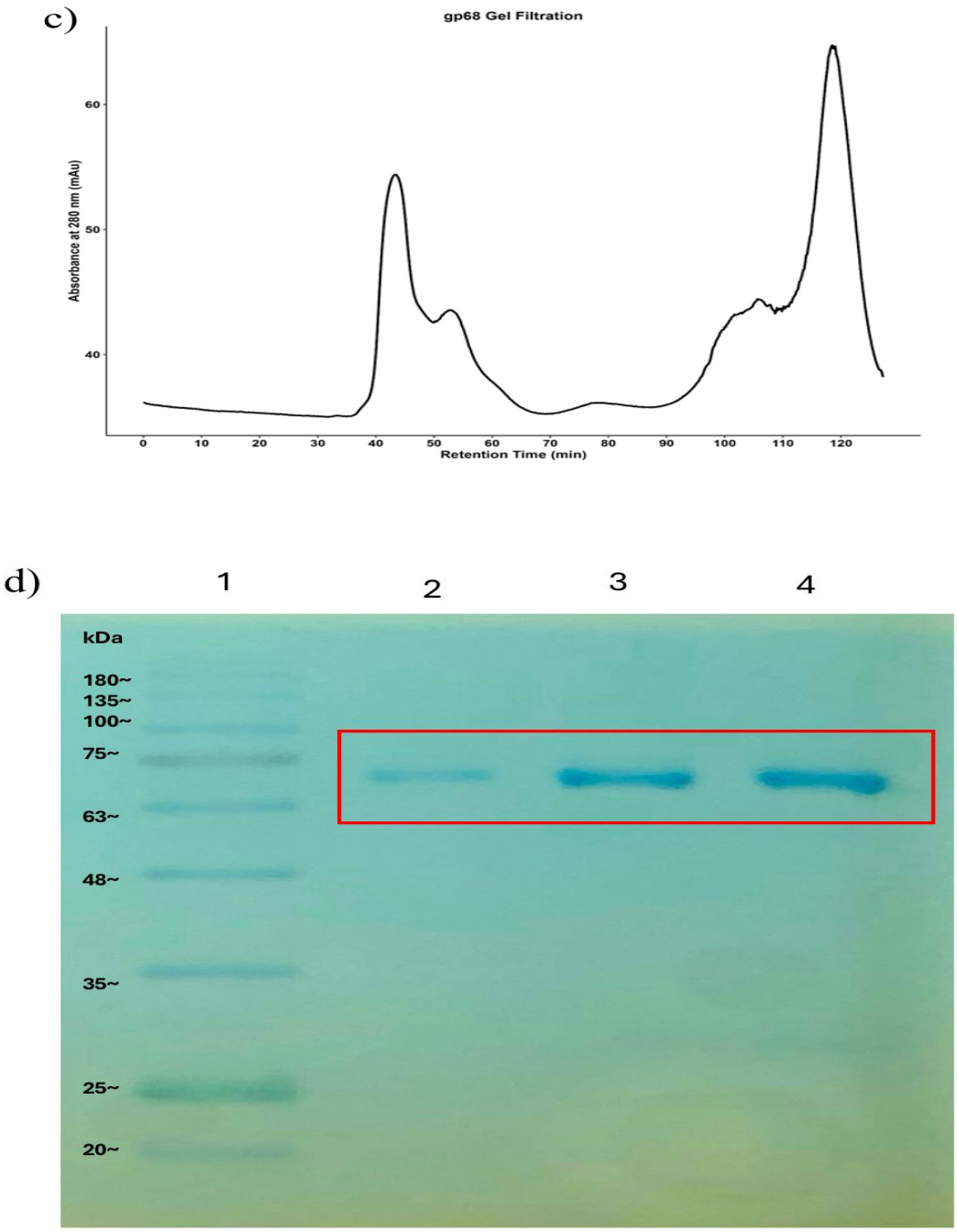


Figure 3.12. Large scale purification for gp68 (pET30a(+) construct). a) Ni^{2+}-NTA purification chromatogram. b) 10% SDS-PAGE gel results for Ni^{2+}-NTA purification (the red box indicates the elution fractions). The lanes represent the following: 1. Pellet, 2. Soluble Fraction, 3. Flow-through, 4. Marker, 6-10. Elution Fractions. c) GF purification chromatogram. d) 10% SDS-PAGE gel results for GF purification (the red box indicates the elution fractions taken from the first peak in the chromatogram)

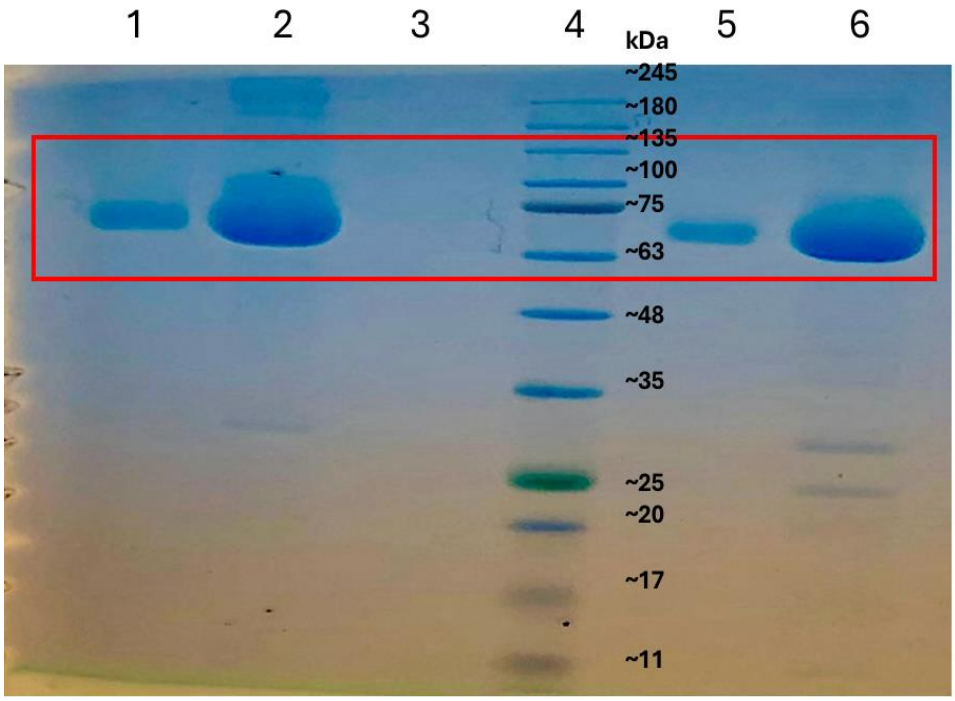


Figure 3.13. 10% SDS-PAGE gel results before and after concentrating both gp40 and gp68 purified proteins. The lanes indicate the following: 1.gp40 before concentration, 2. Gp40 after concentration, 4. Marker, 5. gp68 before concentration, 6. Gp68 after concentration

3.3. Characterization of TnpB-ωRNA complexes

In order to check for presence of nucleic acids in the purified protein samples, the 260/280 values were determined using a Nanodrop spectrophotometer. Values larger than 0.5 in protein samples typically indicate presence of nucleic acids. 0.87 and 0.88 values were present for both gp40 and gp68 respectively. Samples before and after concentration were loaded onto 1% agarose gel. Following electrophoresis and staining with ethidium bromide, nucleic acids were visualized under UV lamp. As shown in Figure 3.14, two distinct bands migrating at 500 bp and 10,000 bp, which are more visible after concentration, were present in the purified protein samples. However, after treating the samples with DNaseI and RNaseA, the DNasI treated samples did not show any band, however, very light band can be seen in the samples treated with RNaseA which would suggest that this represents the bound DNA rather than RNA. However, RNA might have

been present but any problems with handling the samples might have cause it to degrade. (Figure 3.15).

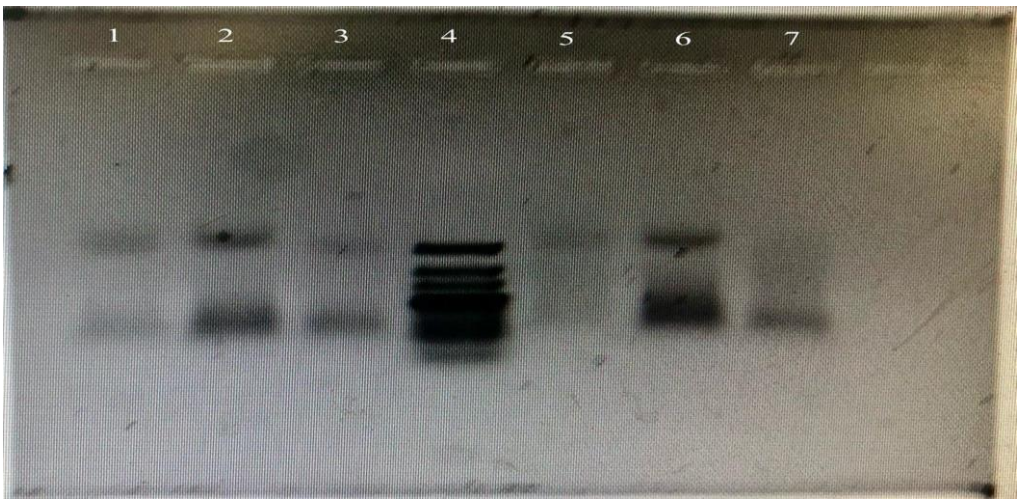


Figure 3.14. 1% agarose gel for the detection of nucleic acid contamination. The lanes indicate the following: 1. gp40 purified protein before concentration, 2. gp40 purified protein after concentration, 3. gp40 protein treated with proteinase K, 4. Marker, 5. gp68 purified protein before concentration, 6. gp68 purified protein after concentration, 7. gp68 protein treated with proteinase K

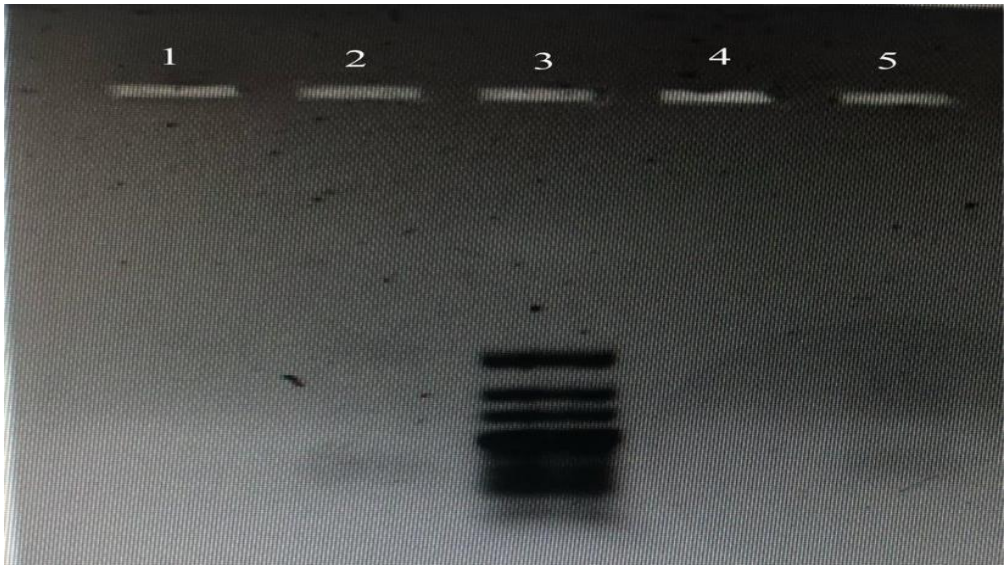


Figure 3.15. 1% agarose gel after treating the samples with DNaseI and RNaseA. The lanes indicate the following: 1. gp40 purified protein treated with DNaseI, 2. gp40 purified protein treated with RNaseA, 3. Marker, 4. gp68 purified protein treated with DNase, 5. gp68 purified protein treated with RNaseA

Although the results reported above suggest that the ATV TnpB proteins were not bound to RNA, it was decided to determine if the ωRNA molecules were expressed at all. Consequently, BL21-A1 cells were transformed with the pET30(a) recombinant plasmids and induced for expression as described previously. Subsequently, the total RNA fractions were extracted and purified using the GeneJet RNA purification kit. An aliquot of the purified RNA was analyzed using agarose gel electrophoresis as shown in Figure 3.16. The plan is to send the samples for sequencing to determine if any of the RNA molecules correspond to the anticipated ωRNA molecules.

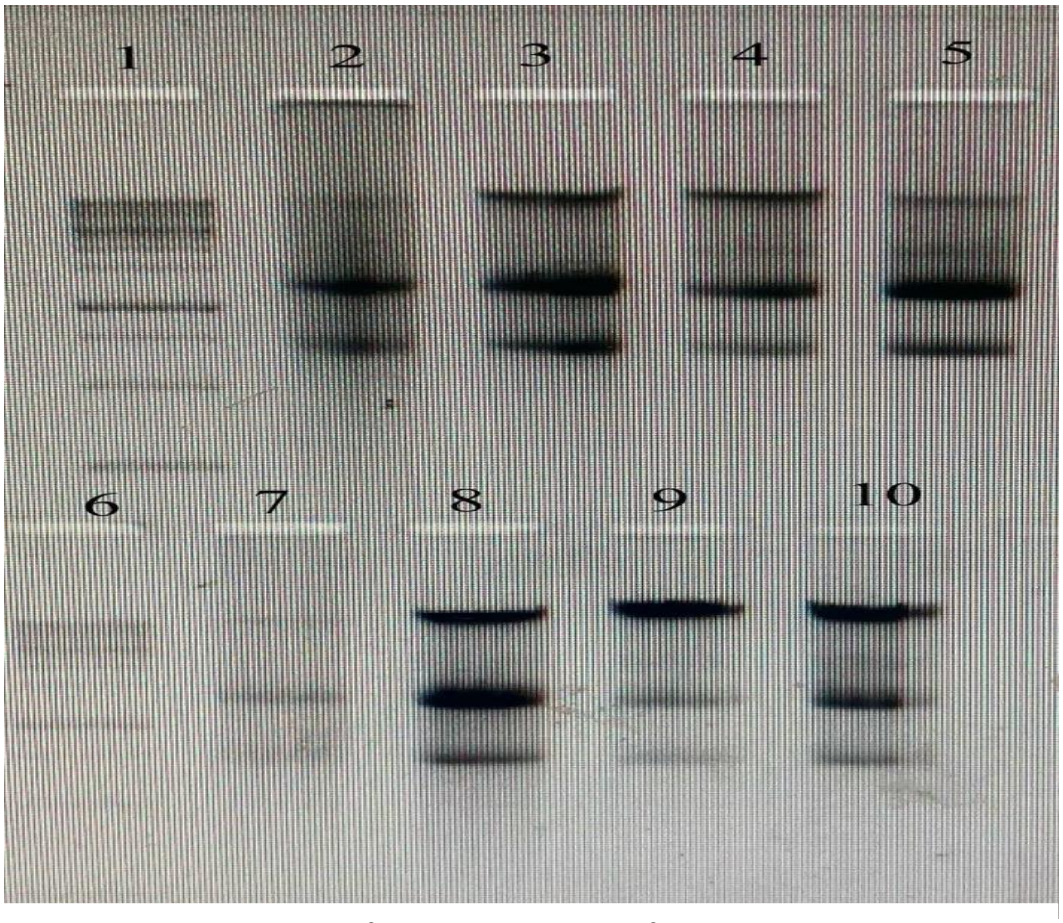


Figure 3.16. 1% agarose gel for the total RNA purification. The lanes indicate the following: 1. Marker, 2. RNA (elution 1) purified from BL21A1 transformed with gp10-pET30a(+), 3. Second elutio , 4. RNA (elution 1) purified from BL21A1 transformed with gp40-pET30a(+), 5. Second elution, 6. Marker, 7. RNA (elution 1) purified from BL21A1 transformed with gp10-pET30a(+), 8. Second elution, 9. A (elution 1) purified from BL21A1 transformed with gp43-pET30a(+), 10. Second elution

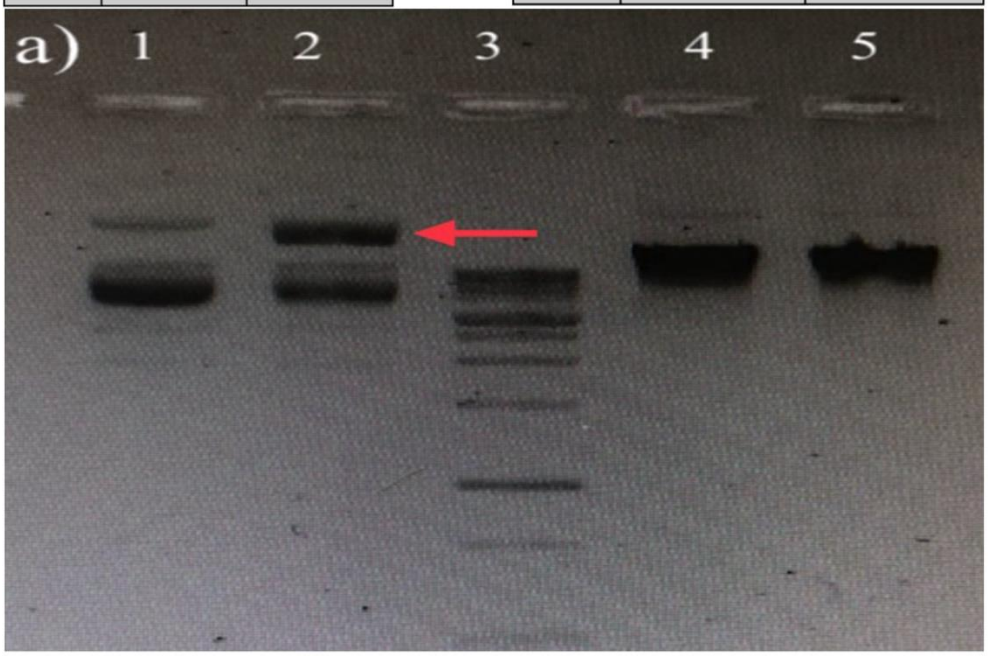
3.4. Plasmid Cleavage Assays

In vitro DNA cleavage assays were used to characterize the nuclease activity of gp68. The 2.9 kbp pTAM was employed as the substrate and the empty pUC19 plasmid (2.6 kbp) as control. Figure 3.17 shows the cleavage assays performed on pUC19 and pTAM with gp68 protein under 65°C temperature. The reason for choosing this temperature is due to the fact that ATV infects hyperthermophilic archaea and therefore, their proteins are expected to be functional at high temperatures. Supercoiled plasmid migrates faster than linear or nicked plasmid forms. Therefore, if any cleavage is taking place, this could be in a form of a nick, or it could linearize the plasmid. Therefore, we expect to see more of the higher band than the lower supercooled band. Also, bands lower than the supercoiled could indicate that double-stranded DNA cleavage has taken place in more than one position.

Results show that gp68 did not display cleavage activity when the protein was incubated with the pTAM plasmid at 65°C irrespective of whether the buffer contained metal ions (Mg^{2+} or Mn^{2+}) or EDTA (Figure 3.18). However, when the protein was incubated with the empty pUC19 plasmid, a slower migrating band was observed in the presence of both metal ions and EDTA, indicative of nickase activity. These results suggest that the nickase activity observed here may be nonspecific since the TAM sequences are not present in the empty pUC19 plasmid.

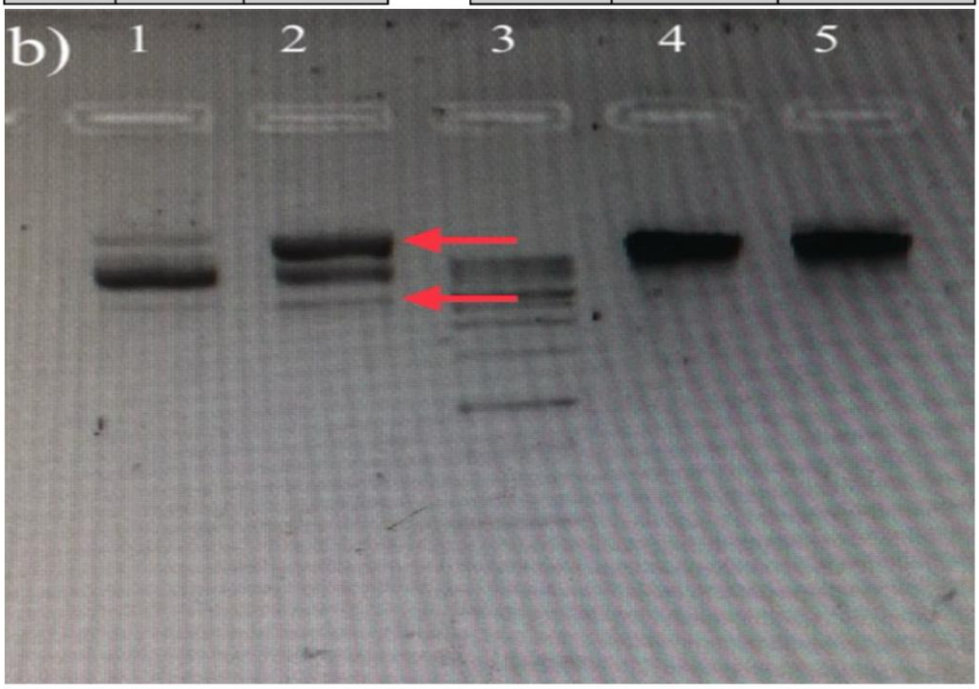
pUC19	+	+
gp68	_	+

pTAM	+	+
gp68	_	+



pUC19	+	+
gp68	_	+

pTAM	+	+
gp68	_	+



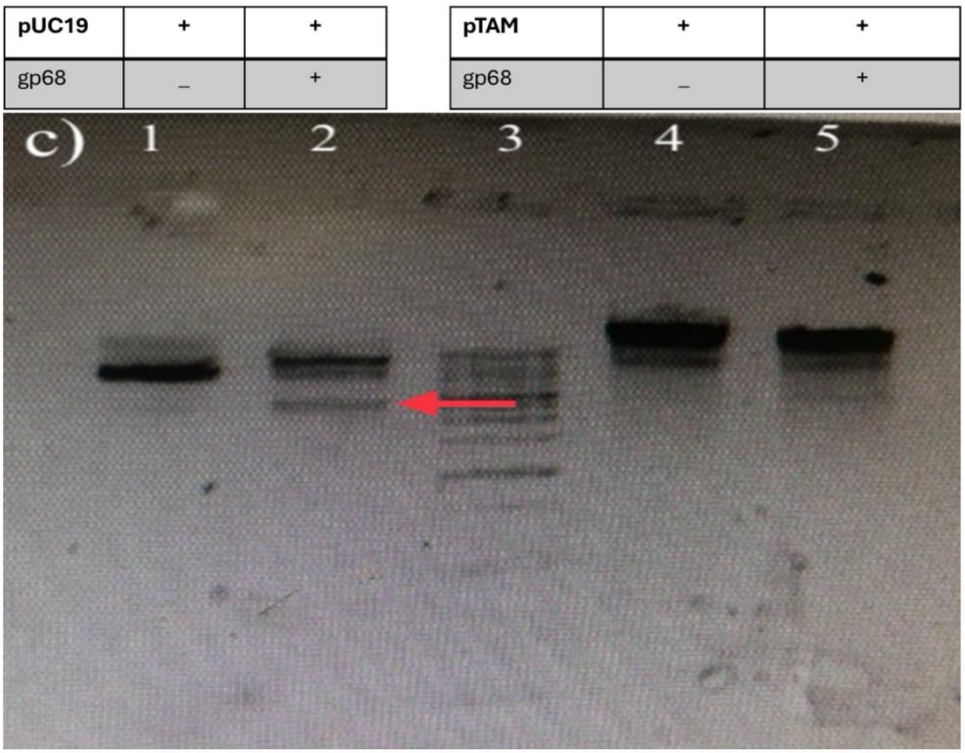


Figure 3.17. 1.5% agarose gels showing cleavage assays of pUC19 and pTAM plasmids with gp68 at 65°C. The presence of cleavage is indicated with a red arrow. a) with EDTA buffer b) with Mg^{2+} buffer c) with Mn^{2+} buffer

Similar to the results obtained at 65°C with pTAM, no nuclease activity was observed at 37°C when the buffer contained EDTA or Mg^{2+} (Figure 3.20). However, when Mn^{2+} was included in the buffer, a slower migrating band was observed.

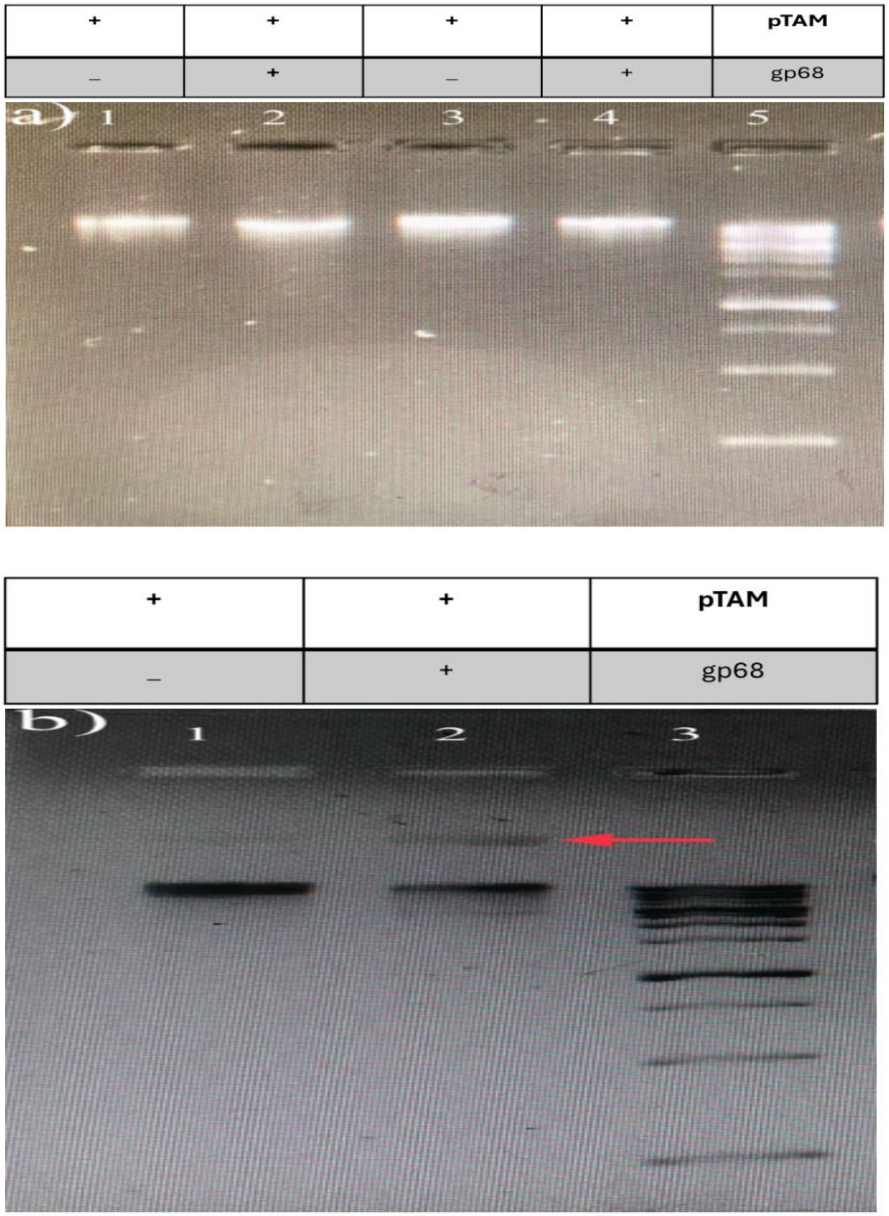


Figure 3.18. 1.5% agarose gels showing cleavage assays of pTAM plasmid with gp68 at 37°C. The presence of cleavage is indicated with a red arrow. a) 1. pTAM only (EDTA buffer), 2. pTAM and gp68 protein (EDTA buffer), 3. pTAM only (Mg^{2+} buffer), 4. pTAM and gp68 protein (Mg^{2+} buffer), 5. Marker. b) pTAM only (Mn^{2+} buffer), 2. pTAM and gp68 protein (Mn^{2+} buffer), 3. Marker

CHAPTER 4

CONCLUSION

Unlike their eukaryotic and bacterial counterparts, the infection cycles of most archaeal viruses are yet to be characterized. This is mainly due to the observation that most of the proteins of these viruses do not possess homology with well-characterized proteins in sequence databases. The ATV genome contains few annotated genes, four of which possess sequence homology with the recently characterized TnpB endonuclease. TnpB proteins are encoded in the IS200/605 transposons, which are widespread in prokaryotes. They possess RNA-guided endonuclease activity and recognize their target DNA sequences using TAM sequences upstream of the target sites.

In this work, all four putative ATV TnpB proteins were analyzed using bioinformatics tools. Results showed that the ATV tnpB genes are not associated with tnpA genes and therefore could be classified as belonging to the IS1341 group. The proteins were expressed and purified in the heterologous E. coli host. Although the expressed proteins appeared to be associated with nucleic acids, nuclease assays suggested that the proteins are associated with DNA and not RNA. This is not surprising since structural studies indicate that TnpB forms contacts with the target DNA in addition to the ωRNA. Therefore, the DNA binding activity observed here may be consistent with expectations, although the absence of RNA binding needs to be confirmed in future studies. The ATV gp68 was demonstrated to possess nonspecific nickase and cleavage activities. It would therefore be interesting to determine if this is indeed the case for other ATV TnpB proteins.

The presence of four tnpB genes in a viral genome is exceptional as viral genomes are highly organized and contain high gene density. The observation that the TnpB-like Fanzors are present in eukaryotic viruses was suggested to indicate that the viruses function as vehicles for the horizontal gene transfer of these elements (Saito et al., 2023). More recently, a viral anti-defense role has been proposed for TnpB proteins encoded on viral genomes. The cyanosiphophage Mic1 infects the bloom-forming cyanobacterium *Microcystis aeruginosa*. In the blooms, several host and cyanophage populations coexist symbiotically, which invariably results in a large-scale exchange of genetic materials

through infection and defense responses. Interestingly, the Mic1 virus displayed a decrease in infectivity when the viral tnpB gene was inactivated by insertion of a host gene encoding an IsrB nickase. Crucially, the Mic1 tnpB gene was determined to be one of a few early genes i.e. expressed immediately following infection (Wang et al., 2024). Such early genes have been determined to be essential for counteracting the host immune system (Stanley et al., 2019). Thus, although it is yet to be determined if the ATV tnpB genes are early genes and therefore function as anti-defense genes, it is interesting to speculate that they the observations reported above for the Mic1 TnpB may also be applicable to ATV TnpB proteins. Future genetic and biochemical characterizations of the ATV TnpB proteins are therefore essential for unraveling the functions of these proteins.

REFERENCES

Altae-Tran, Han, Soumya Kannan, F. Esra Demircioglu, Rachel Oshiro, Suchita P. Nety, Luke J. McKay, Mensur Dlakić, et al. 2021. "The Widespread IS200/IS605 Transposon Family Encodes Diverse Programmable RNA-Guided Endonucleases." *Science* 374 (6563): 57–65. https://doi.org/10.1126/science.abj6856.

Altae-Tran, Han, Sergey Shmakov, Kira S Makarova, Yuri I Wolf, Soumya Kannan, Feng Zhang, and Eugene V Koonin. 2023. "Diversity, Evolution, and Classification of the RNA-Guided Nucleases TnpB and Cas12." *Proceedings of the National Academy of Sciences of the United States of America* 120 (48). https://doi.org/10.1073/pnas.2308224120.

Bao, Weidong, and Jerzy Jurka. 2013. "Homologues of Bacterial TnpB_IS605 Are Widespread in Diverse Eukaryotic Transposable Elements." *Mobile DNA* 4 (1): 12. https://doi.org/10.1186/1759-8753-4-12.

Barabas, Orsolya, Donald R. Ronning, Catherine Guynet, Alison Burgess Hickman, Bao Ton-Hoang, Michael Chandler, and Fred Dyda. 2008. "Mechanism of IS200/IS605 Family DNA Transposases: Activation and Transposon-Directed Target Site Selection." *Cell* 132 (2): 208–20. https://doi.org/10.1016/j.cell.2007.12.029.

Billings, Paul C, John W Orf, Douglas K Palmer, David A Talmage, Cynthia G Pan, and Martin Blumenfeld. 1979. "Anomalous Electrophoretic Mobility of Drosophila Phosphorylated H1 Histone: Is It Related to the Compaction of Satellite DNA into Heterochromatin?" *Nucleic Acids Research* 6 (6): 2151–64. https://doi.org/10.1093/nar/6.6.2151.

Chylinski, Krzysztof, Kira S. Makarova, Emmanuelle Charpentier, and Eugene V. Koonin. 2014. "Classification and Evolution of Type II CRISPR-Cas Systems." *Nucleic Acids Research* 42 (10): 6091–6105. https://doi.org/10.1093/nar/gku241.

Feng, Xu, Ruyi Xu, Jianglan Liao, Jingyu Zhao, Baochang Zhang, Xiaoxiao Xu, Pengpeng Zhao, et al. 2024. "Flexible TAM Requirement of TnpB Enables Efficient Single-Nucleotide Editing with Expanded Targeting Scope." *Nature Communications* 15 (1). https://doi.org/10.1038/s41467-024-47697-4.

Ge, Ruiguang, Rory M. Watt, Xuesong Sun, Julian A. Tanner, Qing-Yu He, Jian-Dong Huang, and Hongzhe Sun. 2005. “Expression and Characterization of a Histidine-Rich Protein, Hpn: Potential for Ni2+ Storage in Helicobacter Pylori.” *Biochemical Journal* 393 (1): 285–93. https://doi.org/10.1042/bj20051160.

Häring, Monika, Gisle Vestergaard, Reinhard Rachel, Lanming Chen, Roger A. Garrett, and David Prangishvili. 2005. “Independent Virus Development Outside a Host.” *Nature* 436 (7054): 1101–2. https://doi.org/10.1038/4361101a.

He, Shuisheng, Alix Corneloup, Catherine Guynet, Laure Lavatine, Anne Caumont-Sarcos, Patricia Siguier, Bernard Marty, Fred Dyda, Michael O Chandler, and B. Ton Hoang. 2015. “The IS *200* /IS *605* Family and ‘Peel and Paste’ Single-Strand Transposition Mechanism” 3 (4). https://doi.org/10.1128/microbiolspec.mdna3-0039-2014.

Hickman, Alison Burgess, Jeffrey A James, Orsolya Barabas, Cécile Pasternak, Bao Ton-Hoang, Michael Chandler, Suzanne Sommer, and Fred Dyda. 2010. “DNA Recognition and the Precleavage State during Single-Stranded DNA Transposition in D. Radiodurans.” *The EMBO Journal* 29 (22): 3840–52. https://doi.org/10.1038/emboj.2010.241.

Isabel Wen Badon, Yeounsun Oh, Ho-Joong Kim, and Seung Hwan Lee. 2024. “Recent Application of CRISPR-Cas12 and OMEGA System for Genome Editing.” *Molecular Therapy* 32 (1): 32–43. https://doi.org/10.1016/j.ymthe.2023.11.013.

Jiang, Kaiyi, Justin Lim, Samantha Sgrizzi, Michael Trinh, Alişan Kayabölen, Natalya Yutin, Weidong Bao, et al. 2023. “Programmable RNA-Guided DNA Endonucleases Are Widespread in Eukaryotes and Their Viruses.” *Science Advances*9 (39). https://doi.org/10.1126/sciadv.adk0171.

Karvelis, Tautvydas, Gytis Druteika, Greta Bigelyte, Karolina Budre, Rimante Zedaveinyte, Arunas Silanskas, Darius Kazlauskas, Česlovas Venclovas, and Virginijus Siksnys. 2021. “Transposon-Associated TnpB Is a Programmable RNA-Guided DNA Endonuclease.” *Nature* 599 (599). https://doi.org/10.1038/s41586-021-04058-1.

Kersulyte, Dangeruta, Natalia S Akopyants, Sandra W Clifton, Bruce A Roe, and Douglas E Berg. 1998. “Novel Sequence Organization and Insertion Specificity of IS605 and IS606: Chimaeric Transposable Elements of Helicobacter Pylori.” *Gene* 223 (1-2): 175–86. https://doi.org/10.1016/s0378-1119(98)00164-4.

Kersulyte, Dangeruta, Billie Velapatiño, Giedrius Dailide, Asish K Mukhopadhyay, Yoshiyuki Ito, Lizbeth Cahuayme, Alan J Parkinson, Robert H Gilman, and Douglas E Berg. 2002. “Transposable Element IS *Hp608* of *Helicobacter Pylori* : Nonrandom Geographic Distribution, Functional Organization, and Insertion Specificity.” *Journal of Bacteriology* 184 (4): 992–1002. https://doi.org/10.1128/jb.184.4.992-1002.2002.

Lee, N., C. Francklyn, and E. P. Hamilton. 1987. “Arabinose-Induced Binding of AraC Protein to AraI2 Activates the AraBAD Operon Promoter.” *Proceedings of the National Academy of Sciences* 84 (24): 8814–18. https://doi.org/10.1073/pnas.84.24.8814.

Meers, Chance, Hoang C. Le, Sanjana R. Pesari, Florian T. Hoffmann, Matt W. G. Walker, Jeanine Gezelle, Stephen Tang, and Samuel H. Sternberg. 2023. “Transposon-Encoded Nucleases Use Guide RNAs to Promote Their Selfish Spread.” *Nature*622 (7984): 863–71. https://doi.org/10.1038/s41586-023-06597-1.

Nakagawa, Ryoya, Hisato Hirano, Satoshi N. Omura, Suchita Nety, Soumya Kannan, Han Altae-Tran, Xiao Yao, et al. 2023. “Cryo-EM Structure of the Transposon-Associated TnpB Enzyme.” *Nature* 616 (April). https://doi.org/10.1038/s41586-023-05933-9.

Nety, Suchita P, Han Altae-Tran, Soumya Kannan, F. Esra Demircioglu, Guilhem Faure, Seiichi Hirano, Kepler Mears, Yugang Zhang, Rhiannon K Macrae, and Feng Zhang. 2023. “The Transposon-Encoded Protein TnpB Processes Its Own MRNA into ΩRNA for Guided Nuclease Activity.” *The CRISPR Journal* 6 (3): 232–42. https://doi.org/10.1089/crispr.2023.0015.

Pasternak, Cécile, Rémi Dulermo, Bao Ton-Hoang, Robert Debuchy, Patricia Siguier, Geneviève Coste, Michaël Chandler, and Suzanne Sommer. 2013. “IS*Dra2*Transposition in*DEinococcus Radiodurans*Is Downregulated by TnpB.” *Molecular Microbiology* 88 (2): 443–55. https://doi.org/10.1111/mmi.12194.

Prangishvili, David, Gisle Vestergaard, Monika Häring, Ricardo Aramayo, Tamara Basta, Reinhard Rachel, and Roger A. Garrett. 2006. "Structural and Genomic Properties of the Hyperthermophilic Archaeal Virus ATV with an Extracellular Stage of the Reproductive Cycle." *Journal of Molecular Biology* 359 (5): 1203–16. https://doi.org/10.1016/j.jmb.2006.04.027.

Saito, Makoto, Peiyu Xu, Guilhem Faure, Samantha Maguire, Soumya Kannan, Han Altae-Tran, Sam Vo, AnAn Desimone, Rhiannon K. Macrae, and Feng Zhang. 2023. "Fanzor Is a Eukaryotic Programmable RNA-Guided Endonuclease." *Nature*, June, 1–3. https://doi.org/10.1038/s41586-023-06356-2.

Sasnauskas, Giedrius, G. Tamulaitiene, Gytis Druteika, Arturo Carabias, Arūnas Šilanskas, Darius Kazlauskas, Česlovas Venclovas, Guillermo Montoya, Tautvydas Karvelis, and Virginijus Šikšnys. 2023. "TnpB Structure Reveals Minimal Functional Core of Cas12 Nuclease Family." *Nature* 616 (7956): 384–89. https://doi.org/10.1038/s41586-023-05826-x.

Sharp, Paul M., and Michael Bulmer. 1988. "Selective Differences among Translation Termination Codons." *Gene* 63 (1): 141–45. https://doi.org/10.1016/0378-1119(88)90553-7.

Shi, Yunhua, Richard A. Mowery, Jonathan Ashley, Michelle Hentz, Alejandro J. Ramirez, Basar Bilgicer, Hilda Slunt-Brown, David R. Borchelt, and Bryan F. Shaw. 2012. "Abnormal SDS-PAGE Migration of Cytosolic Proteins Can Identify Domains and Mechanisms That Control Surfactant Binding." *Protein Science* 21 (8): 1197–1209. https://doi.org/10.1002/pro.2107.

Siguier, Patricia, Edith Gourbeyre, and Mick Chandler. 2014. "Bacterial Insertion Sequences: Their Genomic Impact and Diversity." *FEMS Microbiology Reviews* 38 (5): 865–91. https://doi.org/10.1111/1574-6976.12067.

Siguier, Patricia, Edith Gourbeyre, Alessandro Varani, Bao Ton-Hoang, and Mick Chandler. 2015. "Everyman's Guide to Bacterial Insertion Sequences." *Microbiology Spectrum* 3 (2). https://doi.org/10.1128/microbiolspec.mdna3-0030-2014.

Stanley, Sabrina Y., Adair L. Borges, Kuei-Ho Chen, Danielle L. Swaney, Nevan J. Krogan, Joseph Bondy-Denomy, and Alan R. Davidson. 2019. "Anti-CRISPR-Associated Proteins Are Crucial Repressors of Anti-CRISPR Transcription." *Cell* 178 (6): 1452-1464.e13. https://doi.org/10.1016/j.cell.2019.07.046.

Wang, Xiao-Qian, Kang Du, Chaoyi Chen, Pu Hou, Wei-Fang Li, Yuxing Chen, Qiong Li, and Cong-Zhao Zhou. 2024. "Profiling the Interplay and Coevolution of *Microcystis Aeruginosa* and Cyanosiphophage Mic1." *Microbiology Spectrum*, May. https://doi.org/10.1128/spectrum.00298-24.

Wiegand, Tanner, Florian T Hoffmann, Matt W.G. Walker, Stephen Tang, Egill Richard, Hoang C Le, Chance Meers, and Samuel H Sternberg. 2023. "Emergence of RNA-Guided Transcription Factors via Domestication of Transposon-Encoded TnpB Nucleases." *BioRxiv (Cold Spring Harbor Laboratory)*, November. https://doi.org/10.1101/2023.11.30.569447.

Xu, Ying, Tao Liu, Jing Wang, Binyang Xiong, Ling Liu, and Nan Peng. 2023. "Reprogramming an RNA-Guided Archaeal TnpB Endonuclease for Genome Editing." *Cell Discovery* 9 (1). https://doi.org/10.1038/s41421-023-00615-2.

Ziemer, M A, A Mason, and D M Carlson. 1982. "Cell-Free Translations of Proline-Rich Protein MRNAs." *Journal of Biological Chemistry* 257 (18): 11176–80. https://doi.org/10.1016/s0021-9258(18)33948-6.

APPENDIX A

ATV Putative TnpB DNA and Protein Sequences

A. ATV putative *tnpB* gene sequences:

1. GP10:

ATGCCCGACGTAGGGATTAGGTTTAGGGCTTTCGCAGATGAACCTACAATTAGG
GAGTTAAAATCCCAGTTGAGGTTAGCGTGCGAAATTTACAACACCCTACGTTGG
GCAGATATATATTTCTACCAAAGAGATGGGAAAGGTCTAAGCAAGACTGAATTG
AGACAACTCGCTCTAGATTTGAGAAAACAGGACGATGAATACAAGAGAATCTAT
TCTCAAGCAGTACAACAAATTGCAGACAGATTCTACGACGCTAAGAAGAGATTC
CTCAAAGGGTTAGCACGCTACCCCAGGGAAAAGAAACCACATAAATGGTATTCG
TTAGTGTACCCTCAATCAGGTTGGAAAGTGTTAGAGACGAGAGAAATAAGAACT
AAAAGCAAGAAGAACAAGAAGAAGATAATGACGCTTCAGTTAGCAAACCTGGGA
GTCTTCAAAGTTATTGTTCATAGGGACTTTCCGCTAGACAAGGTAAAGAGGGTG
ATAGTTAAGTTAACACCTTCAGAGAGAGTGTACATTAGTTTCGTTGTGGATCAA
GAATACCCTCAACTCCCTAAGACAAACAAAACGGTTGGAATAGATGTTGGAATA
GAGAAACTTCTCATTACATCAGATGGGGAATACGTGCCAAACTTGCGACCTTAC
GAGAAGGCACTCAATAAGACAAGAAAACTTCATAAAGCTCTCTCGAGGAAGAAG
TTCTTGTCGCACAACTGGTTTAAGGCAAAGATTAAGCTCGCAAGGGCTTATGAG
CATTTGAAGAATCTTAGGAGGGACATATACATGAAACTTGGGAAGTATTTTGCA
GAGCATTATGACGTAGTCGTAATGGAAGATATTCAAGTTAAACAACTTGTTGGT
AAATCCTACAAGAAAATGAGGATGAGGCTTCACGATGTTGCATTTTACGAGCTT
AGGAGCATCATGGAATATCAACTTAGGAAGTACGGGAAGGAGCTTCGTCTCGTT
GATCCTGCTTTTACTTCGATGACTTGTGCTAAATGCGGGTATGTTAAGAAGGAT
TTAACTTTGGCTGATCGTATTTTTGTTTGTCCAAAATGTGGTTGGACTGTGGAT
CGTGACTACAATGCTTCTCTTAACATTCTAAAAAGGTCGGGGTGGGAGCCACCC
TTAGCGCCTGTGGAGCTCCGCCCTCTACCATTGGTAAAAAGCCAATGGCAAGGT
GGGGCTTTGAAGCAGGAAGCCTCGCCCTTTAGGGTGGGGTAG

2. GP40

ATGCCACCCTCCTCTGGTCAACTCTTAGGGGATGAGGAGCGGGAGCCGACTTCT
ACTCCCGCAATACCGGAGGAGGGTGTCTACAAAGTTAAATACTCAAATCGAAGA
ACCAATATAGTTCGCCTTCTTCCAAATGGTTTTCAAGAAAGAAAACTGAGGAGG
TTAGCAGACCTCTCTGCAAAGCTCTTCAACGAAGTTAACTATGAAAGGAGGCAA
CAGTTCTTTCACGAAGGGAAAGTGGACATTAAGGGGACATATAAAAAATATTAT
GAGAAGTATAAGGAAAAATTACGTACTAACGCACAAGCTGTTCTCAATAAGAAC
AACGAAGCGTGGTCGTCATTCTTCTCCTTGCTAAATCTGAAGAAGGAAGGAAAA
TTACCACAACACATAAAACACGTTTCTCCACCCGGGTACTGTAAGGACAGAAAG
ACGAAGAAGAGAAAGCTAATACTAATCGTTAGACAGGATCGTTACAAGGTAGAT
GCTGAAAACAACAAATTAATCCTCAAGGACTTTAACATGGAAATAGAATTCGTC
GGTAGACTTAGGTGGTATGGTAAACAAGGTAGGCTAGAGATAATTTTTGACGAG
ACTAGAAATGCTTGGTACGCTCACATACCAGTTGAGGTAGGTGTTGAGGAAACT
GGAAAGAAGAGTAAGCACGTCGTCAAGGGTGAGAGAAAGAGTATCCAGATTGCC
AAACCAAAGGGAAACAAAGTAGCATCCATCGATTTAGGTATTAACGTAATAGCA

AGTGTAGTGGTAAGTGATGGCACTTGGTTGCTCTACAAGGGTATTAGAACAAAG
GAGGACTACTTCTACTTCCATAAGAGAATTGCAGAAGTACAATCGTTAGCTGAC
AGAACTAGAAACATTGGTGAATATGAGGCATATCTAGAATTGTTAAGGGAAGAG
AGGAGACTGTTTAAGAAACTGAAGAGACGCCTTCTCCATCTCTACCGTAACTTA
GCCTCTCACCTTATTAAAACACTCCACGAACTAGGAGTATCAACAATTTACCTA
GGTAACCCATTCAATATCGTGCAGGAGAAAGGCGACAACTTCATGACAGACAAG
TGGCCTTATCGTAAGTTGATGCATGCCATTGAATTAAAGGCTCAAGAATATGGT
ATGAAGGTGTATGAAGTTGATGAGTACAATACAACAAAGTATTGTGCTTACCAC
GACGTGAAAGTAAAGAGGAACCCAAGAGGAGTAGTTATTTGTCCTAAGGGACAT
AAACTTCACAGCGATCTAAACGGTGCGTTAAACATTCTAAAGAAAGCAGTAGGA
GTGGTCGTTAACGAAGTGAAAAAACCCTTGTCTTTCATTGTGGATCATAACAGA
GTAGCACCCATCAAAGGGGTGTAA

3. GP43

ATGGCTAGGAGGGAGAAAAACCCAATCAGAGCAACGGTTTCGATGAAGATTGGC
TTATCTGACTCCCTCCTAGCCTTCGTGAACAACTACGTTAGAGCACTCCGTTTC
TCCATATTTTGGATGAAAGAGAATGTGAAAAATCCAAACGAGAAGGGCACGCTT
TCTAAAGTGCACGAGGGATTATATGGAAAGCTAAAGGAGGATTATAATCTACCA
CCCAAAGTTTCTGCGGACTGTTATCGTGATGCCCTGGCAATATATAAGAGTTGG
TATAACAATCCCAAGAGGGGTAGATTCCCCCGCGTCTATAAGCCCACTGTATGG
TTAACGCCCAAGCGAAGTTATACTGTAGACTTAGATAGAATGGTAGTTAAGATT
ACCAGTGTTGGGAACTGCCAATTTTAGGCTATCCTAGGAACTTAAAAGAGTATG
CAAACTGGGATATGAAGGAGGCTAGGCTAACAATCAGAGATGGCAAAGCTTTCC
TCAAGGTGGTTTTTGAGAAACCGAAAGTTAAGATACAACCCAAAGGTAGTGTTG
CCGTTGATATTAACATGAGTGAAATTGTAGTAGGGAAGGACGACAGTCACTACG
TTAGGATTCCCACTCGCCTTCACGAGTCTCACCACTTCAAGACATTAGCTGAGA
ATTTGCAAAAGAAGTATCCTAGAAGGTGGAAGGAGAACAAGAGAATTCTACACA
GAATACGCTCTTTTCATCACAAGGCCAAACTAATTATGGAGGACTTCGCTAGGA
AAGTTGGCAAGTGGGTTGTTGAGATTGCTTGGGATTTGGGTGCCAACGTAATCA
AATTGGAGAATCTTAAGAACCTCATCAGGAACGTCAACAAACTGTCAGCCGAGT
TTCGCGATAAACTCTACTTGATGCAATATCATCGTATTCAGTACTGGATAGAAT
GGCAAGCCAGAAAACACGGAATGAATGTGGAGTATGTTAATCCTAGTTACTCTT
CCGTCTCTTGTCCAAAGTGTGGCCGCAAAATGGTTGAGATTGCTCATAGGTACT
TCCACTGTCCCTCGTGTGGTTATGAGAACGATCGTGACGTTATTGCTATCATGA
ATTTAAATGGGAGGGGGTCTCTGACCCTCTCGACTGCCCCTCAGATGAGAGATG
TAATCCCGAATCGATGA

4. GP68

ATGGCTAGGAGGGAGAAAAACCCAATAAGAGCCACAGTTGCGATGAATATTGGC
TTATCTGACTCCCTCCTAGCCTTCGTGAACAACTACGTTAAGGCACTCCGTTTC
TCCTTGTTCTGGATGAAAGAGAATGTGAAAAATCCAAACGAGAAGGGCGCACTC
TCCAAAGTACACGAGGGATTGTATGAAAAGCTGAGGAAAGAATACAATCTACCA
TCTAAAGTTACTGAGGACTGTTATCGTGATGCCCTGGCAATATACAAGAGTTGG
TACAACAACCCGAAAAAAGGTAGATTCCCCCGTGTCTACAAGCCGACAGTGTGG
CTAACGCCCAAGCAAAGTTATACTGTAGACTTAGAGAAAATGACAGTCAGGATA
GCAAGTGTTGGCGAACTACCAATACTAGGTTATCCTAGAAACCTAAAGGAGTAT
GCAAACTGGAAGATGAAGGAGGCTAGGCTAACAATCAAGGATGGCAAGGCTCTC
CTCAAAGTAACTTTTGAGAAGGAAGAAGAGAAAGTTAAACCAAAAGACAGTGTT

GCTGTTGATATAAACATGAATGACATTGTCGTTGGTAAGGATGACACTCACTAC
GTTAGGATTCCCACTCGCCTTCACGACGCTCACCACTTCAAGTCATTAGCTGAG
AATTTGCAGAAGAAGTATCCTAGAAGGTGGAAGCAAAATAGGAGAATTCTACAC
AGGGCACGCTCTTTTCATCAAAAGGCCAAACTAATTATGGAGGACTACGCTAGG
AAGGTTGGTAAGTGGGTTGTTGAGATTGCTGAGGGTTTGGGTGCCAACGTCATA
AAGCTTGAGGACTTGAAGAACCTCATCAAGGACGTTAATAAGCTACCAGCTGAA
TTTCGCGATAAACTATACTTGATGCAATATCGTCGTATTCAGTATTGGATAGAG
TGGCAGGCTAAGAAACACGGAATGATTGTGGAGTTTGTTAATCCTAGTTACTCT
TCCGTTTCTTGCCCAAAGTGTGGCCACAAAATGGTTGAGATTGCTTATAGGTAC
TTTCACTGTCCTTCATGTGGTTATGAGAACGATCGTGATGTTATTGCTATCATG
AATTTAAATGGGAGGGGGTCTCTGACCCTCTCGACTGCCCCTCAAATGAGAGAT
GTAGCTCCGAATCGATGA

B. ATV putative TnpB protein sequences:

1. GP10

MPDVGIRFRAFADEPTIRELKSQLRLACEIYNTLRWADIYFYQRDGKGLSKTEL
RQLALDLRKQDDEYKRIYSQAVQQIADRFYDAKKRFLKGLARYPREKKPHKWYS
LVYPQSGWKVLETREIRTKSKKNKKKIMTLQLANLGVFKVIVHRDFPLDKVKRV
IVKLTPSERVYISFVVDQEYPQLPKTNKTVGIDVGIEKLLITSDGEYVPNLRPY
EKALNKTRKLHKALSRKKFLSHNWFKAKIKLARAYEHLKNLRRDIYMKLGKYFA
EHYDVVVMEDIQVKQLVGKSYKKMRMRLHDVAFYELRSIMEYQLRKYGKELRLV
DPAFTSMTCAKCGYVKKDLTLADRIFVCPKCGWTVDRDYNASLNILKRSGWEPP
LAPVELRPLPLVKSQWQGGALKQEASPFRVG

2. GP40

MPPSSGQLLGDEEREPTSTPAIPEEGVYKVKYSNRRTNIVRLLPNGFQERKLRR
LADLSAKLFNEVNYERRQQFFHEGKVDIKGTYKKYYEKYKEKLRTNAQAVLNKN
NEAWSSFFSLLNLKKEGKLPQHIKHVSPPGYCKDRKTKKRKLILIVRQDRYKVD
AENNKLILKDFNMEIEFVGRLRWYGKQGRLEIIFDETRNAWYAHIPVEVGVEET
GKKSKHVVKGERKSIQIAKPKGNKVASIDLGINVIASVVVSDGTWLLYKGIRTK
EDYFYFHKRIAEVQSLADRTRNIGEYEAYLELLREERRLFKKLKRRLLHLYRNL
ASHLIKTLHELGVSTIYLGNPFNIVQEKGDNFMTDKWPYRKLMHAIELKAQEYG
MKVYEVDEYNTTKYCAYHDVKVKRNPRGVVICPKGHKLHSDLNGALNILKKAVG
VVVNEVKKPLSFIVDHNRVAPIKGV

3. GP43

MARREKNPIRATVSMKIGLSDSLLAFVNNYVRALRFSIFWMKENVKNPNEKGTL
SKVHEGLYGKLKEDYNLPPKVSADCYRDALAIYKSWYNNPKRGRFPRVYKPTVW
LTPKRSYTVDLDRMVVKITSVGELPILGYPRNLKEYANWDMKEARLTIRDGKAF
LKVVFEKPKVKIQPKGSVAVDINMSEIVVGKDDSHYVRIPTRLHESHHFKTLAE
NLQKKYPRRWKENKRILHRIRSFHHKAKLIMEDFARKVGKWVVEIAWDLGANVI
KLENLKNLIRNVNKLSAEFRDKLYLMQYHRIQYWIEWQARKHGMNVEYVNPSYS
SVSCPKCGRKMVEIAHRYFHCPSCGYENDRDVIAIMNLNGRGSLTLSTAPQMRD
VIPNR

4. GP68

MARREKNPIRATVAMNIGLSDSLLAFVNNYVKALRFSLFWMKENVKNPNEKGAL
SKVHEGLYEKLRKEYNLPSKVTEDCYRDALAIYKSWYNNPKKGRFPRVYKPTVW

LTPKQSYTVDLEKMTVRIASVGELPILGYPRNLKEYANWKMKEARLTIKDGKAL
LKVTFEKEEEKVKPKDSVAVDINMNDIVVGKDDTHYVRIPTRLHDAHHFKSLAE
NLQKKYPRRWKQNRRILHRARSFHQKAKLIMEDYARKVGKWVVEIAEGLGANVI
KLEDLKNLIKDVNKLPAEFRDKLYLMQYRRIQYWIEWQAKKHGMIVEFVNPSYS
SVSCPKCGHKMVEIAYRYFHCPSCGYENDRDVIAIMNLNGRGSLTLSTAPQMRD
VAPNR

APPENDIX B

NCBI BLAST Search Results

1. NCBI BLAST search results of all ATV putative TnpB proteins (Top 50 hits)
   A. ATV TnpB GP10

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
transposase [Bicaudavirus pozzuoliense] ATV TnpB GP10	Bicaudavirus pozzuoliense	840	840	100%	0.0	100.00%	409	YP_319841.1
putative transposase [Acidianus two-tailed virus 2]	Acidianus two-tailed virus 2	775	775	100%	0.0	93.40%	409	AQN96422.1
transposase [Sulfolobales archaeon]	Sulfolobales archaeon	692	692	100%	0.0	81.46%	410	MCQ4343084.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	677	677	97%	0.0	78.89%	398	WP_014513621.1
RNA-guided endonuclease TnpB family protein [Acidianus brierleyi]	Acidianus brierleyi	676	676	97%	0.0	80.65%	401	WP_110271413.1
RNA-guided endonuclease TnpB family protein [Acidianus brierleyi]	Acidianus brierleyi	676	676	97%	0.0	80.90%	401	WP_110269789.1
transposase [Sulfolobus islandicus]	Sulfolobus islandicus	675	675	97%	0.0	78.64%	398	WP_272584808.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	675	675	97%	0.0	78.14%	398	WP_014513368.1
transposase [Sulfolobus islandicus]	Sulfolobus islandicus	675	675	97%	0.0	77.89%	398	WP_272584891.1
IS200/IS605 family OrfB protein [Sulfolobus islandicus HVE10/4]	Sulfolobus islandicus HVE10/4	674	674	97%	0.0	79.65%	417	ADX82340.1
transposase [Sulfolobus islandicus]	Sulfolobus islandicus	674	674	97%	0.0	78.14%	398	WP_272584852.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	674	674	97%	0.0	79.65%	401	WP_052891907.1
RNA-guided endonuclease TnpB family protein [Sulfurisphaera ohwakuensis]	Sulfurisphaera ohwakuensis	674	674	100%	0.0	77.56%	410	WP_156015466.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	673	673	97%	0.0	78.14%	398	WP_014513579.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	673	673	97%	0.0	78.14%	398	WP_014512062.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	673	673	97%	0.0	78.39%	398	WP_014513086.1
transposase [Sulfolobus islandicus]	Sulfolobus islandicus	673	673	97%	0.0	77.89%	398	WP_272584914.1
RNA-guided endonuclease TnpB family protein [Acidianus brierleyi]	Acidianus brierleyi	672	672	97%	0.0	80.65%	401	WP_110271150.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	672	672	97%	0.0	77.89%	398	WP_048050170.1
RNA-guided endonuclease TnpB family protein [Acidianus brierleyi]	Acidianus brierleyi	672	672	97%	0.0	78.64%	398	WP_110269896.1
RNA-guided endonuclease TnpB family protein [Acidianus brierleyi]	Acidianus brierleyi	672	672	97%	0.0	79.90%	401	WP_110271670.1
RNA-guided endonuclease TnpB family protein [Sulfurisphaera ohwakuensis]	Sulfurisphaera ohwakuensis	672	672	100%	0.0	77.32%	410	WP_156015061.1
RNA-guided endonuclease TnpB family protein [Acidianus brierleyi]	Acidianus brierleyi	672	672	97%	0.0	79.90%	401	WP_110269557.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	671	671	97%	0.0	79.65%	401	WP_152939228.1
RNA-guided endonuclease TnpB family protein [Sulfurisphaera ohwakuensis]	Sulfurisphaera ohwakuensis	671	671	100%	0.0	77.32%	410	WP_156015096.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	671	671	97%	0.0	79.15%	401	WP_014513968.1
RNA-guided endonuclease TnpB family protein [Acidianus brierleyi]	Acidianus brierleyi	671	671	97%	0.0	80.15%	401	WP_110271266.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	671	671	97%	0.0	77.64%	398	WP_014512015.1
RNA-guided endonuclease TnpB family protein [Acidianus brierleyi]	Acidianus brierleyi	670	670	97%	0.0	79.40%	401	WP_110269973.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	670	670	97%	0.0	79.15%	401	WP_014514382.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	669	669	97%	0.0	77.64%	398	WP_014512091.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	669	669	97%	0.0	78.89%	401	WP_014512187.1
RNA-guided endonuclease TnpB family protein [Acidianus brierleyi]	Acidianus brierleyi	669	669	97%	0.0	80.15%	401	WP_110269487.1
RNA-guided endonuclease TnpB family protein [Acidianus brierleyi]	Acidianus brierleyi	669	669	97%	0.0	79.65%	401	WP_110271249.1
RNA-guided endonuclease TnpB family protein [Sulfurisphaera ohwakuensis]	Sulfurisphaera ohwakuensis	669	669	100%	0.0	77.56%	410	WP_156014956.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	669	669	97%	0.0	78.89%	401	WP_014514154.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	669	669	97%	0.0	78.89%	401	WP_014512677.1
RNA-guided endonuclease TnpB family protein [Sulfurisphaera ohwakuensis]	Sulfurisphaera ohwakuensis	669	669	100%	0.0	77.32%	410	WP_156014799.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	669	669	97%	0.0	78.64%	401	WP_014513663.1
RNA-guided endonuclease TnpB family protein [Sulfurisphaera ohwakuensis]	Sulfurisphaera ohwakuensis	668	668	100%	0.0	77.32%	410	WP_156014298.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	668	668	97%	0.0	78.64%	401	WP_014513771.1
RNA-guided endonuclease TnpB family protein [Sulfurisphaera ohwakuensis]	Sulfurisphaera ohwakuensis	667	667	100%	0.0	77.07%	410	WP_156015548.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	667	667	97%	0.0	78.89%	401	WP_014513907.1
RNA-guided endonuclease TnpB family protein [Sulfurisphaera ohwakuensis]	Sulfurisphaera ohwakuensis	667	667	100%	0.0	77.07%	410	WP_156013951.1
RNA-guided endonuclease TnpB family protein [Sulfurisphaera ohwakuensis]	Sulfurisphaera ohwakuensis	667	667	100%	0.0	77.07%	410	WP_156015603.1
RNA-guided endonuclease TnpB family protein [Sulfurisphaera ohwakuensis]	Sulfurisphaera ohwakuensis	667	667	100%	0.0	76.83%	410	WP_156013978.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	666	666	97%	0.0	77.39%	398	WP_014512503.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	665	665	97%	0.0	78.39%	401	WP_014514182.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	665	665	97%	0.0	78.64%	401	WP_014513687.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	665	665	97%	0.0	78.39%	401	WP_014514302.1

B. ATV TnpB GP40

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
transposase [Bicaudavirus pozzuoliense] **ATV TnpB GP40**	Bicaudavirus pozzuoliense	935	935	100%	0.0	100.00%	457	YP_319871.1
IS605 family OrfB family transposase [Acidianus two-tailed virus 2]	Acidianus two-tailed virus 2	801	801	100%	0.0	89.28%	457	AON96453.1
RNA-guided endonuclease TnpB family protein [Saccharolobus solfataricus]	Saccharolobus solfataricus	783	783	99%	0.0	85.78%	460	WP_020936649.1
RNA-guided endonuclease TnpB family protein [Saccharolobus solfataricus]	Saccharolobus solfataricus	783	783	100%	0.0	84.90%	459	WP_063492729.1
RNA-guided endonuclease TnpB family protein [Acidianus infernus]	Acidianus infernus	764	764	100%	0.0	85.12%	457	WP_155864241.1
transposase [Saccharolobus solfataricus]	Saccharolobus solfataricus	764	764	99%	0.0	85.78%	460	QPG49180.1
RNA-guided endonuclease TnpB family protein [Saccharolobus solfataricus]	Saccharolobus solfataricus	763	763	100%	0.0	83.59%	455	WP_010923094.1
RNA-guided endonuclease TnpB family protein [Saccharolobus solfataricus]	Saccharolobus solfataricus	763	763	99%	0.0	86.00%	460	WP_010923129.1
transposase [Acidianus infernus]	Acidianus infernus	760	760	100%	0.0	82.39%	462	MCY0882576.1
transposase [Acidianus infernus]	Acidianus infernus	757	757	100%	0.0	84.53%	458	MCY0882339.1
transposase [Acidianus infernus]	Acidianus infernus	742	742	100%	0.0	85.87%	475	MCY0883177.1
RNA-guided endonuclease TnpB family protein [Acidianus infernus]	Acidianus infernus	735	735	95%	0.0	84.86%	437	WP_155864309.1
transposase [Caldivirga sp.]	Caldivirga sp.	718	718	100%	0.0	76.47%	479	WP_291998851.1
transposase [Candidatus Marsarchaeota G2 archaeon ECH_B_SAG-C16]	Candidatus Marsarchaeota G...	691	691	100%	0.0	75.60%	459	PSN94592.1
transposase [Thermocladium sp. ECH_B]	Thermocladium sp. ECH_B	680	680	99%	0.0	72.49%	456	KUO91201.1
transposase [Thermocladium sp. ECH_B]	Thermocladium sp. ECH_B	679	679	99%	0.0	71.93%	455	KUO92828.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	679	679	99%	0.0	70.59%	462	WP_218258102.1
transposase [Acidilobus sp.]	Acidilobus sp.	679	679	100%	0.0	74.73%	458	NAZ38625.1
transposase, IS605 OrfB family, central region [uncultured Acidilobus sp. CIS]	uncultured Acidilobus sp. CIS	679	679	100%	0.0	74.73%	458	ESQ20458.1
transposase [Thermocladium sp. ECH_B]	Thermocladium sp. ECH_B	679	679	99%	0.0	72.21%	458	KUO92272.1
transposase [Nitrososphaerota archaeon]	Nitrososphaerota archaeon	679	679	99%	0.0	73.52%	460	NAY81525.1
transposase [Candidatus Marsarchaeota G2 archaeon BE_D]	Candidatus Marsarchaeota G...	677	677	99%	0.0	75.05%	457	PSO06950.1
transposase [Saccharolobus sp.]	Saccharolobus sp.	675	675	100%	0.0	73.42%	461	WP_342815514.1
transposase [Saccharolobus sp.]	Saccharolobus sp.	675	675	100%	0.0	73.42%	461	WP_342810844.1
transposase [Saccharolobus sp.]	Saccharolobus sp.	675	675	100%	0.0	73.42%	461	WP_342813384.1
transposase [Thermocladium sp. ECH_B]	Thermocladium sp. ECH_B	675	675	99%	0.0	71.71%	456	KUO91942.1
transposase [Saccharolobus sp.]	Saccharolobus sp.	674	674	100%	0.0	73.42%	461	WP_342782551.1
transposase [Candidatus Marsarchaeota G2 archaeon ECH_B_SAG-F08]	Candidatus Marsarchaeota G...	673	673	99%	0.0	73.68%	457	PSN97080.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	673	673	99%	0.0	72.71%	478	WP_218266257.1
RNA-guided endonuclease TnpB family protein [Thermocladium modestius]	Thermocladium modestius	672	672	100%	0.0	71.90%	457	WP_188596811.1
transposase [Vulcanisaeta sp.]	Vulcanisaeta sp.	672	672	99%	0.0	72.59%	471	MCG2864476.1
transposase [Thermocladium sp. ECH_B]	Thermocladium sp. ECH_B	672	672	99%	0.0	71.18%	458	KUO91746.1
transposase [Candidatus Marsarchaeota G1 archaeon OSP_D]	Candidatus Marsarchaeota G...	672	672	99%	0.0	76.37%	459	PSN83235.1
transposase [Desulfurococcaceae archaeon]	Desulfurococcaceae archaeon	672	672	96%	0.0	75.51%	441	MCC6023025.1
transposase [Candidatus Marsarchaeota G2 archaeon ECH_B_1]	Candidatus Marsarchaeota G...	671	671	95%	0.0	72.15%	436	PSN98713.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	671	671	99%	0.0	72.49%	478	WP_218266143.1
transposase [Vulcanisaeta souniana]	Vulcanisaeta souniana	670	670	95%	0.0	76.94%	440	WP_349293703.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	670	670	99%	0.0	69.28%	462	WP_218266006.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	669	669	99%	0.0	69.50%	462	WP_218257947.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	667	667	99%	0.0	69.72%	461	WP_218260647.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	665	665	99%	0.0	74.45%	461	WP_218266877.1
transposase [Saccharolobus]	Saccharolobus	664	664	99%	0.0	70.59%	462	WP_240781848.1
transposase [Thermocladium sp. ECH_B]	Thermocladium sp. ECH_B	664	664	99%	0.0	70.96%	457	KUO92865.1
transposase [Nitrososphaerota archaeon]	Nitrososphaerota archaeon	663	663	99%	0.0	75.33%	459	NAY82412.1
transposase [Saccharolobus shibatae]	Saccharolobus shibatae	663	663	99%	0.0	71.62%	479	WP_261310058.1
transposase [Thermocladium sp. ECH_B]	Thermocladium sp. ECH_B	660	660	99%	0.0	71.71%	475	KUO93142.1
transposase [Candidatus Marsarchaeota G2 archaeon BE_D]	Candidatus Marsarchaeota G...	659	659	99%	0.0	75.66%	457	PSO03791.1
transposase [Vulcanisaeta souniana JCM 11219]	Vulcanisaeta souniana JCM 1...	645	645	91%	0.0	77.20%	423	BDR93565.1
transposase [Thermocladium sp. ECH_B]	Thermocladium sp. ECH_B	639	639	95%	0.0	70.55%	435	KUO91523.1
transposase [Saccharolobus sp.]	Saccharolobus sp.	637	637	94%	0.0	73.73%	434	WP_342813535.1

C. ATV TnpB GP43

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
transposase [Bicaudavirus pozzuoliense] ATV TnpB GP43	Bicaudavirus pozzuoliense	793	793	100%	0.0	100.00%	383	YP_319874.1
IS605 family OrfB transposase [Acidianus two-tailed virus 2]	Acidianus two-tailed virus 2	709	709	100%	0.0	87.73%	383	AON96456.1
RNA-guided endonuclease TnpB family protein [Acidianus infernus]	Acidianus infernus	699	699	100%	0.0	85.90%	383	WP_155863494.1
RNA-guided endonuclease TnpB family protein [Acidianus infernus]	Acidianus infernus	694	694	100%	0.0	85.12%	383	WP_155864291.1
transposase [Bicaudavirus pozzuoliense]	Bicaudavirus pozzuoliense	686	686	100%	0.0	86.16%	383	YP_319899.1
RNA-guided endonuclease TnpB family protein [Saccharolobus solfataric...	Saccharolobus solfataricus	677	677	100%	0.0	83.29%	383	WP_020936635.1
RNA-guided endonuclease TnpB family protein [Acidianus infernus]	Acidianus infernus	669	669	100%	0.0	83.55%	383	WP_155864207.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	630	630	100%	0.0	79.11%	381	WP_048052744.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	629	629	100%	0.0	77.55%	382	WP_152939397.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	627	627	100%	0.0	77.02%	382	WP_152941266.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	627	627	100%	0.0	78.59%	381	WP_013776664.1
RNA-guided endonuclease TnpB family protein [Sulfolobales archaeon]	Sulfolobales archaeon	627	627	100%	0.0	77.34%	382	MCG2872300.1
RNA-guided endonuclease TnpB family protein [Sulfuracidifex metallicus]	Sulfuracidifex metallicus	626	626	100%	0.0	77.81%	382	WP_054838422.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	625	625	100%	0.0	76.76%	382	WP_152939166.1
RNA-guided endonuclease TnpB family protein [Sulfolobales archaeon]	Sulfolobales archaeon	625	625	100%	0.0	78.39%	387	MCQ4335705.1
RNA-guided endonuclease TnpB family protein [Sulfuracidifex metallicus]	Sulfuracidifex metallicus	625	625	100%	0.0	78.07%	382	WP_156016674.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	624	624	100%	0.0	76.50%	382	WP_013774993.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	623	623	100%	0.0	76.56%	383	WP_013775100.1
RNA-guided endonuclease TnpB family protein [Saccharolobus solfataric...	Saccharolobus solfataricus	622	622	100%	0.0	78.07%	381	WP_009992247.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	621	621	100%	0.0	77.34%	383	WP_013776434.1
RNA-guided endonuclease TnpB family protein [Saccharolobus sp.]	Saccharolobus sp.	619	619	100%	0.0	74.48%	384	WP_342782880.1
RNA-guided endonuclease TnpB family protein [Sulfuracidifex metallicus]	Sulfuracidifex metallicus	619	619	100%	0.0	75.98%	382	WP_156016140.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	619	619	100%	0.0	75.98%	382	WP_152943027.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	619	619	100%	0.0	76.56%	383	WP_152942809.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	617	617	100%	0.0	75.98%	382	WP_013775197.1
RNA-guided endonuclease TnpB family protein [Acidianus sp. HS-5]	Acidianus sp. HS-5	617	617	100%	0.0	75.46%	383	WP_236752621.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	617	617	100%	0.0	73.70%	384	WP_218259439.1
RNA-guided endonuclease TnpB family protein [Sulfuracidifex metallicus]	Sulfuracidifex metallicus	617	617	100%	0.0	77.81%	381	WP_156016645.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	616	616	100%	0.0	73.96%	384	WP_218267400.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	616	616	100%	0.0	73.96%	384	WP_218261170.1
RNA-guided endonuclease TnpB family protein [Saccharolobus solfataric...	Saccharolobus solfataricus	616	616	100%	0.0	75.98%	382	WP_010923706.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	616	616	100%	0.0	73.96%	384	WP_218260425.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	616	616	100%	0.0	75.98%	382	WP_013776628.1
transposase [Candidatus Marsarchaeota G1 archaeon OSP_C]	Candidatus Marsarchaeota G1 arc...	615	615	100%	0.0	75.46%	381	PSN87618.1
RNA-guided endonuclease TnpB family protein [Saccharolobus]	Saccharolobus	615	615	100%	0.0	73.96%	384	WP_240781180.1
RNA-guided endonuclease TnpB family protein [Sulfuracidifex metallicus]	Sulfuracidifex metallicus	614	614	100%	0.0	76.04%	383	WP_156016515.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	614	614	100%	0.0	77.08%	382	WP_152943038.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	614	614	100%	0.0	76.50%	381	WP_152943392.1
RNA-guided endonuclease TnpB family protein [Sulfuracidifex metallicus]	Sulfuracidifex metallicus	613	613	100%	0.0	75.72%	382	WP_156016659.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	613	613	100%	0.0	75.98%	381	WP_012717770.1
RNA-guided endonuclease TnpB family protein [Sulfurisphaera tokodaii]	Sulfurisphaera tokodaii	613	613	100%	0.0	76.76%	381	WP_010978089.1
RNA-guided endonuclease TnpB family protein [Stygiolobus sp. CP850M]	Stygiolobus sp. CP850M	612	612	100%	0.0	75.52%	383	WP_337720655.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	611	611	100%	0.0	75.78%	382	WP_152942670.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	611	611	100%	0.0	75.98%	381	WP_013774965.1
transposase, IS605 OrfB family [Sulfolobus islandicus M.16.4]	Sulfolobus islandicus M.16.4	610	610	100%	0.0	75.20%	402	ACR41646.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	610	610	100%	0.0	76.50%	385	WP_014512221.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	610	610	100%	0.0	75.20%	386	WP_052885451.1
RNA-guided endonuclease TnpB family protein [Sulfurisphaera tokodaii]	Sulfurisphaera tokodaii	609	609	100%	0.0	76.24%	385	WP_010979004.1
RNA-guided endonuclease TnpB family protein [Saccharolobus caldissim...	Saccharolobus caldissimus	609	609	100%	0.0	76.24%	385	WP_229570262.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	608	608	100%	0.0	74.41%	382	WP_013775032.1

D. ATV TnpB GP68

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
transposase [Bicaudavirus pozzuoliense] **ATV TnpB GP68**	Bicaudavirus pozzuoliense	795	795	100%	0.0	100.00%	383	YP_319899.1
IS605 family OrfB transposase [Acidianus two-tailed virus 2]	Acidianus two-tailed virus 2	704	704	100%	0.0	87.21%	383	AON96456.1
RNA-guided endonuclease TnpB family protein [Saccharolobus solfataric...	Saccharolobus solfataricus	704	704	100%	0.0	87.21%	383	WP_020936635.1
RNA-guided endonuclease TnpB family protein [Acidianus infernus]	Acidianus infernus	701	701	100%	0.0	85.64%	383	WP_155864291.1
transposase [Bicaudavirus pozzuoliense]	Bicaudavirus pozzuoliense	701	701	100%	0.0	86.16%	383	YP_319874.1
RNA-guided endonuclease TnpB family protein [Acidianus infernus]	Acidianus infernus	697	697	100%	0.0	85.64%	383	WP_155863494.1
RNA-guided endonuclease TnpB family protein [Acidianus infernus]	Acidianus infernus	693	693	100%	0.0	87.47%	383	WP_155864207.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	649	649	100%	0.0	79.63%	382	WP_152939166.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	649	649	100%	0.0	79.90%	382	WP_152941266.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	645	645	100%	0.0	79.43%	383	WP_152942809.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	642	642	100%	0.0	78.33%	382	WP_152939397.1
RNA-guided endonuclease TnpB family protein [Sulfuracidifex metallicus]	Sulfuracidifex metallicus	642	642	100%	0.0	78.85%	382	WP_156016674.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	642	642	100%	0.0	78.91%	383	WP_013775100.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	639	639	100%	0.0	80.42%	381	WP_013776664.1
RNA-guided endonuclease TnpB family protein [Sulfuracidifex metallicus]	Sulfuracidifex metallicus	639	639	100%	0.0	79.11%	382	WP_054838422.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	636	636	100%	0.0	78.59%	381	WP_048052744.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	636	636	100%	0.0	77.55%	382	WP_013774993.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	636	636	100%	0.0	78.59%	381	WP_152943392.1
RNA-guided endonuclease TnpB family protein [Sulfuracidifex metallicus]	Sulfuracidifex metallicus	635	635	100%	0.0	77.86%	383	WP_156016515.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	634	634	100%	0.0	78.07%	382	WP_013775197.1
RNA-guided endonuclease TnpB family protein [Sulfolobales archaeon]	Sulfolobales archaeon	634	634	100%	0.0	77.60%	382	MCG2872300.1
RNA-guided endonuclease TnpB family protein [Sulfolobales archaeon]	Sulfolobales archaeon	634	634	100%	0.0	78.91%	387	MCQ4335705.1
RNA-guided endonuclease TnpB family protein [Stygiolobus sp. CP850M]	Stygiolobus sp. CP850M	634	634	100%	0.0	78.39%	383	WP_337720655.1
RNA-guided endonuclease TnpB family protein [Saccharolobus solfataric...	Saccharolobus solfataricus	632	632	100%	0.0	78.33%	381	WP_009992247.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	632	632	100%	0.0	77.02%	382	WP_152943027.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	632	632	100%	0.0	78.39%	382	WP_152943038.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	632	632	100%	0.0	77.28%	382	WP_013776628.1
transposase [Candidatus Marsarchaeota G1 archaeon OSP_C]	Candidatus Marsarchaeota G1 arc...	632	632	100%	0.0	77.55%	381	PSN87618.1
RNA-guided endonuclease TnpB family protein [Sulfuracidifex metallicus]	Sulfuracidifex metallicus	629	629	100%	0.0	77.28%	382	WP_156016659.1
RNA-guided endonuclease TnpB family protein [Acidianus ambivalens]	Acidianus ambivalens	629	629	100%	0.0	78.12%	382	WP_152942670.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	629	629	100%	0.0	75.52%	384	WP_218259439.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	628	628	100%	0.0	75.98%	382	WP_013775032.1
RNA-guided endonuclease TnpB family protein [Saccharolobus solfataric...	Saccharolobus solfataricus	627	627	100%	0.0	77.28%	382	WP_010923706.1
RNA-guided endonuclease TnpB family protein [Saccharolobus sp.]	Saccharolobus sp.	627	627	100%	0.0	75.78%	384	WP_342782880.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	627	627	100%	0.0	75.52%	384	WP_218267400.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	626	626	100%	0.0	75.52%	384	WP_218260425.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	626	626	100%	0.0	75.52%	384	WP_218261170.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	625	625	100%	0.0	77.81%	385	WP_014512221.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	625	625	100%	0.0	77.86%	383	WP_013776696.1
RNA-guided endonuclease TnpB family protein [Sulfuracidifex metallicus]	Sulfuracidifex metallicus	625	625	100%	0.0	77.81%	381	WP_156016645.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	625	625	100%	0.0	77.28%	381	WP_012717770.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	624	624	100%	0.0	77.02%	385	WP_015581754.1
RNA-guided endonuclease TnpB family protein [Sulfuracidifex metallicus]	Sulfuracidifex metallicus	624	624	100%	0.0	75.72%	382	WP_156016140.1
RNA-guided endonuclease TnpB family protein [Saccharolobus]	Saccharolobus	624	624	100%	0.0	75.26%	384	WP_240781180.1
RNA-guided endonuclease TnpB family protein [Acidianus hospitalis]	Acidianus hospitalis	624	624	100%	0.0	77.08%	383	WP_013776434.1
RNA-guided endonuclease TnpB family protein [Sulfurisphaera tokodaii]	Sulfurisphaera tokodaii	623	623	100%	0.0	77.28%	385	WP_010979004.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	623	623	100%	0.0	77.55%	385	WP_014512341.1
Transposase [Saccharolobus shibatae]	Saccharolobus shibatae	623	623	100%	0.0	78.33%	405	QXJ31318.1
RNA-guided endonuclease TnpB family protein [Sulfolobus islandicus]	Sulfolobus islandicus	623	623	100%	0.0	76.76%	385	WP_015581753.1
RNA-guided endonuclease TnpB family protein [Saccharolobus shibatae]	Saccharolobus shibatae	623	623	100%	0.0	78.33%	380	WP_218261281.1

2. NCBI BLAST search results of all genes upstream of ATV TnpB genes

A. ATV gp09

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
protease [Bicaudavirus pozzuoliense] **ATV GP09**	Bicaudavirus pozzuoliense	431	431	100%	3e-152	100.00%	209	YP_319840.1
putative metal-dependent protease [Acidianus two-tailed virus 2]	Acidianus two-tailed virus 2	386	386	99%	1e-134	89.95%	210	AON96421.1
hypothetical protein QIT30_gp34 [Saccharolobus solfataricus rod-shaped virus 1]	Saccharolobus solfataricus...	160	160	98%	1e-44	40.79%	280	YP_010771837.1
protease [Acidianus rod-shaped virus 2]	Acidianus rod-shaped virus 2	158	158	98%	9e-44	40.43%	280	YP_009230247.1
hypothetical protein [Sulfuracidifex metallicus]	Sulfuracidifex metallicus	151	151	99%	7e-41	37.07%	280	WP_277023106.1

B. ATV gp39

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
hypothetical protein ATV_gp39 [Bicaudavirus pozzuoliense] **ATV GP39**	Bicaudavirus pozzuoliense	265	265	100%	9e-89	100.00%	137	YP_319870.1
hypothetical protein [Acidianus infernus]	Acidianus infernus	234	234	95%	2e-76	90.84%	150	WP_155862300.1
hypothetical protein [Acidianus two-tailed virus 2]	Acidianus two-tailed virus 2	227	227	99%	9e-74	86.03%	150	AON96452.1
hypothetical protein [Saccharolobus solfataricus]	Saccharolobus solfataricus	216	216	98%	2e-69	80.74%	134	WP_010923099.1

C. ATV gp42

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
major head protein [Bicaudavirus pozzuoliense] **ATV GP42**	Bicaudavirus pozzuoliense	293	293	100%	2e-99	100.00%	145	YP_319873.1
Chain Z, RNAP inhibitory protein [Bicaudavirus pozzuoliense]	Bicaudavirus pozzuoliense	292	292	100%	3e-99	100.00%	149	7OQ4_Z
archaeal structural protein [Acidianus two-tailed phage variant 1]	Acidianus two-tailed phage variant 1	286	286	97%	4e-97	100.00%	142	AON96520.1
archaeal structural protein [Acidianus two-tailed virus 2]	Acidianus two-tailed virus 2	260	260	97%	8e-87	90.78%	142	AON96455.1
hypothetical protein [Acidianus infernus]	Acidianus infernus	255	255	100%	8e-85	92.41%	145	MCY0882427.1
major head protein [Acidianus tailed spindle virus]	Acidianus tailed spindle virus	72.8	72.8	62%	5e-13	43.96%	116	YP_009230323.1
major head protein [Sulfolobus monocaudavirus SMV2]	Sulfolobus monocaudavirus SMV2	57.4	57.4	57%	4e-07	38.82%	122	YP_009219220.1
major head protein [Sulfolobus monocaudavirus SMV1]	Sulfolobus monocaudavirus SMV1	53.1	53.1	57%	2e-05	36.47%	122	YP_009008073.1
major head protein [Sulfolobus monocaudavirus SMV4]	Sulfolobus monocaudavirus SMV4	49.3	49.3	57%	5e-04	35.29%	122	YP_009218469.1
major head protein [Sulfolobus monocaudavirus SMV3]	Sulfolobus monocaudavirus SMV3	44.3	44.3	31%	0.011	46.67%	58	YP_009226229.1

D. ATV gp67

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
hypothetical protein ATV_gp67 [Bicaudavirus pozzuoliense]	Bicaudaviru...	1080	1080	100%	0.0	100.00%	545	YP_319898.1
membrane protein [Acidianus two-tailed virus 2]	Acidianus tw...	956	956	99%	0.0	88.29%	558	AON96479.1
hypothetical protein [Acidianus infernus]	Acidianus in...	953	953	92%	0.0	95.26%	507	MCY0882336.1
hypothetical protein IX51_09055 [uncultured archaeon]	uncultured a...	105	105	52%	2e-19	31.23%	809	AKA49227.1
hypothetical protein [Candidatus Thermoplasmatota archaeon]	Candidatus ...	103	103	76%	1e-18	26.96%	617	MCL4336403.1
hypothetical protein [Candidatus Thermoplasmatota archaeon]	Candidatus ...	98.6	98.6	75%	4e-17	26.95%	786	MCL4329958.1
hypothetical protein AMDU5_GPLC00012G0016 [Thermoplasmatales archaeon Gpl]	Thermoplas...	94.4	94.4	78%	7e-16	27.94%	758	EQB68376.1
multipass membrane protein [Cuniculiplasma divulgatum]	Cuniculiplas...	94.4	94.4	78%	8e-16	27.94%	758	SIM64454.1
hypothetical protein [Candidatus Thermoplasmatota archaeon]	Candidatus ...	94.0	94.0	78%	9e-16	27.94%	758	MCL4320177.1
hypothetical protein [Candidatus Thermoplasmatota archaeon]	Candidatus ...	94.0	94.0	78%	9e-16	27.94%	758	MCL5787396.1
hypothetical protein RE472_09210 [Thermoplasmatales archaeon]	Thermoplas...	94.0	94.0	78%	1e-15	27.94%	758	WMT49227.1
hypothetical protein [Conexivisphaerales archaeon]	Conexivisph...	93.2	93.2	70%	2e-15	24.43%	859	MDP7982292.1
hypothetical protein [Candidatus Thermoplasmatota archaeon]	Candidatus ...	92.0	92.0	66%	3e-15	29.37%	567	MCL4350827.1
hypothetical protein [Thermogymnomonas acidicola]	Thermogym...	92.4	92.4	55%	3e-15	27.00%	690	WP_188681796.1
hypothetical protein [Nitrososphaerota archaeon]	Nitrosospha...	90.9	90.9	74%	1e-14	24.14%	807	MDG6927815.1
glycosyltransferase 87 family protein [Cuniculiplasma divulgatum]	Cuniculiplas...	90.1	90.1	70%	2e-14	27.35%	682	WP_145983962.1
glycosyltransferase 87 family protein [Cuniculiplasma divulgatum]	Cuniculiplas...	90.1	90.1	70%	2e-14	27.35%	682	WP_148689816.1
hypothetical protein [Candidatus Sysuiplasma acidicola]	Candidatus ...	89.7	89.7	48%	3e-14	31.16%	766	MBX8634857.1
hypothetical protein [Candidatus Sysuiplasma acidicola]	Candidatus ...	89.7	89.7	48%	3e-14	30.82%	766	MBX8636765.1
hypothetical protein [Candidatus Thermoplasmatota archaeon]	Candidatus ...	89.4	89.4	48%	3e-14	28.47%	822	MCL4327002.1
hypothetical protein [Candidatus Sysuiplasma superficiale]	Candidatus ...	89.4	89.4	58%	3e-14	26.93%	837	MBX8643526.1
hypothetical protein [Candidatus Thermoplasmatota archaeon]	Candidatus ...	89.4	89.4	58%	4e-14	26.93%	837	MCL5437488.1
hypothetical protein [Candidatus Sysuiplasma superficiale]	Candidatus ...	89.0	89.0	58%	4e-14	26.93%	747	MBX8632186.1
hypothetical protein [Picrophilus oshimae]	Picrophilus ...	88.2	88.2	48%	7e-14	30.95%	711	WP_153274215.1
Uncharacterized membrane protein [Picrophilus oshimae DSM 9789]	Picrophilus ...	88.2	88.2	48%	8e-14	30.95%	716	SMD30956.1

transporter [Picrophilus oshimae DSM 9789]	Picrophilus…	87.8	87.8	48%	1e-13	30.80%	716	AAT43972.1
hypothetical protein [Picrophilus oshimae]	Picrophilus…	87.4	87.4	48%	1e-13	30.80%	711	WP_153274147.1
hypothetical protein [Candidatus Thermoplasmatota archaeon]	Candidatus…	87.4	87.4	48%	1e-13	27.08%	782	MCL4343076.1
hypothetical protein [Candidatus Thermoplasmatota archaeon]	Candidatus…	87.0	87.0	74%	2e-13	26.87%	624	MCL6089718.1
hypothetical protein [Candidatus Thermoplasmatota archaeon]	Candidatus…	86.7	86.7	74%	2e-13	26.87%	579	MCL4336051.1
glycosyltransferase 87 family protein [Candidatus Thermoplasmatota archaeon]	Candidatus…	86.7	86.7	51%	2e-13	29.24%	595	MCL5679209.1
hypothetical protein [Candidatus Thermoplasmatota archaeon]	Candidatus…	86.3	86.3	78%	3e-13	24.79%	669	MCL5802823.1
hypothetical protein [Candidatus Thermoplasmatota archaeon]	Candidatus…	85.1	85.1	48%	7e-13	28.47%	783	MCL5786333.1
hypothetical protein AMDU2_EPLC00005G0007 [Thermoplasmatales archaeon E-plas…	Thermoplas…	84.7	84.7	48%	8e-13	28.82%	762	EQB67171.1
hypothetical protein [Candidatus Thermoplasmatota archaeon]	Candidatus…	84.7	84.7	48%	9e-13	28.18%	783	MCL4328410.1
TPA: hypothetical protein [Acidimicrobiia bacterium]	Acidimicrobii…	83.2	83.2	64%	2e-12	24.31%	587	HET8740769.1
hypothetical protein B2I18_03350 [Cuniculiplasma sp. C_DKE]	Cuniculiplas…	82.8	82.8	84%	4e-12	24.95%	758	OWP54329.1
hypothetical protein [Candidatus Thermoplasmatota archaeon]	Candidatus…	82.0	82.0	70%	6e-12	26.01%	749	MCL6014763.1
glycosyltransferase 87 family protein [Candidatus Thermoplasmatota archaeon]	Candidatus…	80.1	80.1	73%	2e-11	26.77%	587	MCL5441514.1
hypothetical protein [Ferroplasma sp.]	Ferroplasma…	80.1	80.1	64%	2e-11	25.86%	616	WP_337860763.1
hypothetical protein [Chloroflexota bacterium]	Chloroflexot…	78.2	78.2	53%	6e-11	26.56%	451	MDE3094651.1
glycosyltransferase family 87 protein [Thermoplasma sp. Kam2015]	Thermoplas…	78.2	78.2	48%	1e-10	27.02%	689	WP_110641552.1
putative integral membrane protein [Thermoplasmatales archaeon]	Thermoplas…	77.8	77.8	60%	1e-10	27.92%	780	QRF75126.1
hypothetical protein [Candidatus Thermoplasmatota archaeon]	Candidatus…	77.0	77.0	74%	2e-10	24.41%	619	MCL5881974.1
hypothetical protein DRJ98_04125 [Thermoprotei archaeon]	Thermoprot…	76.3	76.3	27%	2e-10	31.33%	410	RLF11222.1
glycosyltransferase 87 family protein [Candidatus Thermoplasmatota archaeon]	Candidatus…	76.6	76.6	72%	2e-10	25.27%	624	MCL5782963.1
hypothetical protein B2I17_05885 [Thermoplasmatales archaeon B_DKE]	Thermoplas…	77.0	77.0	60%	2e-10	27.92%	782	OWP56585.1
TPA: hypothetical protein [Gaiellaceae bacterium]	Gaiellaceae …	77.0	77.0	51%	3e-10	27.19%	699	HEX4518784.1
hypothetical protein DRN06_01800 [Thermoprotei archaeon]	Thermoprot…	75.5	75.5	27%	3e-10	30.67%	410	RLF18334.1
hypothetical protein HS7_04510 [Sulfolobales archaeon HS-7]	Sulfolobales…	76.3	76.3	65%	4e-10	24.74%	602	BCU67014.1

APPENDIX C

Plasmid Sequences and Map

A. pET42b(+) cloned sequences:

TwinStrep-10×His-MBP-TEV (cleavage site)-TnpBGP10

1. GP10

ATGGGGGGTTCAGCTTGGTCGCACCCGCAGTTTGAAAAGGGTGGAGGTTCGGGC
GGTGGGAGCGGCGGCAGTGCGTGGTCGCATCCTCAGTTTGAAAAGGGTTCGATG
GGGGGTTCTCATCATCATCATCATCATCATCATCATCATGGTATGGCTAGCATG
AAAATCGAAGAAGGTAAACTGGTAATCTGGATTAACGGCGATAAAGGCTATAAC
GGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACC
GTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGC
GATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAA
TCTGGCCTGTTGGCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTAT
CCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATC
GCTGTTGAAGCGTTATCGCTGATTTATAACAAAGATCTGCTGCCGAACCCGCCA
AAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAG
AGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCT
GCTGACGGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGAC
GTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTG
ATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCC
TTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAAC
ATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGT
CAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGT
CCGAACAAAGAGCTGGCAAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAA
GGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCT
TACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACTATGGAAAACGCC
CAGAAAGGTGAAATCATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCC
GTGCGTACTGCGGTGATCAACGCCGCCAGCGGTCGTCAGACTGTCGATGAAGCC
CTGAAAGACGCGCAGACTAATTCGAGCTCGAACAACAACAACAATAACAATAAC
AACAACCTCGGGATCGAGGAAAACCTGTACTTCCAATCCAATGCAGGTGGTGGT
GGTATGCCCGACGTAGGGATTAGGTTTAGGGCTTTCGCAGATGAACCTACAATT
AGGGAGTTAAAATCCCAGTTGAGGTTAGCGTGCGAAATTTACAACACCCTACGT
TGGGCAGATATATATTTCTACCAAAGAGATGGGAAAGGTCTAAGCAAGACTGAA
TTGAGACAACTCGCTCTAGATTTGAGAAAACAGGACGATGAATACAAGAGAATC
TATTCTCAAGCAGTACAACAAATTGCAGACAGATTCTACGACGCTAAGAAGAGA
TTCCTCAAAGGGTTAGCACGCTACCCCAGGGAAAAGAAACCACATAAATGGTAT
TCGTTAGTGTACCCTCAATCAGGTTGGAAAGTGTTAGAGACGAGAGAAATAAGA
ACTAAAAGCAAGAAGAACAAGAAGAAGATAATGACGCTTCAGTTAGCAAACCTG
GGAGTCTTCAAAGTTATTGTTCATAGGGACTTTCCGCTAGACAAGGTAAAGAGG
GTGATAGTTAAGTTAACACCTTCAGAGAGAGTGTACATTAGTTTCGTTGTGGAT
CAAGAATACCCTCAACTCCCTAAGACAAACAAAACGGTTGGAATAGATGTTGGA
ATAGAGAAACTTCTCATTACATCAGATGGGGAATACGTGCCAAACTTGCGACCT
TACGAGAAGGCACTCAATAAGACAAGAAAACTTCATAAAGCTCTCTCGAGGAAG
AAGTTCTTGTCGCACAACTGGTTTAAGGCAAAGATTAAGCTCGCAAGGGCTTAT

GAGCATTTGAAGAATCTTAGGAGGGACATATACATGAAACTTGGGAAGTATTTT
GCAGAGCATTATGACGTAGTCGTAATGGAAGATATTCAAGTTAAACAACTTGTT
GGTAAATCCTACAAGAAAATGAGGATGAGGCTTCACGATGTTGCATTTTACGAG
CTTAGGAGCATCATGGAATATCAACTTAGGAAGTACGGGAAGGAGCTTCGTCTC
GTTGATCCTGCTTTTACTTCGATGACTTGTGCTAAATGCGGGTATGTTAAGAAG
GATTTAACTTTGGCTGATCGTATTTTTGTTTGTCCAAAATGTGGTTGGACTGTG
GATCGTGACTACAATGCTTCTCTTAACATTCTAAAAAGGTCGGGGTGGGAGCCA
CCCTTAGCGCCTGTGGAGCTCCGCCCTCTACCATTGGTAAAAAGCCAATGGCAA
GGTGGGGCTTTGAAGCAGGAAGCCTCGCCCTTTAGGGTGGGGTAG

2. GP40

ATGGGGGGTTCAGCTTGGTCGCACCCGCAGTTTGAAAAGGGTGGAGGTTCGGGC
GGTGGGAGCGGCGGCAGTGCGTGGTCGCATCCTCAGTTTGAAAAGGGTTCGATG
GGGGGTTCTCATCATCATCATCATCATCATCATCATCATGGTATGGCTAGCATG
AAAATCGAAGAAGGTAAACTGGTAATCTGGATTAACGGCGATAAAGGCTATAAC
GGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACC
GTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGC
GATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAA
TCTGGCCTGTTGGCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTAT
CCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATC
GCTGTTGAAGCGTTATCGCTGATTTATAACAAAGATCTGCTGCCGAACCCGCCA
AAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAG
AGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCT
GCTGACGGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGAC
GTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTG
ATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCC
TTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAAC
ATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGT
CAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGT
CCGAACAAAGAGCTGGCAAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAA
GGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCT
TACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACTATGGAAAACGCC
CAGAAAGGTGAAATCATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCC
GTGCGTACTGCGGTGATCAACGCCGCCAGCGGTCGTCAGACTGTCGATGAAGCC
CTGAAAGACGCGCAGACTAATTCGAGCTCGAACAACAACAACAATAACAATAAC
AACAACCTCGGGATCGAGGAAAACCTGTACTTCCAATCCAATGCAGGTGGTGGT
GGTATGCCACCCTCCTCTGGTCAACTCTTAGGGGATGAGGAGCGGGAGCCGACT
TCTACTCCCGCAATACCGGAGGAGGGTGTCTACAAAGTTAAATACTCAAATCGA
AGAACCAATATAGTTCGCCTTCTTCCAAATGGTTTTCAAGAAAGAAAACTGAGG
AGGTTAGCAGACCTCTCTGCAAAGCTCTTCAACGAAGTTAACTATGAAAGGAGG
CAACAGTTCTTTCACGAAGGGAAAGTGGACATTAAGGGGACATATAAAAAATAT
TATGAGAAGTATAAGGAAAAATTACGTACTAACGCACAAGCTGTTCTCAATAAG
AACAACGAAGCGTGGTCGTCATTCTTCTCCTTGCTAAATCTGAAGAAGGAAGGA
AAATTACCACAACACATAAAACACGTTTCTCCACCCGGGTACTGTAAGGACAGA
AAGACGAAGAAGAGAAAGCTAATACTAATCGTTAGACAGGATCGTTACAAGGTA
GATGCTGAAAACAACAAATTAATCCTCAAGGACTTTAACATGGAAATAGAATTC
GTCGGTAGACTTAGGTGGTATGGTAAACAAGGTAGGCTAGAGATAATTTTTGAC
GAGACTAGAAATGCTTGGTACGCTCACATACCAGTTGAGGTAGGTGTTGAGGAA
ACTGGAAAGAAGAGTAAGCACGTCGTCAAGGGTGAGAGAAAGAGTATCCAGATT
GCCAAACCAAAGGGAAACAAAGTAGCATCCATCGATTTAGGTATTAACGTAATA

GCAAGTGTAGTGGTAAGTGATGGCACTTGGTTGCTCTACAAGGGTATTAGAACA
AAGGAGGACTACTTCTACTTCCATAAGAGAATTGCAGAAGTACAATCGTTAGCT
GACAGAACTAGAAACATTGGTGAATATGAGGCATATCTAGAATTGTTAAGGGAA
GAGAGGAGACTGTTTAAGAAACTGAAGAGACGCCTTCTCCATCTCTACCGTAAC
TTAGCCTCTCACCTTATTAAAACACTCCACGAACTAGGAGTATCAACAATTTAC
CTAGGTAACCCATTCAATATCGTGCAGGAGAAAGGCGACAACTTCATGACAGAC
AAGTGGCCTTATCGTAAGTTGATGCATGCCATTGAATTAAAGGCTCAAGAATAT
GGTATGAAGGTGTATGAAGTTGATGAGTACAATACAACAAAGTATTGTGCTTAC
CACGACGTGAAAGTAAAGAGGAACCCAAGAGGAGTAGTTATTTGTCCTAAGGGA
CATAAACTTCACAGCGATCTAAACGGTGCGTTAAACATTCTAAAGAAAGCAGTA
GGAGTGGTCGTTAACGAAGTGAAAAAACCCTTGTCTTTCATTGTGGATCATAAC
AGAGTAGCACCCATCAAAGGGGTGTAA

3. GP43

ATGGGGGGTTCAGCTTGGTCGCACCCGCAGTTTGAAAAGGGTGGAGGTTCGGGC
GGTGGGAGCGGCGGCAGTGCGTGGTCGCATCCTCAGTTTGAAAAGGGTTCGATG
GGGGGTTCTCATCATCATCATCATCATCATCATCATCATGGTATGGCTAGCATG
AAAATCGAAGAAGGTAAACTGGTAATCTGGATTAACGGCGATAAAGGCTATAAC
GGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACC
GTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGC
GATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAA
TCTGGCCTGTTGGCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTAT
CCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATC
GCTGTTGAAGCGTTATCGCTGATTTATAACAAAGATCTGCTGCCGAACCCGCCA
AAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAG
AGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCT
GCTGACGGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGAC
GTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTG
ATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCC
TTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAAC
ATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGT
CAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGT
CCGAACAAAGAGCTGGCAAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAA
GGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCT
TACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACTATGGAAAACGCC
CAGAAAGGTGAAATC**ATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCC**
GTGCGTACTGCGGTGATCAACGCCGCCAGCGGTCGTCAGACTGTCGATGAAGCC
CTGAAAGACGCGCAGACTAATTCGAGCTCGAACAACAACAACAATAACAATAAC
AACAACCTCGGGATCGAGGAAAACCTGTACTTCCAATCCAATGCAGGTGGTGGT
GGTATGGCTAGGAGGGAGAAAAACCCAATCAGAGCAACGGTTTCGATGAAGATT
GGCTTATCTGACTCCCTCCTAGCCTTCGTGAACAACTACGTTAGAGCACTCCGT
TTCTCCATATTTTGGATGAAAGAGAATGTGAAAAATCCAAACGAGAAGGGCACG
CTTTCTAAAGTGCACGAGGGATTATATGGAAAGCTAAAGGAGGATTATAATCTA
CCACCCAAAGTTTCTGCGGACTGTTATCGTGATGCCCTGGCAATATATAAGAGT
TGGTATAACAATCCCAAGAGGGGTAGATTCCCCCGCGTCTATAAGCCCACTGTA
TGGTTAACGCCCAAGCGAAGTTATACTGTAGACTTAGATAGAATGGTAGTTAAG
ATTACCAGTGTTGGGAACTGCCAATTTTAGGCTATCCTAGGAACTTAAAAGAGT
ATGCAAACTGGGATATGAAGGAGGCTAGGCTAACAATCAGAGATGGCAAAGCTT
TCCTCAAGGTGGTTTTTGAGAAACCGAAAGTTAAGATACAACCCAAAGGTAGTG

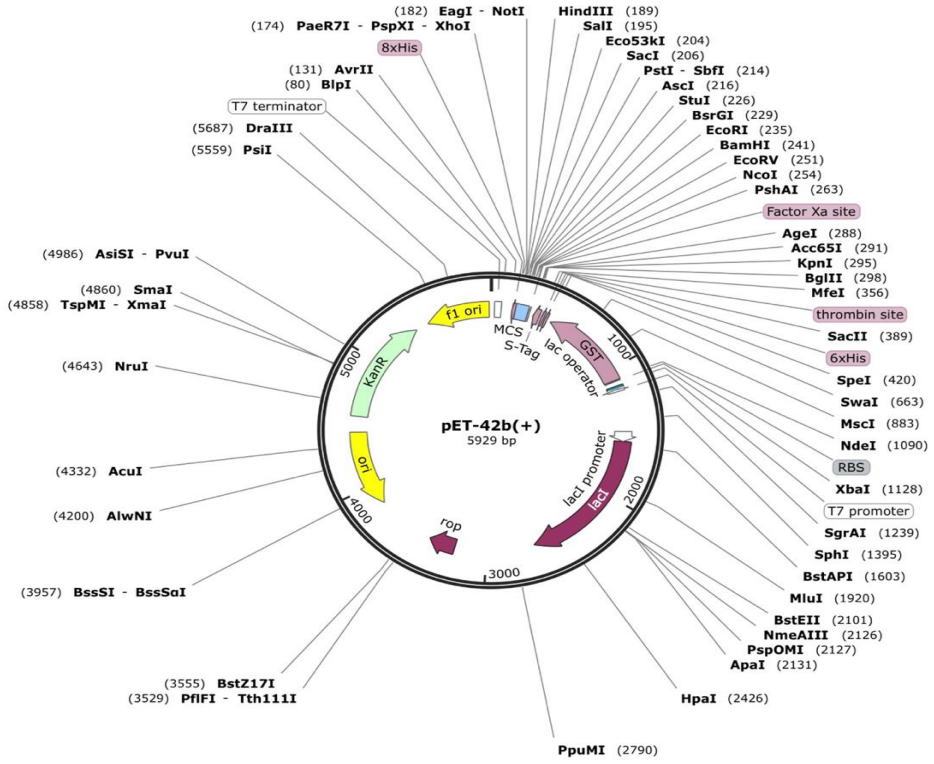
TTGCCGTTGATATTAACATGAGTGAAATTGTAGTAGGGAAGGACGACAGTCACT
ACGTTAGGATTCCCACTCGCCTTCACGAGTCTCACCACTTCAAGACATTAGCTG
AGAATTTGCAAAAGAAGTATCCTAGAAGGTGGAAGGAGAACAAGAGAATTCTAC
ACAGAATACGCTCTTTTCATCACAAGGCCAAACTAATTATGGAGGACTTCGCTA
GGAAAGTTGGCAAGTGGGTTGTTGAGATTGCTTGGGATTTGGGTGCCAACGTAA
TCAAATTGGAGAATCTTAAGAACCTCATCAGGAACGTCAACAAACTGTCAGCCG
AGTTTCGCGATAAACTCTACTTGATGCAATATCATCGTATTCAGTACTGGATAG
AATGGCAAGCCAGAAAACACGGAATGAATGTGGAGTATGTTAATCCTAGTTACT
CTTCCGTCTCTTGTCCAAAGTGTGGCCGCAAAATGGTTGAGATTGCTCATAGGT
ACTTCCACTGTCCCTCGTGTGGTTATGAGAACGATCGTGACGTTATTGCTATCA
TGAATTTAAATGGGAGGGGGTCTCTGACCCTCTCGACTGCCCCTCAGATGAGAG
ATGTAATCCCGAATCGATGA

4. GP68

ATGGGGGGTTCAGCTTGGTCGCACCCGCAGTTTGAAAAGGGTGGAGGTTCGGGC
GGTGGGAGCGGCGGCAGTGCGTGGTCGCATCCTCAGTTTGAAAAGGGTTCGATG
GGGGGTTCTCATCATCATCATCATCATCATCATCATCATGGTATGGCTAGCATG
AAAATCGAAGAAGGTAAACTGGTAATCTGGATTAACGGCGATAAAGGCTATAAC
GGTCTCGCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACC
GTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGC
GATGGCCCTGACATTATCTTCTGGGCACACGACCGCTTTGGTGGCTACGCTCAA
TCTGGCCTGTTGGCTGAAATCACCCCGGACAAAGCGTTCCAGGACAAGCTGTAT
CCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATC
GCTGTTGAAGCGTTATCGCTGATTTATAACAAAGATCTGCTGCCGAACCCGCCA
AAAACCTGGGAAGAGATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAG
AGCGCGCTGATGTTCAACCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCT
GCTGACGGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATTAAAGAC
GTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTGACCTTCCTGGTTGACCTG
ATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAGAAGCTGCC
TTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAAC
ATCGACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGT
CAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGT
CCGAACAAAGAGCTGGCAAAAGAGTTCCTCGAAAACTATCTGCTGACTGATGAA
GGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCT
TACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACTATGGAAAACGCC
CAGAAAGGTGAAATCATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCC
GTGCGTACTGCGGTGATCAACGCCGCCAGCGGTCGTCAGACTGTCGATGAAGCC
CTGAAAGACGCGCAGACTAATTCGAGCTCGAACAACAACAACAATAACAATAAC
AACAACCTCGGGATCGAGGAAAACCTGTACTTCCAATCCAATGCAGGTGGTGGT
GGTATGGCTAGGAGGGAGAAAAACCCAATAAGAGCCACAGTTGCGATGAATATT
GGCTTATCTGACTCCCTCCTAGCCTTCGTGAACAACTACGTTAAGGCACTCCGT
TTCTCCTTGTTCTGGATGAAAGAGAATGTGAAAAATCCAAACGAGAAGGGCGCA
CTCTCCAAAGTACACGAGGGATTGTATGAAAAGCTGAGGAAAGAATACAATCTA
CCATCTAAAGTTACTGAGGACTGTTATCGTGATGCCCTGGCAATATACAAGAGT
TGGTACAACAACCCGAAAAAAGGTAGATTCCCCCGTGTCTACAAGCCGACAGTG
TGGCTAACGCCCAAGCAAAGTTATACTGTAGACTTAGAGAAAATGACAGTCAGG
ATAGCAAGTGTTGGCGAACTACCAATACTAGGTTATCCTAGAAACCTAAAGGAG
TATGCAAACTGGAAGATGAAGGAGGCTAGGCTAACAATCAAGGATGGCAAGGCT
CTCCTCAAAGTAACTTTTGAGAAGGAAGAAGAGAAAGTTAAACCAAAAGACAGT

GTTGCTGTTGATATAAACATGAATGACATTGTCGTTGGTAAGGATGACACTCAC
TACGTTAGGATTCCCACTCGCCTTCACGACGCTCACCACTTCAAGTCATTAGCT
GAGAATTTGCAGAAGAAGTATCCTAGAAGGTGGAAGCAAAATAGGAGAATTCTA
CACAGGGCACGCTCTTTTCATCAAAAGGCCAAACTAATTATGGAGGACTACGCT
AGGAAGGTTGGTAAGTGGGTTGTTGAGATTGCTGAGGGTTTGGGTGCCAACGTC
ATAAAGCTTGAGGACTTGAAGAACCTCATCAAGGACGTTAATAAGCTACCAGCT
GAATTTCGCGATAAACTATACTTGATGCAATATCGTCGTATTCAGTATTGGATA
GAGTGGCAGGCTAAGAAACACGGAATGATTGTGGAGTTTGTTAATCCTAGTTAC
TCTTCCGTTTCTTGCCCAAAGTGTGGCCACAAAATGGTTGAGATTGCTTATAGG
TACTTTCACTGTCCTTCATGTGGTTATGAGAACGATCGTGATGTTATTGCTATC
ATGAATTTAAATGGGAGGGGGTCTCTGACCCTCTCGACTGCCCCTCAAATGAGA
GATGTAGCTCCGAATCGATGA

B. pET42b(+) plasmid map



C. pET30a(+) cloned sequences:

5' intergenic region—ATG—10XHisTag—TnpB—3'intergenic region

1. GP10

CCCCCTTATAGATTTAAAAACTTCTATTCACAATATTCTATTTAATGCATCATC
ATCATCATCATCATCATCATCATATGCCCGACGTAGGGATTAGGTTTAGGGCTT
TCGCAGATGAACCTACAATTAGGGAGTTAAAATCCCAGTTGAGGTTAGCGTGCG
AAATTTACAACACCCTACGTTGGGCAGATATATATTTCTACCAAAGAGATGGGA
AAGGTCTAAGCAAGACTGAATTGAGACAACTCGCTCTAGATTTGAGAAAACAGG
ACGATGAATACAAGAGAATCTATTCTCAAGCAGTACAACAAATTGCAGACAGAT
TCTACGACGCTAAGAAGAGATTCCTCAAAGGGTTAGCACGCTACCCCAGGGAAA
AGAAACCACATAAATGGTATTCGTTAGTGTACCCTCAATCAGGTTGGAAAGTGT
TAGAGACGAGAGAAATAAGAACTAAAAGCAAGAAGAACAAGAAGAAGATAATGA
CGCTTCAGTTAGCAAACCTGGGAGTCTTCAAAGTTATTGTTCATAGGGACTTTC
CGCTAGACAAGGTAAAGAGGGTGATAGTTAAGTTAACACCTTCAGAGAGAGTGT
ACATTAGTTTCGTTGTGGATCAAGAATACCCTCAACTCCCTAAGACAAACAAAA
CGGTTGGAATAGATGTTGGAATAGAGAAACTTCTCATTACATCAGATGGGGAAT
ACGTGCCAAACTTGCGACCTTACGAGAAGGCACTCAATAAGACAAGAAAACTTC
ATAAAGCTCTCTCGAGGAAGAAGTTCTTGTCGCACAACTGGTTTAAGGCAAAGA
TTAAGCTCGCAAGGGCTTATGAGCATTTGAAGAATCTTAGGAGGGACATATACA
TGAAACTTGGGAAGTATTTTGCAGAGCATTATGACGTAGTCGTAATGGAAGATA
TTCAAGTTAAACAACTTGTTGGTAAATCCTACAAGAAAATGAGGATGAGGCTTC
ACGATGTTGCATTTTACGAGCTTAGGAGCATCATGGAATATCAACTTAGGAAGT
ACGGGAAGGAGCTTCGTCTCGTTGATCCTGCTTTTACTTCGATGACTTGTGCTA
AATGCGGGTATGTTAAGAAGGATTTAACTTTGGCTGATCGTATTTTTGTTTGTC
CAAAATGTGGTTGGACTGTGGATCGTGACTACAATGCTTCTCTTAACATTCTAA
AAAGGTCGGGGGTGGGAGCCACCCTTAGCGCCTGTGGAGCTCCGCCCTCTACCAT
TGGTAAAAAGCCAATGGCAAGGTGGGGCTTTGAAGCAGGAAGCCTCGCCCTTTA
GGGTGGGGTAGCTCACAAGGGCGAAGAGAGGAAAGTCGGTGAGGATACTCTA

2. GP40

GAACACAGCTGTTTTTAACGAAACTAACTTCTTCAAACTTTTATAAAACTGAGG
GTTAAAGTGGGCGGAAAGGGAATGGATAGTCCCTTATAGATTTTTAAACTTCAG
CAATAACGAGTAGTAATGAGCAGTTATGCATCATCATCATCATCATCATCATCA
TCATATGCCACCCTCCTCTGGTCAACTCTTAGGGGATGAGGAGCGGGAGCCGAC
TTCTACTCCCGCAATACCGGAGGAGGGTGTCTACAAAGTTAAATACTCAAATCG
AAGAACCAATATAGTTCGCCTTCTTCCAAATGGTTTTCAAGAAAGAAAACTGAG
GAGGTTAGCAGACCTCTCTGCAAAGCTCTTCAACGAAGTTAACTATGAAAGGAG
GCAACAGTTCTTTCACGAAGGGAAAGTGGACATTAAGGGGACATATAAAAAATA
TTATGAGAAGTATAAGGAAAAATTACGTACTAACGCACAAGCTGTTCTCAATAA
GAACAACGAAGCGTGGTCGTCATTCTTCTCCTTGCTAAATCTGAAGAAGGAAGG
AAAATTACCACAACACATAAAACACGTTTCTCCACCCGGGTACTGTAAGGACAG
AAAGACGAAGAAGAGAAAGCTAATACTAATCGTTAGACAGGATCGTTACAAGGT
AGATGCTGAAAACAACAAATTAATCCTCAAGGACTTTAACATGGAAATAGAATT
CGTCGGTAGACTTAGGTGGTATGGTAAACAAGGTAGGCTAGAGATAATTTTTGA
CGAGACTAGAAATGCTTGGTACGCTCACATACCAGTTGAGGTAGGTGTTGAGGA
AACTGGAAAGAAGAGTAAGCACGTCGTCAAGGGTGAGAGAAAGAGTATCCAGAT
TGCCAAACCAAAGGGAAACAAAGTAGCATCCATCGATTTAGGTATTAACGTAAT

AGCAAGTGTAGTGGTAAGTGATGGCACTTGGTTGCTCTACAAGGGTATTAGAAC
AAAGGAGGACTACTTCTACTTCCATAAGAGAATTGCAGAAGTACAATCGTTAGC
TGACAGAACTAGAAACATTGGTGAATATGAGGCATATCTAGAATTGTTAAGGGA
AGAGAGGAGACTGTTTAAGAAACTGAAGAGACGCCTTCTCCATCTCTACCGTAA
CTTAGCCTCTCACCTTATTAAAACACTCCACGAACTAGGAGTATCAACAATTTA
CCTAGGTAACCCATTCAATATCGTGCAGGAGAAAGGCGACAACTTCATGACAGA
CAAGTGGCCTTATCGTAAGTTGATGCATGCCATTGAATTAAAGGCTCAAGAATA
TGGTATGAAGGTGTATGAAGTTGATGAGTACAATACAACAAAGTATTGTGCTTA
CCACGACGTGAAAGTAAAGAGGAACCCAAGAGGAGTAGTTATTTGTCCTAAGGG
ACATAAACTTCACAGCGATCTAAACGGTGCGTTAAACATTCTAAAGAAAGCAGT
AGGAGTGGTCGTTAACGAAGTGAAAAAACCCTTGTCTTTCATTGTGGATCATAA
CAGAGTAGCACCCATCAAAGGGGTGTAACCCTTAAGACCTCGGGGAACCCTCGC
CCTTCAGGGCGGGGAGGAGGTCGGTACTACTAGGATGCCTTTGGCGCTAACTCT
TCTTTCTCACTTAAGACACAAGACTTAAGACAGGACTTAAGACAGCTGTTTTTA
ACGAT

3. GP43

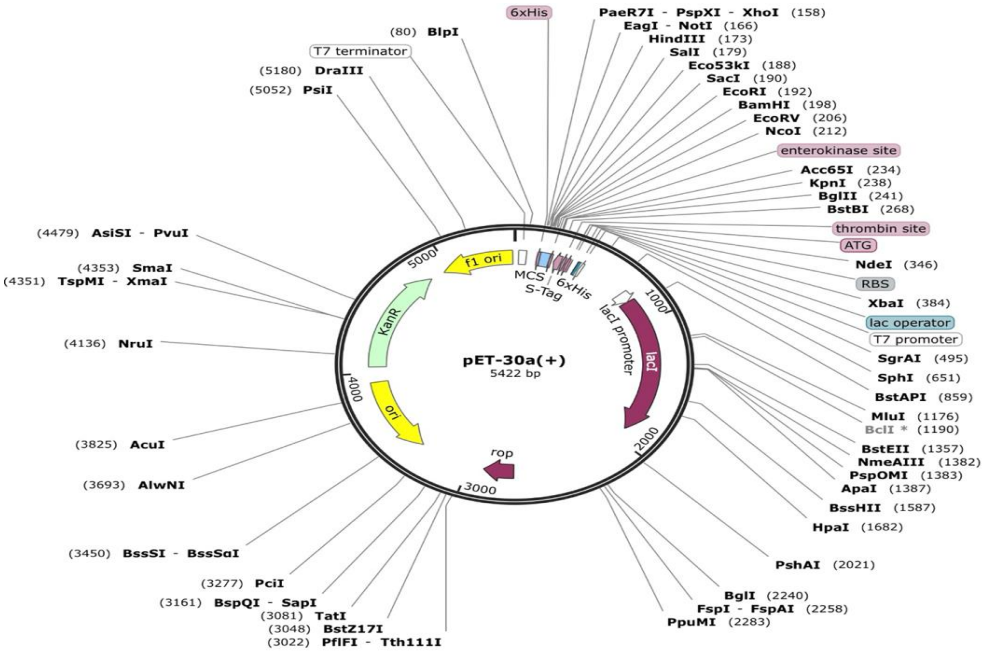
ACGAGAAGAGAATGCGTGAAAATCTTAAAAACAAAACTGGGGGTTAAGGGGCGG
AAAGCCTCGTAGGCGGGTATGGATAACCCCTTTATAGATTTAAAAACTTCTTTT
CCTATTGATACCCAATGCATCATCATCATCATCATCATCATCATCATATGGCTA
GGAGGGAGAAAAACCCAATCAGAGCAACGGTTTCGATGAAGATTGGCTTATCTG
ACTCCCTCCTAGCCTTCGTGAACAACTACGTTAGAGCACTCCGTTTCTCCATAT
TTTGGATGAAAGAGAATGTGAAAAATCCAAACGAGAAGGGCACGCTTTCTAAAG
TGCACGAGGGATTATATGGAAAGCTAAAGGAGGATTATAATCTACCACCCAAAG
TTTCTGCGGACTGTTATCGTGATGCCCTGGCAATATATAAGAGTTGGTATAACA
ATCCCAAGAGGGGTAGATTCCCCCGCGTCTATAAGCCCACTGTATGGTTAACGC
CCAAGCGAAGTTATACTGTAGACTTAGATAGAATGGTAGTTAAGATTACCAGTG
TTGGTGAACTGCCAATTTTAGGCTATCCTAGGAACTTAAAAGAGTATGCAAACT
GGGATATGAAGGAGGCTAGGCTAACAATCAGAGATGGCAAAGCTTTCCTCAAGG
TGGTTTTTGAGAAACCGAAAGTTAAGATACAACCCAAAGGTAGTGTTGCCGTTG
ATATTAACATGAGTGAAATTGTAGTAGGGAAGGACGACAGTCACTACGTTAGGA
TTCCCACTCGCCTTCACGAGTCTCACCACTTCAAGACATTAGCTGAGAATTTGC
AAAAGAAGTATCCTAGAAGGTGGAAGGAGAACAAGAGAATTCTACACAGAATAC
GCTCTTTTCATCACAAGGCCAAACTAATTATGGAGGACTTCGCTAGGAAAGTTG
GCAAGTGGGTTGTTGAGATTGCTTGGGATTTGGGTGCCAACGTAATCAAATTGG
AGAATCTTAAGAACCTCATCAGGAACGTCAACAAACTGTCAGCCGAGTTTCGCG
ATAAACTCTACTTGATGCAATATCATCGTATTCAGTACTGGATAGAATGGCAAG
CCAGAAAACACGGAATGAATGTGGAGTATGTTAATCCTAGTTACTCTTCCGTCT
CTTGTCCAAAGTGTGGCCGCAAAATGGTTGAGATTGCTCATAGGTACTTCCACT
GTCCCTCGTGTGGTTATGAGAACGATCGTGACGTTATTGCTATCATGAATTTAA
ATGGGAGGGGGTCTCTGACCCTCTCGACTGCCCCTCAGATGAGAGATGTAATCC
CGAATCGATGAGGGGAACTATGAACCCCTTGGGAGAGAACCCTCACCCTTCCAA
GGCGGGGAGGAAGTCAGTGGCCC

4. GP68

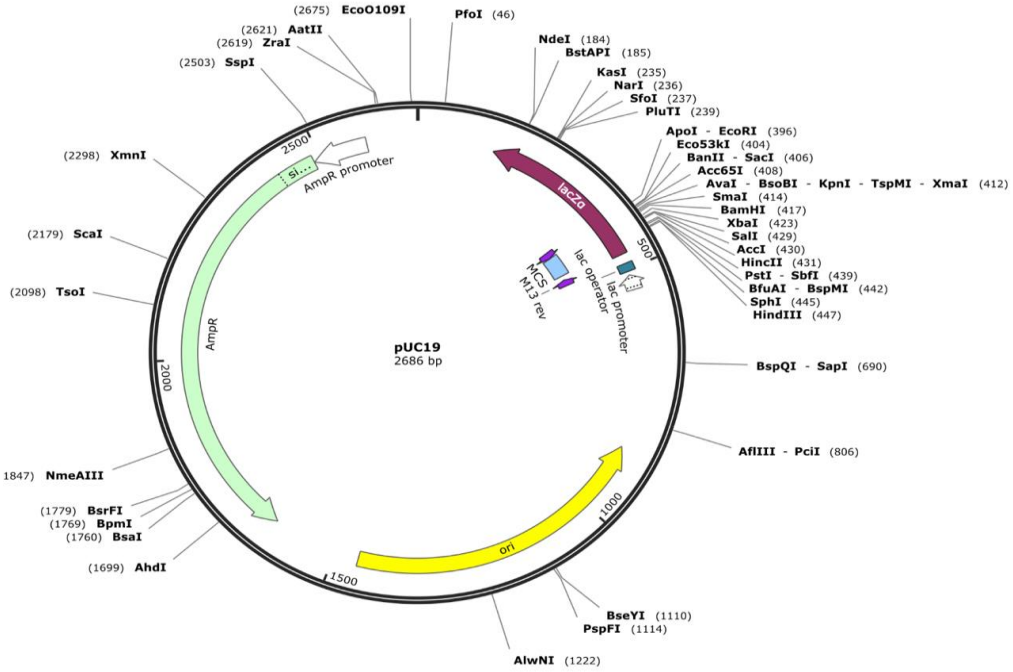
GGGTTAAGGGAGTAGAAAGGGGGTGGATAGCCCCCTTATAGATTTAAAAACTTC
TTTTTCTACCAATCTTCAATGCATCATCATCATCATCATCATCATCATCATATG
GCTAGGAGGGAGAAAAACCCAATAAGAGCCACAGTTGCGATGAATATTGGCTTA

TCTGACTCCCTCCTAGCCTTCGTGAACAACTACGTTAAGGCACTCCGTTTCTCC
TTGTTCTGGATGAAAGAGAATGTGAAAAATCCAAACGAGAAGGGCGCACTCTCC
AAAGTACACGAGGGATTGTATGAAAAGCTGAGGAAAGAATACAATCTACCATCT
AAAGTTACTGAGGACTGTTATCGTGATGCCCTGGCAATATACAAGAGTTGGTAC
AACAACCCGAAAAAAGGTAGATTCCCCCGTGTCTACAAGCCGACAGTGTGGCTA
ACGCCCAAGCAAAGTTATACTGTAGACTTAGAGAAAATGACAGTCAGGATAGCA
AGTGTTGGCGAACTACCAATACTAGGTTATCCTAGAAACCTAAAGGAGTATGCA
AACTGGAAGATGAAGGAGGCTAGGCTAACAATCAAGGATGGCAAGGCTCTCCTC
AAAGTAACTTTTGAGAAGGAAGAAGAGAAAGTTAAACCAAAAGACAGTGTTGCT
GTTGATATAAACATGAATGACATTGTCGTTGGTAAGGATGACACTCACTACGTT
AGGATTCCCACTCGCCTTCACGACGCTCACCACTTCAAGTCATTAGCTGAGAAT
TTGCAGAAGAAGTATCCTAGAAGGTGGAAGCAAAATAGGAGAATTCTACACAGG
GCACGCTCTTTTCATCAAAAGGCCAAACTAATTATGGAGGACTACGCTAGGAAG
GTTGGTAAGTGGGTTGTTGAGATTGCTGAGGGTTTGGGTGCCAACGTCATAAAG
CTTGAGGACTTGAAGAACCTCATCAAGGACGTTAATAAGCTACCAGCTGAATTT
CGCGATAAACTATACTTGATGCAATATCGTCGTATTCAGTATTGGATAGAGTGG
CAGGCTAAGAAACACGGAATGATTGTGGAGTTTGTTAATCCTAGTTACTCTTCC
GTTTCTTGCCCAAAGTGTGGCCACAAAATGGTTGAGATTGCTTATAGGTACTTT
CACTGTCCTTCATGTGGTTATGAGAACGATCGTGATGTTATTGCTATCATGAAT
TTAAATGGGAGGGGGTCTCTGACCCTCTCGACTGCCCCTCAAATGAGAGATGTA
GCTCCGAATCGATGAGGGGAACCCTCGCTGACGGGGAGGAAGTCAGGCCTTCCT
AAGGGATGGATAGCCATTATGTTTTTAAACTTTCTATCAAGATTCCCTGGGT

D. pET30a(+) plasmid map:



E. pUC19 plasmid map:



F. pTAM inserted sequence (TAM+Target DNA)

TAM (uppercase) + guide complementary (lowercase) sequences
(The insert is cloned between BamHI and HindIII)

GP40: TTATAggtactactaggatgc; TTTAAggtactactaggatgc; TTTATggtactactaggatgc
GP68: TTATAaggccttcctaaggga; TTTAAaggccttcctaaggga; TTTATaggccttcctaaggg
a
GP43: TTATAagtggccc; TTTAAagtggccc; TTTATagtggccc
GP10: TTATAggtgaggatactcta; TTTAAggtgaggatactcta; TTTATggtgaggatactcta