

Identification of non-coding RNAs in the hyper thermophilic bacterium *Thermotoga maritima* MSB8 through comparative genomics and *in-silico* analyses

Rahma Alshamrani¹, Muhammad Huzaifa Arshad², Munazza Gull³, Muhammad Zohaib Nawaz⁴, Sana Malik^{2*}

¹Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

²Bioenergy Research Centre, Department of Bioinformatics & Biotechnology, Government College University, Faisalabad, Pakistan

³Biochemistry Department, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

⁴State Key Laboratory of Marine Environmental Science, Institute of Marine Microbes and Ecospheres, College of Ocean and Earth Sciences, Xiamen University, Xiamen, China

Received:

August 23, 2021

Accepted:

September 14, 2021

Online First:

November 03, 2021

Published:

January 09, 2022

Abstract

Thermotoga maritima is a hyperthermophile with the potential to produce thermostable commercial enzymes which can be used for saccharification of plant biomass for subsequent fermentation to bioproducts. The molecular mechanism involved in the hyper thermostability in this bacterium is still not well-understood. It is known that small non-coding RNAs (ncRNAs) regulate and modulate the gene expression of various biological processes at transcriptional and post-transcriptional levels, hence coordinate the adaptation processes in response to environmental stimuli in both prokaryotes and eukaryotes. To understand the role of small ncRNAs in the hyper thermostability of *T. maritima*, an *in silico*-based approach was employed involving the identification of the ncRNAs in this bacterium on a genome-wide scale. A novel pipeline was constructed which involved a combination of various bioinformatics algorithms. In total, 20804 orthologous groups were predicted on the genome of *T. maritima* and 20 other bacteria (reference genomes) by the OrthoMCL tool. By using the “Perl” and “Bash” languages 258 orthologous IGR datasets were created. Among these datasets, small ncRNAs were identified by employing RNaz and RNA Infernal tools. Total 28 ncRNA candidates were predicted by the RNaz tool and 9 candidates were confirmed as novel *cis*-regulatory small ncRNAs in *T. maritima* MSB8 by Infernal tool and were named as Tmn (*T. maritima* ncRNAs). This work provides novel insights into the role of ncRNAs in the stress adaptability of MSB8 and can give a much better understanding of the lifestyle of this bacterium after validation of the data through wet-lab approaches. Having a clear understanding of the thermo-tolerance mechanism, the MSB8 can be exploited in the future for the commercial production of thermostable compounds and biohydrogen.

Keywords: *Thermotoga maritima*, Small non-coding RNAs, Genome-wide detection, Thermostability, Plant biomass, Saccharification

How to cite this:

Alshamrani R, Arshad MH, Nawaz MZ, Gull M and Malik S, 2022. Identification of non-coding RNAs in the hyper thermophilic bacterium *Thermotoga maritima* MSB8 through comparative genomics and *in-silico* analyses. Asian J. Agric. Biol. 2022(1): 202105224. DOI: <https://doi.org/10.35495/ajab.2021.05.224>

*Corresponding author email:
sanabiotech94@gmail.com

This is an Open Access article distributed under the terms of the Creative Commons Attribution 3.0 License. (<https://creativecommons.org/licenses/by/3.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Introduction

Chemically reactive environment of the hydrothermal vents supports different ecosystems and provides optimum conditions for the stable synthesis of prebiotics (Martin et al., 2008). From these hydrothermal vents, a hyperthermophilic rod like bacterium *Thermotoga maritima* was isolated (Singh et al., 2015) which can grow between the temperature range of 60-90°C with optimum growth at 80°C (Huber et al., 1986). The hyperthermophilic nature of *T. maritima* makes it an attention grabber for the production of various industrially important biocatalysts including esterases, lipases, and amylases, because of their extremely high thermal stability and efficiency over other enzymes for high-temperature industrial processes (Levisson et al., 2007; Mehmood et al., 2014). This bacterium has been the focus of many evolutionary biologists because it harbors the characteristics of early microorganisms on earth and can also produce a high amount of biohydrogen by fermentation (Singh et al., 2015), along with acetic acid and carbon dioxide. Like different other organisms which produce heat shock proteins (Lindquist and Craig, 1988), MSB8 tends to incorporate more “charged” amino acids on the surfaces of their proteins, to form an impenetrable hydrophobic core (Kumar and Nussinov, 2001). This bacterium has a sheet-like structure or toga present on its poles, made up of porin (*OmpB*), and an alpha-helical protein (*OmpA1*) (Ranjit and Noll, 2016) forming a trimer of proteins, enclosing the whole bacterium. This toga helps in nutrient uptake when MSB8 is in hostile conditions as it expands the surface-to-volume ratio of the cell (Jiang et al., 2006). Understanding the molecular mechanism behind this unique property of hyper thermostability in *T. maritima* is an interesting and critical research challenge. Different studies have shown that understanding the exact molecular mechanism behind stress tolerance and increasing or decreasing this tolerance in any organism is very challenging (Zhang et al., 2017; Zhang et al., 2019). Over the past few years, researchers have put a lot of effort to unravel the mechanism involved in the hyper thermostability in *T. maritima* following computational and experimental approaches. Different structural and genetic properties have been associated with the thermostability of *T. maritima*, but due to the huge and complex structural and genomic data, the real picture is still unclear.

During the past few years, research involving the detection of small non-coding RNAs is very much accelerated because they play a key role in the regulation of prime biological activities including cellular metabolism, pathogenicity, stress adaptation, and virulence in both prokaryotes and eukaryotes (Vanderpool et al., 2011). The small non-coding RNAs are encoded in intergenic regions (IGRs) of the genome and were unnoticed for a long period because of their difficult detection following wet-lab approaches (Tsai et al., 2015). They can modify the activities of protein by binding to them and can also affect the translation and stability of mRNAs (Storz et al., 2004). Studies have shown the regulatory roles of different ncRNA in *Escherichia coli*, including the translation of a stress response transcription factor, *RpoS*, stimulated by *DsrA* (ncRNA) which regulate its gene expression (McCullen et al., 2010), *DicF* plays a role as an antisense cell division inhibitor (Balasubramanian et al., 2016), while *RhyB* controls the homeostasis of iron, that act as a cofactor in various enzymatic reactions (Masse et al., 2007). The defense mechanism of prokaryotes and eukaryotes comprise CRISPR (Clustered Regulatory Interspaced Short Palindromic Repeats) and RNAi (RNA interference) respectively and both are the analogs of each other (Felden and Paillard, 2017). Jørgensen et al. (2013) reported the role of a small ncRNA in the regulation of biofilm production in *E. coli*. The multicellular adhesive ncRNA (*McaS*) was found to trigger the synthesis of an exopolysaccharide named b-1,6 N-acetyl-D-glucosamine (PGA) when an RNA binding protein, CsrA (carbon storage regulator A) binds to *McaS*. They showed that the inactivation of CsrA binding sites in *McaS* RNA affects the biofilm production, the binding ability of CsrA, and the regulation of PGA (Jørgensen et al., 2013). Similarly, Massé and Gottesman (2002) studied a small ncRNA named *RhyB* in *E. coli* that is involved in the regulation of iron metabolism. When bacteria are in non-favorable conditions and the mineral supply is limited, the *RhyB* downregulates the iron-storage ability in the bacteria (Massé and Gottesman, 2002). Given the above studies highlighting that the small ncRNAs coordinate the processes of adaptation in response to different environmental factors, integrate the environmental stimuli, thus control the target gene expression (Storz et al., 2011), we hypothesized that small ncRNAs might be involved in the hyper thermostability of *T. maritima* MSB8. To validate our hypothesis, we employed an *in silico* approach based



on comparative genomics to detect the small non-coding RNAs in the genome of *T. maritima* MSB8 and twenty other bacterial genomes using different online available tools and databases. The predicted ncRNAs were further characterized to evaluate their functions and sequence conservation. To our knowledge, this is the first study based on the *in silico* identification of small ncRNAs in the genome of *T. maritima* MSB8. Discovery of the novel small ncRNAs in MSB8 and analyzing their sequence conservation patterns across the Thermotogales will facilitate the elucidation of gene regulation and molecular mechanism involved in the extreme thermostability of this group of bacteria and will also give insights into the origin of microbial life on the planet earth.

Material and Methods

A unique pipeline was constructed involving different tools/software step by step to identify the novel ncRNAs in the genome of *T. maritima* MSB8 (Figure-1).

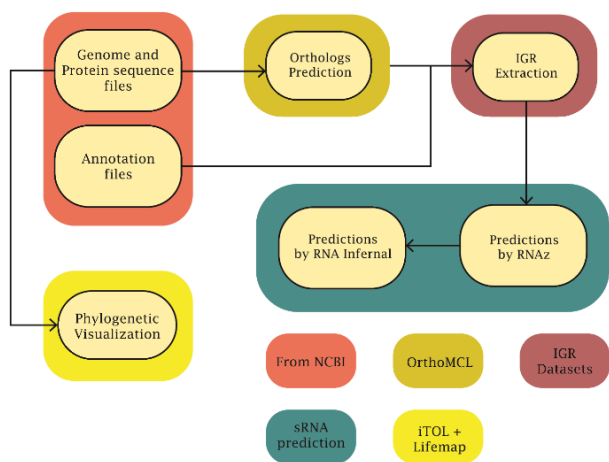


Figure-1: The biomolecular pipeline developed in the current study to identify the novel ncRNAs in MSB8.

Prediction of orthologous groups of genes

To identify the taxonomically related sequences, the prediction of the orthologous relationship of genes among different genomes is important. In this study, 20 fully sequenced bacterial genomes along with the genome of *T. maritima* MSB8 were retrieved from the NCBI Genome database in FASTA and GFF format. Using OrthoMCL, an online tool based on the Markov Cluster algorithm, the orthologous groups

were created across the whole genome of MSB8 and reference genomes by following a comparative genomics-based approach. The OrthoMCL tool uses a scalable method to group the orthologs and paralogs of different genomes of interest.

Extraction of IGR candidates for ncRNA genes

The intergenic regions of the prokaryotic and eukaryotic genomes are most likely to contain the non-coding RNA genes; therefore, it is important to extract the potential IGR candidates for ncRNA genes. To create each IGR dataset, the sequences from both 3' (*trans*) and 5' (*cis*) untranslated regions (UTRs) of each gene of MSB8 and other reference genomes were extracted following the scripts written in “Perl” and “Bash” languages.

For 5'-end datasets, 250 nt upstream and 20 nt downstream of the transcription start site of the genes present in a positive direction were selected and considered as one set of 5'-UTRs. For the genes in the negative direction, 250 nt downstream and 20 nt upstream of the transcription start site were taken. To attain amplified sensitivity and specificity, three overlapping windows (−250~ −100, −200~ −50, −150~ +20) were created for IGRs in both positive and negative directions. Likewise, for 3'-end datasets, 250 nt downstream and 20 nt upstream of the transcription start site of the genes present in a positive direction were selected and referred to as one set of 3'-UTRs. For the genes in the negative direction, 250 nt upstream and 20 nt downstream of the transcription start site were taken. For this IGR dataset, overlapping windows were −20~150, 50 −200, 100 −250 (Nawaz et al., 2017).

Prediction of non-coding RNAs by RNAz

A comparative genomics-based approach was employed to predict the bacterial non-coding RNAs using a popular and reliable tool RNAz (Washietl et al., 2005) individually hosted by two different web servers namely ViennaRNA Web Services and Galaxy server. Unlike other RNA-structure prediction tools, RNAz follows a different approach based on two prime facts: i) structural conservation and ii) unusual thermodynamic stability of the ncRNA which allows for the steadfast *de novo* detection of these non-coding RNAs in different genomes (Gruber et al., 2010). It compares the “minimum free energy” (MFE) of a query sequence with the MFEs of a huge number of random multiple-aligned sequences having a similar length and base

composition, using a highly sensitive support vector machine (SVM) build on Rfam and RNAalifold algorithm for structure prediction. As mentioned above, the IGR datasets were created from sequences of both 3' (*trans*) and 5' (*cis*) UTRs in the target genome and their orthologs in the reference genomes. Then multiple sequence alignment of these IGR datasets was performed by CLUSTALW, a comparatively reliable alignment tool based on the PAM substitution matrix. The multiple aligned sequences were then subjected to secondary structure prediction with default parameters using the RNAz tool.

Secondary structure conservation of ncRNAs by RNA Infernal

Once the secondary structure of non-coding RNA is predicted, it is essential to identify the homologs of ncRNA having the conserved secondary structure in other closely related species. For this purpose, a robust RNA motif-searching software Infernal (INFERence of RNA Alignment) (<http://eddylab.org/infernal>) was used. This software allows the identification of RNA homologs having a more conserved secondary structure than the primary sequence using the Rfam database (Nawrocki, 2014). Infernal applies statistical models of secondary structure and sequence of RNA families known as “Covariance models” (CMs), that are well-matched for finding RNA similarity. These CMs are created from known homologs of bacterial species by doing their multiple sequence alignment and using the *cmbuild* program (Nawrocki and Eddy, 2013; Barquist et al., 2016). After building CMs, the next step was to “calibrate” the model using the *cmcalibrate* program which gives the approximate probabilistic parameters required to calculate the *E-values* (expectation values). Next, the sequence database was searched for homologs of the RNA family using the *cmscan* program. The CM score greater than 10 represents the extremely confident hits (Nawrocki et al., 2009). Finally, these predicted ncRNAs are considered as conserved ncRNAs as they are orthologous to each other present at the 5'-UTR of the orthologous genes.

Phylogenetic analyses

To represent the evolutionary relationships among the bacterial species used in this study, 16S rRNA-based phylogenetic tree was constructed using the multiple-sequence-alignment data from whole-

genome sequence alignment through the “Mauve” tool. This tool takes FASTA files as input and constructs a tree based on the neighbor-joining method (Darling et al., 2004). The output tree was in *Newick* format, which was then visualized using iTOL (Interactive Tree Of Life), an online freely available tool. It supports *nexus*, *Newick*, and *phyloXML* formats of trees (Letunic and Bork, 2019).

Results

In closely related species, the secondary structures of small non-coding RNAs are considered to be conserved facilitating the identification of homologous genes in evolutionarily conserved organisms. This study was based on a comparative-genomics approach for the genome-wide detection of small ncRNAs in *T. maritima* and different bacterial genomes. Here, we selected 20 reference bacterial genomes based on the evolutionary relatedness with *T. maritima* (Figure-2). by retrieving data from NCBI's Genome database. The sequence genome file (.fna) contains a genomic sequence in FASTA format. Similarly, the sequence protein file (.faa) contains proteins annotated on the genome assembly in FASTA format. The annotation GFF file (.gff) contains annotation of the genomic sequence in Generic Feature Format Version 3. Likewise, the annotation GenBank file (.gbff) is a flat-file format for genomic sequences in the assembly and has genome sequences and their annotation. The novel ncRNA candidates were detected in *T. maritima* and 20 bacterial genomes after subjecting them to a series of bioinformatics tools and following a comparative genomics-based approach. Only the positive hits from the result of previous tools were further subjected to analyses.

The identification of orthologous groups among different genomes of bacteria under study was carried out using OrthoMCL. The output was a tab-delimited text file with seven different columns, each representing a particular thing. Out of a total of seven columns in the output file, the first three columns are of particular interest, representing the IDs of proteins, IDs of orthologous groups present in OrthoMCL-DB (database), and the IDs of those proteins in OrthoMCL-DB which were selected as the best hits, respectively. Out of all the entries, 20804 entries of orthologous groups along with the names of bacteria and the protein ID of the corresponding orthologous group were filtered out and stored in a separate text file.



The genes for small RNAs are encoded within intergenic regions (IGRs) of the bacterial genomes (Koo et al., 2011). So, the best strategy to detect sRNAs on the genomic scale by using an *in silico*-based approach is the extraction of the IGRs from whole genomes. After applying scripts written in Perl and Bash languages, a total of 956 files were created, out of which, 129 entries were selected and stored in a separate text file. Then using this data, 258 datasets were created (with window size as mentioned in the methods section) for both 3' and 5' UTRs (129 datasets for both UTRs).

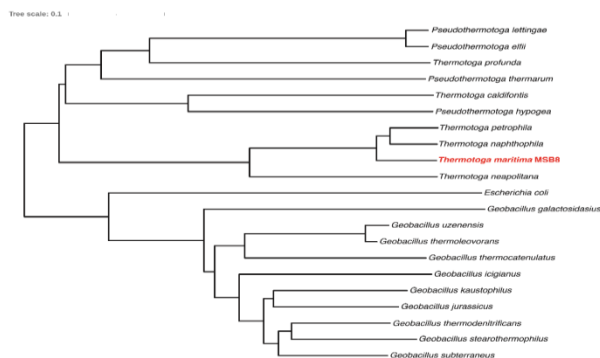


Figure-2: Phylogenetic relationship among all the bacterial species discussed in this study.

Orthologous group of ncRNA genes and IGR datasets

The identification of orthologous groups among different genomes of bacteria under study was carried out using OrthoMCL. The output was a tab-delimited text file with seven different columns, each representing a particular thing. Out of a total of seven columns in the output file, the first three columns are of particular interest, representing the IDs of proteins, IDs of orthologous groups present in OrthoMCL-DB (database), and the IDs of those proteins in OrthoMCL-DB which were selected as the best hits, respectively. Out of all the entries, 20804 entries of orthologous groups along with the names of bacteria and the protein ID of the corresponding orthologous group were filtered out and stored in a separate text file.

The genes for small RNAs are encoded within intergenic regions (IGRs) of the bacterial genomes (Koo et al., 2011). So, the best strategy to detect sRNAs on the genomic scale by using an *in silico*-based approach is the extraction of the IGRs from whole genomes. After applying scripts written in Perl and Bash languages, a total of 956 files were created, out of which, 129 entries were selected and stored in a separate text file. Then using this data, 258 datasets were created (with window

size as mentioned in the methods section) for both 3' and 5' UTRs (129 datasets for both UTRs).

Prediction of ncRNAs by RNAz

Secondary structural motifs of conserved ncRNAs were predicted by employing the RNAz tool. Using the CLUSTALW tool, 231 sequence datasets were successfully aligned out of a total of 258 IGR datasets. The aligned datasets were then subjected to the RNAz tool: 87 datasets for 3'-UTRs and 144 datasets for 5'-UTRs, respectively. The output file of the RNAz tool included the positively predicted ncRNAs, their minimum free energy (MFE), mean pairwise identity, GC content, and the Shannon entropy value. If the prediction displays “RNA”, that dataset is considered to have functional RNA secondary structure and if the prediction displays “OTHER”, that means no RNA secondary structure is predicted. Out of these datasets, only 28 IGR datasets were positively predicted by RNAz to have a functional RNA secondary structural element. The output files of both RNAz tools hosted by the Galaxy server and ViennaRNA Web Services are identical except for the HTML output of the latter, which provides a more user-friendly and pictorial display (Figure-3A and 3B).

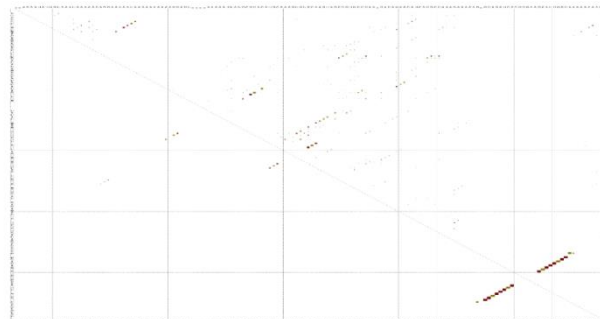


Figure-3A: Secondary structure diagram of RNA predicted by the RNAz tool.



Figure-3B: Probabilities of base-pairing for sequences in the dataset as Dot plot.

Conservation of predicted ncRNAs by RNA Infernal

To determine the sequence conservation of predicted ncRNAs among MSB8 and the closely related bacterial species, the conserved motifs in reliable ncRNA candidates were searched using the RNA Infernal tool. Among the 28 confident hits of RNAz, 19 were *trans-acting* ncRNAs and 9 were *cis-acting* ncRNAs (CM score > 10). Analysis of the conservation pattern of predicted ncRNAs reveals that 9 *cis-acting* ncRNAs were MSB8-specific ncRNAs, conserved in the genome of MSB8. However, all the *trans-acting* ncRNAs when searched in the Rfam database were found to be conserved in other closely related species under study and involved in various biological functions. The *trans-acting* ncRNAs regulate the expression of the genes present on the different loci, while the *cis-acting* ncRNAs involve in the regulation of genes present on the same genetic locus. *Trans-acting* ncRNAs are ubiquitous, usually present at the 3'-UTR of the genes, and are linked to every global response in bacteria, because of which these are well-documented and extensively characterized. Conversely, *cis-acting* ncRNAs are less in number, commonly found at the 5'-UTR of the genes, and are usually associated with the responses to different environmental stimuli (Papenfort et al., 2015). In this study, the predicted *cis-acting* ncRNAs are less in number than the *trans-acting* RNAs so, we only focused on the former ones because of the limited information available about them.

Annotation of ncRNAs for function inference

To classify all the predicted *cis-acting* ncRNAs, the Rfam database was used, which provides a collection of ncRNAs families mainly represented by their conserved secondary structure, CM score, and multiple sequence alignment. The ncRNAs were annotated for their functional characterization. The information about the proteins regulated by these predicted sRNAs was then searched in NCBI and UniProt databases (Table-1). Among the 9 *cis-acting* ncRNAs, *Tmn2* was annotated as *mir-96*, a micro RNA that plays a key role in the regulation of genes involved in the development of sensory organs (Geng et al., 2018). *Tmn3* was annotated as *FsrA*, a small RNA involves in the iron regulation (Smaldone et al., 2012), *Tmn7* was annotated as *Afu_294*, a small nucleolar RNA, involves in the methylation of specific rRNA (Jöchl et al., 2008), and the *Tmn9* was

annotated as *TB10Cs2H2*, another small nucleolar RNA, acts as the guide RNA for the pseudouridylation of nucleotides (Liang et al., 2005). There were no orthologs documented in the Rfam database for the remaining 5 *cis-acting* ncRNAs (SSRC34, snoR27, Afu_203, SpF41_sRNA, and TnuHACA3) which implies that these are the novel ncRNA candidates in the genome of MSB8.

Table-1: Name of the predicted ncRNAs, the gene, and the proteins which these ncRNAs regulate along with the function performed by those proteins.

ncRNAs	Gene	Target protein	Size of protein (aa)	Molecular function of protein
<i>Tms1</i>	TM_0667	Hypothetical protein TM0667	256	Deoxyribonuclease activity
<i>Tms2</i>	TM_0562	Hypothetical protein TM0562	182	Integral part of membranes
<i>Tms3</i>	dnaA	Chromosomal replication Initiation protein dnaA	440	ATP binding DNA replication origin binding
<i>Tms4</i>	TM_1183	Aldo/keto reductase Oxidoreductase	379	Iron-sulfur cluster binding Oxidoreductase activity
<i>Tms5</i>	TM_1183	Aldo/keto reductase Oxidoreductase	379	Iron-sulfur cluster binding Oxidoreductase activity
<i>Tms6</i>	TM_0337	UPF0313 protein TM_0337	558	4 iron, 4 sulfur cluster binding Catalytic activity Iron ion-binding
<i>Tms7</i>	TM_1791	Hypothetical protein TM1791	244	Unknown (RAMPs) Repair associated mysterious proteins
<i>Tms8</i>	TM_0179	Hypothetical protein TM0179	92	Unknown
<i>Tms9</i>	TM_0179	Hypothetical protein TM0179	92	Unknown

Discussion

Small non-coding RNAs are known to be novel biomarkers that play a key role in the regulation and expression of the different genes. However, there is still a huge research gap in the detection of ncRNAs and the identification of their roles in various metabolic processes in all organisms. The wet lab-based methods have not been developed to identify the non-coding RNA up to now. In the present study, we tried to bridge this gap by detecting the small ncRNAs following a bioinformatics tools-based scheme in the genome of a thermophilic anaerobic eubacterium, *T. maritima* MSB8, which might have a



role in the extreme thermostability of this bacterium. From the orthologous genes present in the genome of MSB8 and reference bacterial genomes, we created the IGR datasets (sequences that exist among the annotated open reading frames) by taking the sequences from 5' and 3'-UTRs. We subjected these datasets to two different reliable bioinformatics tools to predict ncRNAs in our target bacteria and the reference genomes. Based on the thermodynamic stability and structural conservation, few ncRNAs were considered as positive hits by the RNAz tool, which was then searched in the RNA infernal, the motif searching tool. The positively predicted small ncRNAs present at the 5'-UTRs of the genes by RNAz and showing an infernal score of ≥ 10 bit were regarded as "confident hits". Sequence conservation pattern reveals that all the identified *cis-acting* ncRNAs were conserved in the genome of MSB8 and the *trans-acting* ncRNAs were conserved in the other closely related species presented here. The *trans-acting* ncRNAs are the abundantly existing and extensively studied class of regulatory RNAs in bacteria, as they are involved in almost every metabolic response in bacteria. Similarly, in this study, 68% out of the total identified ncRNAs, belong to the *trans-acting* ncRNAs, a dominant class of regulatory RNAs. All the identified *cis-acting* ncRNAs were annotated using the Rfam database against the known regulatory motifs for their functional characterization. For some of these *cis-acting* ncRNAs, no orthologs were found in the Rfam database, which makes them novel candidates. By using RNAfold and Infernal tools, 13 putative ncRNAs were identified in the genome of MSB8, but the secondary structures for these regulatory RNAs could not be predicted (Latif et al., 2013). Amin et al. (2016) detected the differential expression of 63 sRNAs in the genome of *Salmonella enterica* in response to the starvation stress during different growth phases (Amin et al., 2016). In another study, a similar approach was employed as ours to detect the small RNAs in the genome of *Shewanella piezotolerans*, a deep-sea psychrotolerant bacterium (Nawaz et al., 2017). Transcriptome data analysis of *Agrobacterium* strains has shown the involvement of 384 identified small regulatory RNAs in the virulence induction cycle (Raja et al., 2018). They identified 16 *cis-acting* small RNAs in the genome of *Shewanella* which might play a role in the adaptation of this bacterium to extreme environmental conditions. Recently, genome-scale analysis of *Azospirillum*

brasilense has revealed the presence of 59 small RNAs, 53 of which are specified as the novel, because of the unavailability of their homologs in Rfam and BSRD (bacterial sRNAs database). These novel sRNAs are found to be associated with potential mRNA targets linked to the stress adaptability mechanism of this bacteria (Koul et al., 2020). In line with these findings, some of the *cis-acting* ncRNAs identified in our study have shown their roles in different key regulatory functions in MSB8 which may or may not be associated with the adaptation of MSB8 to extreme environmental conditions. However, these findings require validation through qPCR and microarray-based wet-lab approaches and a more detailed study is required to decipher the functioning of all the ncRNAs in the genome of MSB8.

Conclusion

Thermotoga maritima is a hyperthermophile with the potential to produce thermostable commercial enzymes and biohydrogen. However, the molecular mechanism involved in the hyper thermotolerance of this bacterium is still not well-understood. The small non-coding RNAs (ncRNAs) have been shown to regulate the gene expression of various biological processes at transcriptional and post-transcriptional levels and coordinate in the adaptation processes in response to environmental stimuli. However, molecular biology or biotechnology-based techniques are not developed to directly investigate the non-coding RNAs. Here, a bioinformatics-based novel pipeline using "Perl" and "Bash" languages was developed to identify the small ncRNAs with their potential roles in the hyper thermotolerance of *T. maritima*. In total 28 ncRNA candidates were predicted by the RNAz tool and 9 candidates were confirmed as novel *cis-regulatory* small ncRNAs in *T. maritima* MSB8 by Infernal tool and were named as Tmn (*T. maritima* ncRNAs). This work provided novel insights into the role of ncRNAs in the stress adaptability of MSB8 and can give a much better understanding of the lifestyle of this bacterium after validation of the data through wet-lab approaches. Having a clear understanding of the thermo-tolerance mechanism, the MSB8 can be exploited to further understand thermotolerance for its subsequent industrial applications.



Acknowledgment

Authors are highly obliged to the Higher Education Commission (HEC), Pakistan for sponsoring the Ph.D. of Ms. Sana Malik under HEC's Indigenous Ph.D. Fellowship program.

Disclaimer: None.

Conflict of Interest: None.

Source of Funding: This research was financially supported by the Higher Education Commission (HEC) Pakistan.

References

- Amin SV, Roberts JT, Patterson DG, Coley AB, Allred JA, Denner JM, Johnson JP, Mullen GE, O'Neal TK and Smith JT, 2016. Novel small RNA (sRNA) landscape of the starvation-stress response transcriptome of *Salmonella enterica* serovar typhimurium. *RNA Biol.* 13: 331-342.
- Balasubramanian D, Ragunathan PT, Fei J and Vanderpool CK, 2016. A prophage-encoded small RNA controls metabolism and cell division in *Escherichia coli*. *Msystems.* 1: 00021-15.
- Barquist L, Burge SW and Gardner PP, 2016. Studying RNA homology and conservation with infernal: from single sequences to RNA families. *Curr. Protoc. Bioinformatics.* 54: 12.13.1-12.13.25.
- Darling AC, Mau B, Blattner FR and Perna NT, 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14: 1394-1403.
- Felden B and Paillard L, 2017. When eukaryotes and prokaryotes look alike: the case of regulatory RNAs. *FEMS Microbiol. Rev.* 41: 624-639.
- Geng R, Furness DN, Muraleedharan CK, Zhang J, Dabdoub A, Lin V and Xu S, 2018. The microRNA-183/96/182 cluster is essential for stereociliary bundle formation and function of cochlear sensory hair cells. *Sci. Rep.* 8: 1-13.
- Gruber AR, Findeiß S, Washietl S, Hofacker IL and Stadler PF, 2010. RNAz 2.0: improved noncoding RNA detection. *Biocomputing*, World Scientific. 69-79.
- Huber R, Langworthy TA, König H, Thomm M, Woese CR, Sleytr UB and Stetter KO, 1986. *Thermotoga maritima* sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90 C. *Arch. Microbiol.* 144: 324-333.
- Jiang Y, Zhou Q, Wu K, Li X-Q and Shao W-L, 2006. A highly efficient method for liquid and solid cultivation of the anaerobic hyperthermophilic eubacterium *Thermotoga maritima*. *FEMS Microbiol Lett.* 259: 254-259.
- Jöchl C, Rederstorff M, Hertel J, Stadler PF, Hofacker IL, Schrettl M, Haas H and Hüttenhofer A, 2008. Small ncRNA transcriptome analysis from *Aspergillus fumigatus* suggests a novel mechanism for regulation of protein synthesis. *Nucleic Acids Res.* 36: 2677-2689.
- Jørgensen MG, Thomason MK, Havelund J, Valentin-Hansen P and Storz G, 2013. Dual function of the McaS small RNA in controlling biofilm formation. *Genes Dev.* 27: 1132-1145.
- Koo JT, Alleyne TM, Schiano CA, Jafari N and Lathem WW, 2011. Global discovery of small RNAs in *Yersinia pseudotuberculosis* identifies Yersinia-specific small, noncoding RNAs required for virulence. *Proc. Nat. Acad. Sci.* 108: E709-E717.
- Koul V, Srivastava D, Singh PP and Kochar M, 2020. Genome-wide identification of *Azospirillum brasilense* Sp245 small RNAs responsive to nitrogen starvation and likely involvement in plant-microbe interactions. *BMC Genomics.* 21: 1-16.
- Kumar S and Nussinov R, 2001. How do thermophilic proteins deal with heat?. *Cell. Mol. Life Sci.* 58: 1216-1233.
- Latif H, Lerman JA, Portnoy VA, Tarasova Y, Nagarajan H, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Lee D-H and Qiu Y, 2013. The genome organization of *Thermotoga maritima* reflects its lifestyle. *PLoS Genet.* 9(4): e1003485.
- Letunic I and Bork P, 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47: W256-W259.
- Levisson M, van der Oost J and Kengen SWM, 2007. Characterization and structural modeling of a new type of thermostable esterase from *Thermotoga maritima*. *The FEBS J.* 274: 2832-2842.
- Liang XH, Uliel S, Hury A, Barth S, Doniger T, Unger R and Michaeli S, 2005. A genome-wide analysis of C/D and H/ACA-like small nucleolar RNAs in *Trypanosoma brucei* reveals a trypanosome-specific pattern of rRNA modification. *RNA.* 11: 619-645.
- Lindquist S and Craig EA, 1988. The heat-shock proteins. *Annl. Rev. Gen.* 22: 631-677.
- Martin W, Baross J, Kelley D and Russell MJ, 2008. Hydrothermal vents and the origin of life. *Nat. Rev. Microbiol.* 6: 805-814.



- Massé E and Gottesman S, 2002. A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. Proc. Nat. Acad. Sci. 99: 4620-4625.
- Masse E, Salvail H, Desnoyers G and Arguin M, 2007. Small RNAs controlling iron metabolism. Curr. Opin. Microbiol. 10: 140-145.
- Mehmood MA, Shahid I, Hussain K, Latif F and Rajoka MI, 2014. Thermodynamic properties of the β -glucosidase from *Thermotoga maritima* extend the upper limit of thermophilicity. Protein Pept. Lett. 21(12): 1282-1288.
- McCullen CA, Benhammou JN, Majdalani N and Gottesman S, 2010. Mechanism of positive regulation by DsrA and RprA small noncoding RNAs: pairing increases translation and protects rpoS mRNA from degradation. J. Bacteriol. 192: 5559-5571.
- Nawaz MZ, Jian H, He Y, Xiong L, Xiao X and Wang F, 2017. Genome-wide detection of small regulatory RNAs in deep-sea bacterium *Shewanella piezotolerans* WP3. Front. Microbiol. 8: 1093.
- Nawrocki EP, 2014. Annotating functional RNAs in genomes using Infernal, pp. 163-197. In Gorodkin J., Ruzzo W. (eds), RNA sequence, structure, and function: computational and bioinformatic methods, vol. 1097. Humana Press, Totowa, NJ.
- Nawrocki EP and Eddy SR, 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 29: 2933-2935.
- Nawrocki EP, Kolbe DL and Eddy SR, 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics. 25: 1335-1337.
- Papenfort K, Förstner KU, Cong JP, Sharma CM and Bassler BL, 2015. Differential RNA-seq of *Vibrio cholerae* identifies the VqmR small RNA as a regulator of biofilm formation. Proc. Nat. Acad. Sci. 112: E766-E775.
- Raja I, Kumar V, Sabapathy H, Kumariah M, Rajendran K and Tennyson J, 2018. Prediction and identification of novel sRNAs involved in Agrobacterium strains by integrated genome-wide and transcriptome-based methods. FEMS Microbiol. Lett. 365: fny247.
- Ranjit C and Noll KM, 2016. Distension of the toga of *Thermotoga maritima* involves continued growth of the outer envelope as cells enter the stationary phase. FEMS Microbiol. Lett. 363: fnw218.
- Singh A, Sevda S, Abu Reesh IM, Vanbroekhoven K, Rathore D and Pant D, 2015. Biohydrogen production from lignocellulosic biomass: technology and sustainability. Energies. 8: 13062-13080.
- Singh R, Gradnigo J, White D, Lipzen A, Martin J, Schackwitz W, Moriyama E and Blum P, 2015. Complete genome sequence of an evolved *Thermotoga maritima* isolate. Genome Announc. 3: e00557-15.
- Smaldone GT, Antelmann H, Gaballa A and Helmann JD, 2012. The FsrA sRNA and FbpB protein mediate the iron-dependent induction of the *Bacillus subtilis* lutABC iron-sulfur-containing oxidases. J. Bacteriol. 194: 2586-2593.
- Storz G, Opdyke JA and Zhang A, 2004. Controlling mRNA stability and translation with small, noncoding RNAs. Curr. Opin. Microbiol. 7: 140-144.
- Storz G, Vogel J and Wassarman KM, 2011. Regulation by small RNAs in bacteria: expanding frontiers. Mol. Cell. 43: 880-891.
- Tsai CH, Liao R, Chou B, Palumbo M and Contreras LM, 2015. Genome-wide analyses in bacteria show small-RNA enrichment for long and conserved intergenic regions. J. Bacteriol. 197: 40-50.
- Vanderpool CK, Balasubramanian D and Lloyd CR, 2011. Dual-function RNA regulators in bacteria. Biochimie. 93: 1943-1949.
- Washietl S, Hofacker IL and Stadler PF, 2005. Fast and reliable prediction of noncoding RNAs. Proc. Nat. Acad. Sci. 102: 2454-2459.
- Zhang MM, Xiong L, Tang YJ, Mehmood MA, Zhao ZK, Bai FW and Zhao XQ, 2019. Enhanced acetic acid stress tolerance and ethanol production in *Saccharomyces cerevisiae* by modulating expression of the de novo purine biosynthesis genes. Biotechnol. Biofuel. 12(1): 1-13.
- Zhang M, Zhang K, Mehmood MA, Zhao ZK, Bai FW and Zhao X, 2017. Deletion of acetate transporter gene *ADY2* improved tolerance of *Saccharomyces cerevisiae* against multiple stresses and enhanced ethanol production in the presence of acetic acid. Bioresour. Technol. 245: 1461-1468.

Contribution of Authors

Alshamrani R: Reviewed literature and wrote the first draft of manuscript

Arshad MH: Data collection, analysis and interpretation

Nawaz MZ & Malik S: Conceptualization, funding acquisition and research supervision and manuscript write up

Gull M: Literature review and editing of manuscript

